基于交互引导的问答对联合生成模型

刘 杰 林绍鑫 王善鹏
（南开大学人工智能学院 天津 300350）

摘 要 大规模问答对的自动生成在知识问答库构建和机器阅读理解等许多应用具有关键价值.尽管其重要性已得到广泛认可,现有问答对生成方法仍面临着严峻挑战.首先,在传统的问答对生成模型中,抽取式的答案获取方法难以适用于复杂的自然交互场景.相比较而言,生成式模型通过对文本的语义理解,能够自动生成表述更加自然的答案.其次,对于问答对生成任务来说,为了防止生成的答案和问题出现语义上的不匹配,需要更全面地捕捉并增强答案生成和问题生成两个子任务之间的交互.最后,由于答案抽取和问题生成存在任务难度的差异,这两个任务在联合训练的过程中会出现任务之间的优化不平衡问题.为此,本文提出了一个基于交互引导的问答对联合生成模型(Interaction-Guided Joint Abstractive QAPs Generation Model, IGJA-QAP).具体而言,本文设计了一个带有答案引导的多头门机制的联合生成模型,同时对两个子任务进行统一建模并有效地捕获和增强它们之间的信息交互,从而可以生成语义上匹配的问答对.本文在三个大规模数据集SQuAD、NewQA和CoQA上进行了综合全面的实验分析.本文提出的模型在答案生成任务上METEOR值平均分别超出其他最佳方法3.0%、5.9%和4.3%,问题生成任务上METEOR值平均分别超出其他最佳方法1.5%、0.5%和2.1%.实验结果表明,本文提出的模型达到了目前最高的性能.

关键词 问答对生成;统一生成式模型;答案引导的多头门;指针网络;相互优化
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.00251

Question-Answer Pairs Generation Based on Interaction-Guided Joint Abstractive Model

LIU Jie LIN Shao-Xin WANG Shan-Peng
(*College of Artificial Intelligence*, *Nankai University*, *Tianjin* 300350)

Abstract Automatically generating large-scale question-answer pairs is valuable for many applications such as knowledge base construction and machine reading comprehension. Although its importance has been widely recognized, existing approaches to question-answer pair generation still face serious challenges. First, in traditional question-answer pair generation models, extractive answer acquisition methods are difficult to apply to complex natural interaction scenarios. In contrast, generative models can automatically generate answers with more natural expressions through semantic understanding of text. Second, for the question-answer pair generation task, the interaction between the two subtasks of answer generation and question generation needs to be captured and enhanced more comprehensively in order to prevent semantic mismatches between the generated answers and questions. Finally, due to the difference in task difficulty between answer extraction and question generation, the joint learning of the two tasks of answer extraction and question generation can lead to an optimization imbalance between the two subtasks during training. For this reason, this paper proposes an Interaction-Guided Joint

收稿日期:2022-12-05;在线发布日期:2023-08-29.本课题得到国家自然科学基金(61976119)资助.**刘 杰**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、数据挖掘、机器学习.E-mail:jliu@nankai.edu.cn.**林绍鑫**,硕士研究生,主要研究领域为自然语言处理.**王善鹏**,学士,主要研究领域为自然语言处理.

Abstractive QAPs Generation Model (IGJA-QAP). Specifically, this paper designs a joint generation model with an answer-guided multiheaded gate mechanism that simultaneously models two subtasks in a unified manner and efficiently captures and enhances the information interactions between them, so that semantically matching question-answer pairs can be generated. In this paper, a comprehensive experimental analysis is conducted on three large-scale datasets, SQuAD, NewQA and CoQA. The proposed model outperforms the best methods by 3.0%, 5.9% and 4.3% on average for the answer generation task and 1.5%, 0.5% and 2.1% on average for the question generation task, respectively. The experimental results demonstrate that our model achieves state-of-the-art performance.

Keywords question-answer pairs generation; unified abstractive model; multi-head answer-guide gate; pointer network; mutual optimization

1 引 言

大规模问答对在许多应用程序中起着至关重要的作用,例如辅助构建知识库、改进搜索引擎[1],以及训练聊天机器人进行流畅的对话[2-3]. 然而,根据给定文档通过人工标注获得问答对的方式是耗时且代价昂贵的. 因此,自动地从给定文档中生成高质量的问答对是至关重要的.

现有大部分关于生成问答对[1,4-5]生成的方法主要分为流水线式结构和端到端结构. 流水线式结构是将问题抽取和答案生成任务依次训练,首先生成答案或问题,然后生成问题或答案. 该结构将前者获取的结果作为后者模型的部分输入,从而实现问题或答案引导生成的效果. 最近,一些研究人员针对流水线式结构存在的问题提出了端到端的模型[6-9],这种联合训练框架可以同时完成答案抽取和问题生成两个子任务的训练,实现答案抽取任务和问题生成任务之间的交互,从而生成匹配的问答对.

尽管关于问答对生成方面的研究已经取得一定的进展,但这一任务仍然面临着重要的挑战. 首先,无论是在流水线式结构还是端到端方法中,传统方法的抽取式答案获取方式不足以生成自然并且全面的问答对. 抽取式的方法意味着从文章中截断一段连续的文字片段作为答案,但它不能自然地表达复杂的语义. 其次,答案抽取和问题生成两个任务在语义理解上具有相关性,而现有的方法在训练过程中限制了两种任务之间的语义交互,不足以生成更加匹配且与文章密切相关的问答对. 虽然生成的答案及其对应的问题在语法和内容上的表示截然不同,但它们在原文中都对应着同一位置的内容,因此具有相同的潜在语义信息. 然而,流水线式结构的方法忽略了答案抽取和问题生成之间的关系. 即使端到端方法[5]通过共享一个公共编码器在两个子任务训练时交换了语义信息,但是这不足以确保解码器输出的答案和问题具有共同潜在语义信息. 最后,由于答案抽取和问题生成在任务难度的差异,答案抽取和问题生成两个任务的联合学习会导致在训练时两个子任务之间的优化不平衡. 如图1所示,答案提取旨在预测长度为N的输入文本上的边界标签分布,而问题生成产生解码序列中每个位置在长度为V的整个词汇表上的分布. 两任务在输出空间上的显著差异带来了任务难度的不平等,导致优化中对两个任务的重视程度不对等. 而这种不对等会使得模型在梯度优化时偏向更简单的任务类型,难以达到两个子任务共同优化的效果[10-11].

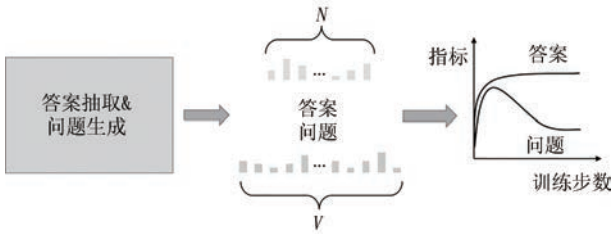


图1 传统端到端模型的不平衡优化效果展示(中间部分是生成的问题和答案分别在词典和输入序列上的分布. N表示输入序列的长度,V表示词典的大小)

为了应对这些挑战,本文提出了基于交互引导的问答对联合生成式模型(Interaction-Guided Joint Abstractive QAPs Generation Model, IGJA-QAP). 本文提出的模型在端到端方法的基础上,利用生成式方法替代抽取式方法来生成答案,以便为复杂的语义提供更好的表达能力. 与此同时,利用完全生成式的方法来获取答案和问题可以改善任务难度的

不平等，减少答案抽取和问题生成之间的不平衡. 另外，引入指针网络机制[12]可以让答案生成器从上下文信息中直接复制，从而保留了答案抽取的优点. 为了能在答案和问题的解码过程中能保持相同的上下文语义信息，本文提出将问答对生成的解码过程集成到同一解码器中. 通过这种方式，答案生成和问题生成之间可以实现信息交互并相互增益，以生成匹配且与上下文密切相关的高质量问答对. 因此，联合生成式模型可以为问答对的生成带来相互优化，避免生成相关性较弱的答案与问题. 此外，Liu等人[1]提出，当从一段文本中获取问答对时，答案可被视为指导问题生成的线索. 因此，本文利用答案引导的多头门机制将跨任务信息从答案生成转移到问题生成，从而更好地实现任务之间的信息交互并引导问题的生成.

为了验证模型的有效性，本文在三个基准数据集上进行了大量的实验：SQuAD、NewsQA和CoQA. 实验结果表明模型IGJA-QAP达到了目前最高的性能指标. 本文的工作总结如下：

(1)本文提出一种基于交互引导的问答对联合生成模型IGJA-QAP. 统一式结构使得问题生成和答案生成在训练时能够保持相同的上下文语义信息，而联合训练可以实现信息的交互. 通过针对性设计的答案引导的多头门机制，两个子任务可以有效地传递跨任务信息，从而实现相互优化. 此外，该模型可以自然地平衡子任务的复杂性并解决优化不平衡问题. 因此，本文的模型IGJA-QAP能够生成匹配且与上下文密切相关的问答对.

(2)本文设计了一个带有指针网络的生成式答案生成模块，它使答案生成模块不仅能生成自然且新颖的文本，而且能从文档中精确地复制关键信息到答案中，从而生成自然且与上下文密切相关的答案.

(3)本文在三个基准数据集上进行了大量的实验，以证明模型IGJA-QAP的有效性. 对比实验结果表明本文提出的联合学习框架可以显著提高阅读理解. 消融实验的结果也证明了模型中各部分模块设计的有效性，同时人工评估也显示了该模型的高质量生成能力.

2　相关工作

问答对生成任务作为自然语言处理领域的重要任务之一，其目标为学习和理解给定的文本，并能够从中自动地生成高质量的问答对. 问答对生成任务可以分为两个相关的对偶任务，其中包括问答(Question Answering，QA)和问题生成(Question Generation，QG).

(1)问答任务

问答任务旨在根据问题在上下文中搜索相关的信息并从中找到最优答案[13]. 在问答任务中，Rajpurkar等人[14]提出首个知识问答大规模数据集SQuAD. 该数据集定义了一个抽取式问答任务，其中答案来自对应文档中的连续片段. 抽取式问答任务是从文档中预测子序列片段来回答问题，包括指出答案边界或选择文档中连续单词的跨度. 在过去几年中，大部分关于知识问答的工作都集中在答案抽取上，其中包括QANet[15]、BiDAF[16]和VQAP[7]. 但是，这种边界预测或者抽取式机制对于文档的综合表达能力不强，不适合更开放的任务[17-21]，比如生成式问答任务和问题生成任务. 目前，基于预训练的生成式模型，如BART[3]、UNILM[22]和T5[23]，具有强大的生成和理解能力，能够生成高质量的答案，因此可以有效地解决问答任务现存的难题. 为了提高生成答案的准确性和相关性，一些工作[24-27]利用基于人工反馈的强化学习框架(Reinforcement Learning from Human Feedback，RLHF)来增强模型的生成能力. Nakano等人[25]使用人类反馈替代自动化评测方法(如ROUGE、BLEU)，并将人类反馈作为奖励进行强化学习，从而构造目标优化函数对生成模型进行训练. Ouyang等人[27]使用来自人类偏好的强化学习来训练QA模型引用具体论据来论证答案，有助于提高答案的可信度.

(2)问题生成任务

大多数早期的问题生成工作主要采用基于人工设计的模板或基于预先设计的规则将局部文本转换为不同形式的问题[28-30]. 这种方法首先预处理给定的文本，从中抽取出可用于生成问题的局部文本，然后按照规则或模板对抽取出来的文本生成相关的问题[3,13,31]. 但是这种方式需要语言领域的专家提前设计规则或模板，消耗了大量的人力资源. 随着深度学习的发展，基于神经网络的端到端问题生成任务的研究逐渐成为热点，其中包括答案引导的问题生成和辅助信息引导的问题生成. 答案引导的问题生成需要上下文信息和答案信息作为输入，其中答案信息一般是答案的位置信息. 大多数研究工作利用带有注意力机制的编码解码框架去解决这种问题[4,32-33]，不同点在于融入答案信息的策略不同，例

如利用答案位置的嵌入表示[34-35]、答案单独编码[36]、表示上下文与答案之间的相对距离[37]等. 除此之外,为了提高生成答案的质量,有很多的工作在编码器中加入了额外的辅助信息[2,36,38-39],从而控制生成难度和质量. 另外,引入了额外的实体和标记信息[40]也可以辅助模型选择部分文本从而生成高质量的问题. Du等人[41]提出了一种分层神经句子级序列标记模型来识别值得提问的句子. 但是,现有的工作难以自动地直接从原始文本中生成无答案引导的问题,并且这些技术大多包含难以调整整体性能的独立组件.

(3) 问答对生成任务

目前,问答对生成的研究工作主要集中在流水线式的模型上[1,4-5]. 流水线式的模型包括答案引导的问答对生成模型和问题引导的问答对生成模型. 答案引导的问答对生成模型首先在文中选择可以被提问的片段,然后模型基于现在选择出来的片段学习如何提问,最后检测出已经生成问题相对应的答案位置. 而问题引导的问答对生成模型与上述方法相似,只是生成问题的先后顺序不一样. 例如,Du等人[4]提出了一个新的神经网络模型,它通过一种新颖的门控机制结合了共指知识,以检测有价值的答案,然后生成一个有答案的问题. 与此相似,Golub等人[42]提出了一个包含答案标注模块和问题合成模块的两阶段同步网络SynNet,用于问答对生成. 另外,Liu等人[1]模仿人类提出问题的方式,通过预先定义的各种信息抽取方式抽取出answer-clue-style-aware等一系列信息,从而引导多样性问题的生成. 虽然流水线式的架构符合人类的生成逻辑,但是这种方式在两阶段的训练过程中会产生严重的累积误差,以至于生成不匹配的问答对. 由于问题生成和答案抽取两个任务在形式上是对偶的,有些研究通过训练同一个模型分别解决两个任务[17,43-44],即用同一个模型分别生成问题和答案. 但是由于这种双重约束的执行形式,两个任务在同一模型上的训练目标函数会对彼此产生消极的影响. 为了克服这些缺点,Cui等人[6]为问答对生成引入了一站式模型OneStop. 该模型将问题生成和答案抽取集成到一个统一的框架中,同时生成问题和答案. 然而,由于答案提取和问题生成难度上的差异性,在两任务上的联合训练会导致优化不平衡现象,即该模型会偏向于优化简单的任务. 此外,仅仅通过同一个编码器实现的QA和QG之间的交互并不能保证解码输出的共同语义. Lyu等人[45]提出了基于BERT的流水线式模型,该模型从摘要中启发式地生成答案和问题. Dugan等人[46]借助摘要生成模型简化输入的上下文信息,从而提高问答对与上下文的相关性. Back等人[47]通过学习恢复包含答案的句子来预训练问题生成模型,之后采用流水线式模型生成问答对. Shakeri等人[8]提出了端到端的问答对生成模型,在一个生成器中同时预测问题和答案,并将生成问答的可能性作为衡量标准. 本文提出的基于交互引导的联合生成式模型IGJA-QAP可以有效地解决上述问题. IGJA-QAP模型首先将答案抽取任务替换为答案生成任务,很好地平衡任务之间的难度,减少不平衡的优化带来的影响. 其次,该模型融入两个任务之间的对偶性,将两个任务嵌入到同一个编码器和解码器中同时训练,让两个任务在训练时能够相互交流和相互促进. 通过这样的方式,解码器解码出来的信息也能保持一致,更好地保证答案和问题之间的语义相似度,从而通过不同的生成器生成匹配的知识问答对.

3 问答对生成模型

本节将详细介绍基于交互引导的联合问答对生成模型IGJA-QAP. 其中3.1节概述任务目标和整体模型结构;3.2节和3.3节分别介绍该模型包含的答案生成任务和问题生成任务;3.4节详细阐述了目标函数的组成形式.

3.1 模型框架

本文所研究的问答对生成任务针对N个词d_i构成的文档文字序列$D=(d_i)_{i=1}^{N}$,分别生成长度为M的问题文字序列$Q=(q_t)_{t=1}^{M}$和长度为L的答案文字序列$A=(a_l)_{l=1}^{L}$. 针对此任务,本文所提出的问答对生成模型的学习目标如下:

$$\begin{aligned} Q,A &= \operatorname{argmax}_{Q,A} P(Q,A|D) \\ &= \operatorname{argmax}_{Q,A} \mathrm{P}(A|D;\theta)\mathrm{P}(Q|A,D;\theta) \end{aligned} \quad (1)$$

为了生成多样性的知识问答对,本文采用了集束搜索策略[46]进行解码生成.

考虑到答案和问题之间是具有相同的潜在语义信息,并且答案生成和问题生成可以作为对偶任务. 如图2所示,本文提出的模型由编码器-解码器、答案生成模块和问题生成模块构成,其中编码器-解码器采用Transformer结构,答案生成模块包含着指针网络和答案生成器,问题生成模块包含着答案引

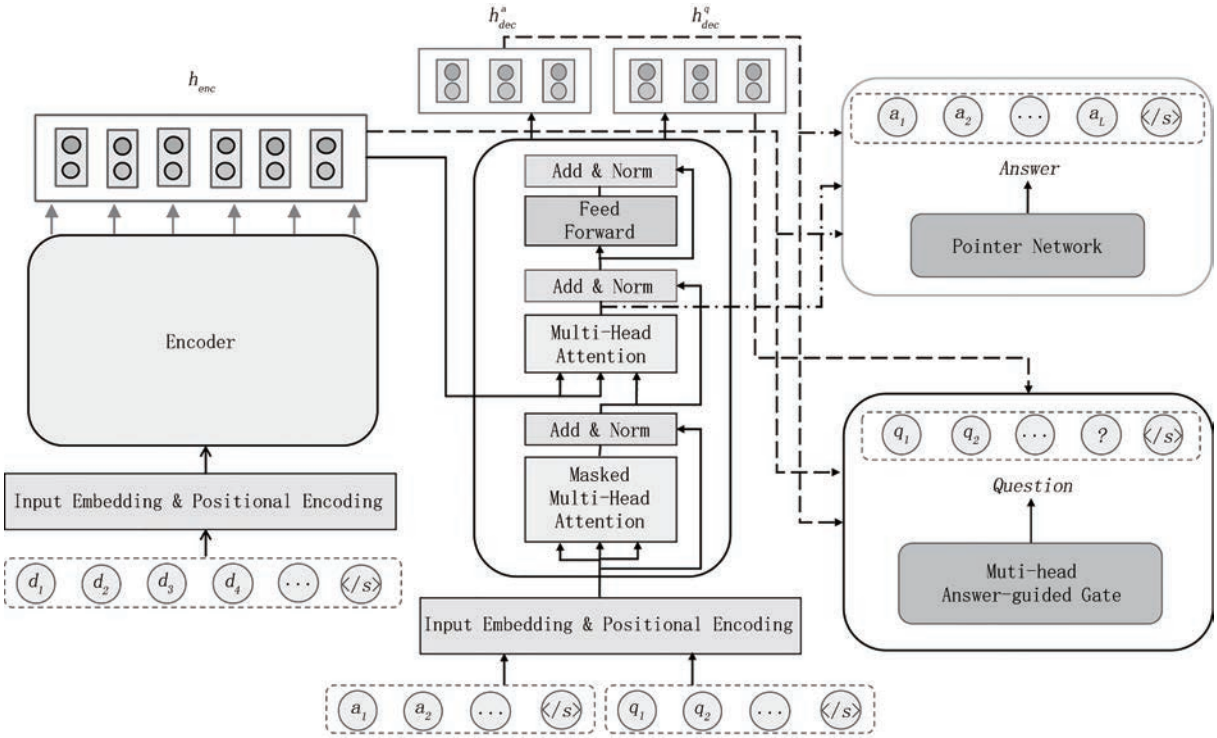


图2　模型框架图

导的多头门机制和问题生成器. 本文在对上下文编码后，将答案和问题的解码过程融合到一个统一的解码器中，然后通过指针网络生成答案. 最后，答案引导的多头门机制利用生成的答案引导对应问题的生成，形成问答对. 这种方式能够更好地促进两任务之间的交互，并且能够使得解码后的答案和问题能保持相同的上下文语义信息，从而在解码过程中增强两任务之间的跨任务信息交流. 如图2所示，IGJA-QAP模型包含两个任务：答案生成和问题生成. 首先，编码器接受输入文本$D=(d_i)_{i=1}^{N}$生成上下文的编码向量h_{enc}，并将其送入到解码器解码得到答案的解码信息h^a_{dec}. 为了结合抽取式方法的优势，本文借助指针网络，不仅能在词典上生成答案序列，也能直接从上下文中复制信息，从而生成最终的答案序列. 在获取答案的解码信息后，本文以同样的方式获取问题的解码信息h^q_{dec}. 为了提高生成的答案与问题之间的相关性，本文设计了答案引导的多头门机制，以此融合来自答案的解码信息. 这种方式使得问题的解码信息更关注于生成的答案信息，从而实现任务之间的引导生成和跨任务信息交流. 本文将获得的融合信息送入问题生成器中，生成最终的问题序列. 由于基于Transformer[48]编码-解码框架的预训练模型T5[49]在生成方面展现了强大的性能，同时它也在知识问答任务上表现优异，因此本文工作采用T5作为预训练模型. 另外，生成模型在解码时容易发生解码不终止问题[50]，因此本文在输入的末尾加入了</s>符号以防止不间断生成的问题.

3.2　答案生成

不同于流水线式的框架或传统端到端的模型将答案获取的过程转换为位置预测的抽取式任务，本文将其定义为序列到序列的生成任务. 假设字典长度为$|V|$，隐状态的维度为z. 编码器读取输入序列$D=(d_i)_{i=1}^{N}$，并且产生隐状态序列$h_{enc}=(h_i)_{i=1}^{N}$，其中$h_i\in R^d$. 然后h_{enc}被传到解码器中解码成$h^a_{dec}=(h^a_l)_{l=1}^{L}$，解码过程中生成交叉注意力序列$\alpha^a_{dec}=(\alpha^a_l)_{l=1}^{L}$. 之后单个字符的解码表示$h^a_l$经过生成器得到在词典上的分布$P_{voc}$，如下面公式所示：

$$P_{voc}(w_l)=\mathrm{softmax}(W_a h^a_l+b^a) \tag{2}$$

其中$W_a\in R^{|V|\times z}$和$b^a\in R^{|V|}$是可学习的参数.

事实上，在大多数生成场景中，答案的部分片段往往是可以直接从输入文本中获取的. 因此本文提出的模型借助指针网络直接复制输入文本的信息到输出分布中. 如图3所示，在指针网络的帮助下，答案生成部分不仅可以从字典中直接生成，而且还可以复制上下文内容. 通过这种方式，生成的答案不仅能够准确并且自然地复述上下文中的内容，而且

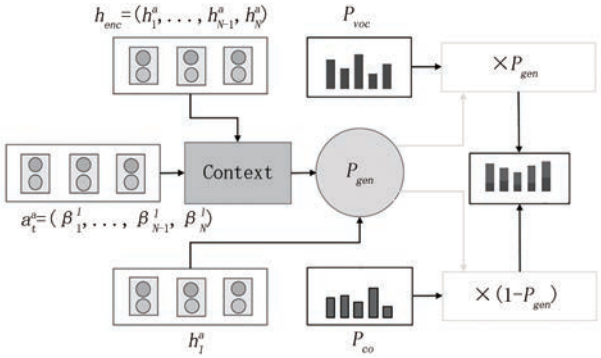


图3 生成答案的指针网络

在生成的同时具有抽取式的优点. 在指针网络中，首先编码隐状态 $h_{\text{enc}}=(h_i)_{i=1}^N$ 在第 l 个词的交叉注意力分数 $\alpha_l^a=(\beta_1^l,\cdots,\beta_{N-1}^l,\beta_N^l)$ 上获取加权平均和，如下式所示，得到上下文向量 $\boldsymbol{c}_l$：

$$\boldsymbol{c}_l=\sum_{i=1}^{N}\beta_i^l h_i \tag{3}$$

上下文向量 $\boldsymbol{c}_l$ 可以被认为是从上下文中获取的固定长度的表示，与第 l 个词解码后的隐状态 h_l^a 相连接，经过一层非线性变换得到软开关 $P_{\text{gen}}\in[0,1]$，如下式所示：

$$P_{\text{gen}}=\sigma\left(W_{\text{gen}}\left[h_l^a;\boldsymbol{c}_l\right]+b_{\text{gen}}\right) \tag{4}$$

其中 W_{gen} 和 b_{gen} 都是可学习变量，σ 是 sigmoid 函数. 软开关 P_{gen} 可以用来决定从生成或者上下文中复制的权重，通过下式得到最终第 l 个词的概率分布 $P_a(w_l)$：

$$\gamma_t^l=\sum_{i:w_i=w_t}^{N}\beta_i^l \tag{5}$$

$$P_{\text{co}}(w_l)=\left[\gamma_1^l,\gamma_2^l,\cdots,\gamma_{|V|}^l\right] \tag{6}$$

$$P_a(w_l)=P_{\text{gen}}P_{\text{voc}}(w_l)+\left(1-P_{\text{gen}}\right)P_{\text{co}}(w_l) \tag{7}$$

上式 γ_t^l 表示词典中第 t 个词所对应的交叉注意力和，$P_{co}(w_l)\in R^{|V|}$ 表示从上下文复制的词在字典上的分布.

3.3 问题生成

在获取答案之后，IGJA-QAP 模型将生成的答案作为线索来引导问题的生成. 问题生成模型借助答案解码后隐藏向量 h_{dec}^a，利用答案引导的多头门机制，从而帮助问题的生成，使得问题和答案更加匹配.

如图4所示，答案引导的多头注意力机制包括一个多头的注意力模型和一个答案引导的门机制. 问题生成模型在获取问题解码的隐状态 $h_{\text{dec}}^q=(h_1^q,\cdots,h_{M-1}^q,h_M^q)$ 后，借助多头的注意力模型去捕获

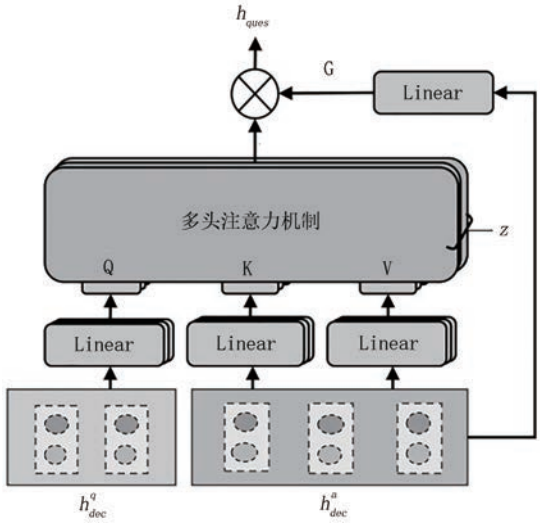


图4 答案引导的多头门机制

答案和问题之间语义相关性，如下式所示：

$$Q_j,K_j,V_j=W_j^Q h_{\text{dec}}^q,W_j^K h_{\text{dec}}^a,W_j^V h_{\text{dec}}^a \tag{8}$$

$$\text{head}_j=\text{softmax}\left(\frac{Q_j\times K_j}{\sqrt{k}}\right)V_j \tag{9}$$

$$S_{\text{ques}}=\text{Concat}\left(\text{head}_1,\cdots,\text{head}_y\right) \tag{10}$$

公式中 $W_j^Q\in R^{k\times z}$，$W_j^K\in R^{k\times z}$，$W_j^V\in R^{k\times z}$ 都是可学习的权重矩阵，在经过交叉注意力后，可以得到融合答案的问题状态表示 S_{ques}.

为了防止不匹配的问题生成，本文利用一个门机制 G 去更好地吸收答案的状态信息，公式如下所示：

$$G=W_g h_{\text{dec}}^a \tag{11}$$

$$h_{\text{ques}}=S_{\text{ques}}\odot G \tag{12}$$

这里 $\odot$ 表示元素积.

在获取答案最终的隐藏层表示后，利用问题生成器去生成在词典上的状态分布 $P_q(w_m)$，如下式所示：

$$P_q(w_m)=\text{softmax}\left(W_q h_m^q+b_q\right) \tag{13}$$

其中，$W_g\in R^{M\times L}$，$W_q\in R^{|V|\times z}$，$b_q\in R^{|V|}$ 都是可学习的参数.

3.4 损失函数

基于公式(1)，可以优化生成序列的对数似然函数来更新模型参数 θ，损失函数如下：

$$\Phi=-\log P_{\text{a}}(\omega)-\log P_{\text{q}}(\omega)=\Phi_a+\Phi_q \tag{14}$$

其中 Φ_a 和 Φ_q 分别表示答案和问题的交叉熵损失.

利用公式(14)中的损失去更新模型参数，使用集束搜索策略[51]生成多样性且语义匹配的问答对.

由于在本文模型中相同的问答生成方式可以有效解决生成空间的差异，改善任务难度的不平等，与

OneStop[6]模型的损失函数形式不同，本文直接将答案和问题的损失函数相加，而无需引入超参数λ来平衡问题生成和答案生成之间的任务难度. 后面的实验也证明了这种方案的可行性.

4 实　验

4.1 数据集

本文在三个阅读理解数据集上做了大量的实验：SQuAD[14]，NewsQA[52]和CoQA[53]. 在数据源处理上，本文首先将SQuAD和NewsQA数据集的上下文切分成多段，使得每段都只包含一个问答对，即一条数据包含上下文、一个答案和一个问题，然后过滤掉答案不在上下文的数据. 由于CoQA数据的答案是标注人员根据上下文总结得到，仅用于完全生成式模型，本文可以用这个数据集来验证本文提出模型的生成能力. 考虑到SQuAD和CoQA数据集的测试集都没有对外公布，因此需要从它们的验证集中截取部分数据作为测试集. 表1展现了处理后所有数据集的统计信息.

表1　处理后数据集统计信息

数据集	SQuAD	NewsQA	CoQA
训练集	36 078	92 449	108 647
验证集	1584	5166	2395
测试集	4009	5122	5588
上下文的平均长度	25.77	36.5	10.6
问题的平均长度	11.6	7.7	6.4
答案的平均长度	3.8	5.5	2.9

4.2 对比模型和评价指标

本文在问题生成和答案生成两个任务上做了大量的实验. 考虑到问答对生成的任务类型，本文选取了问题生成和答案生成的模型作为各自任务的对比方法. 为了评价问题生成的质量，本文将采用下面几种模型作为对比模型：

DeepNQG[49]. 一种基于注意力机制的序列到序列问题生成模型，即输入上下文序列输出问题序列. 端到端的生成模型，不依赖任何人工定义的规则.

T5-QG. 一种基于大型预训练模型T5[23]的问题生成模型，采用端到端的框架，输入上下文序列输出问题序列. T5[23]采用统一的编码解码框架，将问题都转换成text-to-text格式，并在大量跨语言数据C4上进行训练. 该预训练模型在许多基准任务上获得了最好的表现，这些基准涵盖了摘要、问题回答、文本分类等等.

T5-A2QG. Golub等人[42]提出一种两阶段流水线式方法. 作为流水线式方法，第一阶段首先将上下文信息作为输入生成答案；然后在第二阶段将生成答案的嵌入向量和上下文信息的嵌入向量相连接生成所对应的问题. 作为对比模型，本文将该模型的基础编码解码模型换成了T5[23].

OneStop[6]. 本文基于预训练模型T5，复现了Cui[6]提出的端到端式的模型OneStop. 该模型是一种端到端的模型，输入上下文信息后，同时生成问题和预测对应答案的起始位置. 该模型编码上下文输入，之后直接送到解码器中进行解码，生成问题. 之后将编码器和解码器的隐状态向量送入自注意力机制，然后预测答案对应的起始位置.

Info-QAP. Lee等人[9]提出了一种新颖的概率生成框架用来生成多样化的问答对. 该模型使用分层条件变分自动编码器学习给定上下文的问题和答案的联合分布，同时最大化它们的交互信息来增强生成的问答对之间的一致性. 为了与本文所提出的模型进行对比，我们去除分层条件变分自动编码器，只保留该模型的问答对生成部分，即问题生成和答案抽取两个子模块.

QA2QG. Dugan等人[46]提出了一种基于预训练模型T5的流水线式问答对生成模型. 该模型先应用摘要生成模型从输入的上下文中提取摘要，然后利用提示学习的方式在预训练模型T5上同时微调问题生成，知识问答和答案抽取三个不同的任务. 最终以流水线的方式先抽取出答案，然后基于抽取出的答案和上下文信息生成对应的问题. 在本文句子级的问答对生成任务中，我们仅使用该模型的问答对生成模型作为对比模型.

为了评价答案生成任务的生成质量，本文除了与OneStop、T5-Info-QAP和QA2QG模型作对比，还与下面的一些模型进行比较：

T5-QA. Alberti等人[54]通过微调预训练模型BERT[55]，输入上下文，直接预测答案的起始位置. 本文将BERT替换成T5[23]作为预训练模型，将抽取式任务变成生成式任务.

T5-MPQA. 按照Song[43]的训练模式，该模型在预训练模型T5上依次训练问题生成和答案生成，即将两个任务整合到一个模型中. 通过这种方式，该模型可以通过结合来自问题生成的信息来提高答案生成的性能.

为了评价生成的问答对与标准问答对之间的相关性，本文采用机器翻译常用的BLEU[56]和METROE[57]以及摘要生成任务中ROUGE-L[58]指标. 其中BLEU-1将平均一元语法精度作为评估相似度的一种手段，而ROUGE-L测量最长的公共子序列以显示流畅性. 考虑到METEOR计算n元语法精度和召回率的调和平均值，本文在分析实验结果时也会将它当作主要分析指标. 在评估抽取式模型的效果之前，本文首先将预测位置转换为相应的文本，之后用生成任务的指标评价抽取式任务.

4.3 实验设置

在本文的实验中，采用来自Google的T5-base作为预训练模型，它具有12层的Transformer结构并且隐状态的维度z为768. 答案引导多头门的头数y为12，隐状态维度为768，因此每个头的k为64. 本文的模型在训练时，样本批量大小设置为16，选择学习率为0.000 01的Adam优化器进行梯度下降. 所有模型都在v100 GPU上进行了300 000步的训练.

4.4 实验结果

表2和表3展现了模型IGJA-QAP与其他基准模型在三个数据集上的实验结果. 从两个表中可以看出，IGJA-QAP同时在问题生成和答案生成上取得了最好的结果. 下面分别从问题生成和答案生成两个任务上进行分析.

表2 在问题生成任务上的对比实验结果

数据集	模型	BLEU-1	ROUGE-L	METEOR
SQuAD	DeepNQG	22.0	41.2	16.2
	T5-QG	37.3	40.5	26.7
	T5-A2QG	34.1	37.9	23.5
	OneStop	35.8	35.4	25.4
	Info-QAP	36.4	35.6	26.6
	QA2QG	36.4	34.2	24.7
	IGJA-QAP	**38.4**	**41.6**	**28.2**
NewsQA	DeepNQG	12.9	36.8	13.4
	T5-QG	30.0	43.5	16.9
	T5-A2QG	30.2	30.9	16.6
	OneStop	28.3	30.0	15.4
	Info-QAP	**35.9**	38.1	15.9
	QA2QG	25.2	36.1	14.6
	IGJA-QAP	30.3	**44.1**	**17.4**
CoQA	DeepNQG	11.4	35.5	11.5
	T5-QG	30.5	41.8	14.2
	T5-A2QG	27.7	40.3	13.0
	QA2QG	13.9	36.6	13.4
	IGJA-QAP	**32.3**	**43.2**	**16.3**

表3 在答案生成任务上的对比实验结果

数据集	模型	BLEU-1	ROUGE-L	METEOR
SQuAD	T5-QA	23.7	54.0	21.2
	T5-MPQG	18.3	55.9	21.0
	OneStop	29.1	43.2	30.0
	Info-QAP	25.9	41.6	28.8
	QA2QG	**30.2**	40.7	25.1
	IGJA-QAP	25.8	**56.1**	**33.0**
NewsQA	T5-QA	31.8	57.0	38.7
	T5-MPQG	18.3	55.9	29.0
	OneStop	29.7	48.9	40.0
	Info-QAP	**35.2**	52.8	42.5
	QA2QG	24.2	38.5	29.4
	IGJA-QAP	27.2	**59.0**	**45.9**
CoQA	T5-QA	18.5	41.0	21.3
	T5-MPQG	21.8	46.3	24.8
	QA2QG	**25.6**	40.4	27.5
	IGJA-QAP	24.3	**48.9**	**29.1**

问题生成. 如表2所示，IGJA-QAP模型在ROUGE-L和METEOR两个指标上取得了最好的表现. 为了评价生成问题的综合质量，本文考虑到METEOR在计算n语法时会同时加入精确率和召回率，利用调和平均值来平衡两个指标的相互影响程度. 因此在下面的分析中，本文将METEOR作为问题生成的主要度量指标. 表2列出了在问题生成任务上的对比实验结果. 与T5-QG相比，由于IGJA-QAP方法利用答案引导问题的生成，有助于生成的问题能够准确地对应答案. 因此在SQuAD、NewsQA和CoQA三个数据集上取得了1.5%、0.5%、2.1%的提高. 对于T5-A2QG的流水线式模型方法，IGJA-QAP模型在SQuAD数据集上分别比T5-A2QG高4.7%，在NewsQA上高出0.8%，在CoQA上高出3.3%，这说明我们的统一生成式模型可以通过问题和答案之间的交互来改进问题生成，从而避免了流水线式方法由于两阶段训练带来的累积误差. 同样对于流水线式方法QA2QG，由于结构上带来的累积误差，该模型在三个数据集上METEOR分别低于IGJA-QAP模型3.5%、2.8%和2.9%. IGJA-QAP在SQuAD数据集上超过OneStop模型2.8%，在NewsQA数据集上超过2.0%. 该实验结果也有效地证明了将问题生成和答案生成融入到同一解码器中可以有效地提高答案和问题之间的跨任务信息交流，使得生成问题更贴近答案和原文. 另外完全生成式的问题答案获取方式能够缓解由于任务难度差异带来的不平衡优化问题，在模型

训练时带来了显著的提升效果. 相比于模型Info-QAP，我们的模型在前两个数据集上METEOR值分别高出1.4%、1.5%. 该模型结构上与OneStop相似，因此同样存在交互不足以及优化不平衡的问题. 然而，由于该模型通过优化问题和答案之间的交互信息提高了它们之间的匹配性和一致性，模型性能上要优于OneStop.

综上，对比实验的结果验证了IGJA-QAP模型可以通过任务之间的充分交互和完全生成式的问答对框架有效地提高了问题生成的质量.

答案生成. 正如在表3中观察到的，与其他的基准模型相比，IGJA-QAP模型在促进答案生成方面得到了显著的提升. 在SQuAD、NewsQA和CoQA三个数据集上，IGJA-QAP模型在ROUGE-L和METEOR指标上都获得了较好的表现. IGJA-QAP在METEOR指标上的平均分数超过了基准模型T5-QA 8.9%，其中包括SQuAD数据集上11.8%的提升、NewsQA数据集上7.2%的提升以及CoQA数据集上7.8%的提升. 与T5-QA模型相比，IGJA-QAP在三个数据集上优秀的表现也证明了将答案生成和问题生成结合在一起时，能够促进任务之间的共同优化，使得生成的答案和问题能够紧贴上下文，弥补了T5-QA模型无答案信息引导的缺点. 而与T5-MPQG模型相比，IGJA-QAP在SQuAD数据集上METEOR值提高了12%，在NewsQA上提高了16.9%，在CoQA上对T5-MPQG提高了4.3%. 针对T5-MPQG的改善证明，IGJA-QAP的统一架构通过增强信息交互有效地提高了知识问答的能力，利用不同的生成器可以避免对偶性带来的负面影响. 我们可以注意到，与OneStop的答案抽取方式相比，IGJA-QAP的生成式方案可以很好地平衡与答案生成之间的难度，从而改善联合训练中存在的不平衡优化问题. 因此相对于OneStop，IGJA-QAP得到了显著的改进，在SQuAD和NewsQA数据集上METEOR值分别提升了3.0%和5.9%. 相比于模型Info-QAP，我们的模型在SQuAD和NewsQA数据集上METEOR值分别提升了4.2%和3.4%. 由于Info-QAP模型在结构上与OneStop上相似，也存在优化不平衡问题. 相比于流水线式模型QG2QA，由于缺少与问题生成之间的交互，在性能上要比我们交互式模型差，在三个数据集上METEOR值分别低7.9%，16.5%和1.6%. 在CoQA数据集上的对比实验结果，可以证明IGJA-QAP模型无论是对于提取式或生成式答案都具有很强的鲁棒性，这也展现出IGJA-QAP强大的生成能力. 尽管IGJA-QAP模型在SQuAD上达到了33.0%、在NewsQA上45.9%、在CoQA上29.1%的最高METEOR分数，但它在BLEU-1上并没有获得最好的成绩. 从案例研究中，可以了解到IGJA-QAP生成的答案比基准模型的更长. 与其他指标相比，由于BLEU值是计算生成的结果与真实结果之间的精确率，生成的答案更长可能会导致BLEU-1较低.

问答对生成. 为了展示问题生成和答案生成相互优化的能力，本文在SQuAD数据集上将IGJA-QAP模型与OneStop模型进行了比较. 在训练OneStop模型时，本文将问题和答案的损失函数用超参数λ相连接，如下所示：

$$\Phi = \lambda \Phi_a + (1-\lambda)\Phi_q \tag{15}$$

公式中Φ_a和Φ_q分别表示答案和问题的损失函数.

虽然OneStop模型尝试将答案抽取和问题生成两个任务统一训练，但是两个任务之间的难度不均衡问题会使得训练倾向于难度相对较小的任务. 如图5所示，我们可以观察到OneStop的问题生成任务的METEOR值迅速达到最高值，随后开始下降，但是答案抽取继续上升. 这种现象源于任务难度的不平等，如图1所示答案抽取任务的生成空间远小于问题生成任务的生成空间，所以答案抽取任务在任务难度上是小于问题生成任务的. 而对于联合模型的训练而言，模型会倾向于易于优化的答案抽取任务，以至于忽视了困难任务的优化[10-11]. 相比之下，IGJA-QAP模型由于将答案抽取式任务转化成答案生成类任务，在任务难度上削弱了这种不平衡现象. 其次，IGJA-QAP模型将两个任务融合到统一的模型中，利用一个解码器和不同的生成器进行生成. 这种统一的架构能够确保解码后问题和答案之间的语义一致性，也能够促进相互之间的优化. 因此，从图5的右半部分可以看出，IGJA-QAP在训练时问题生成和答案生成呈现出相互增长的趋势. 我们还可以注意到，在OneStop模型中，调整权重λ可以减少两个任务的不平衡优化. 比较的结果表明，我们的统一框架在不引入超参数的情况下也可以为问题生成和答案生成带来相互优化.

4.5　人工评价

由于使用自动评估指标BLEU、ROUGE-L以及METEOR等指标只能评估生成数据与真实数据之间的字符相似性，而无法度量生成问答对与目标问答对之间的语义相似度. 目前也没有可用的指标

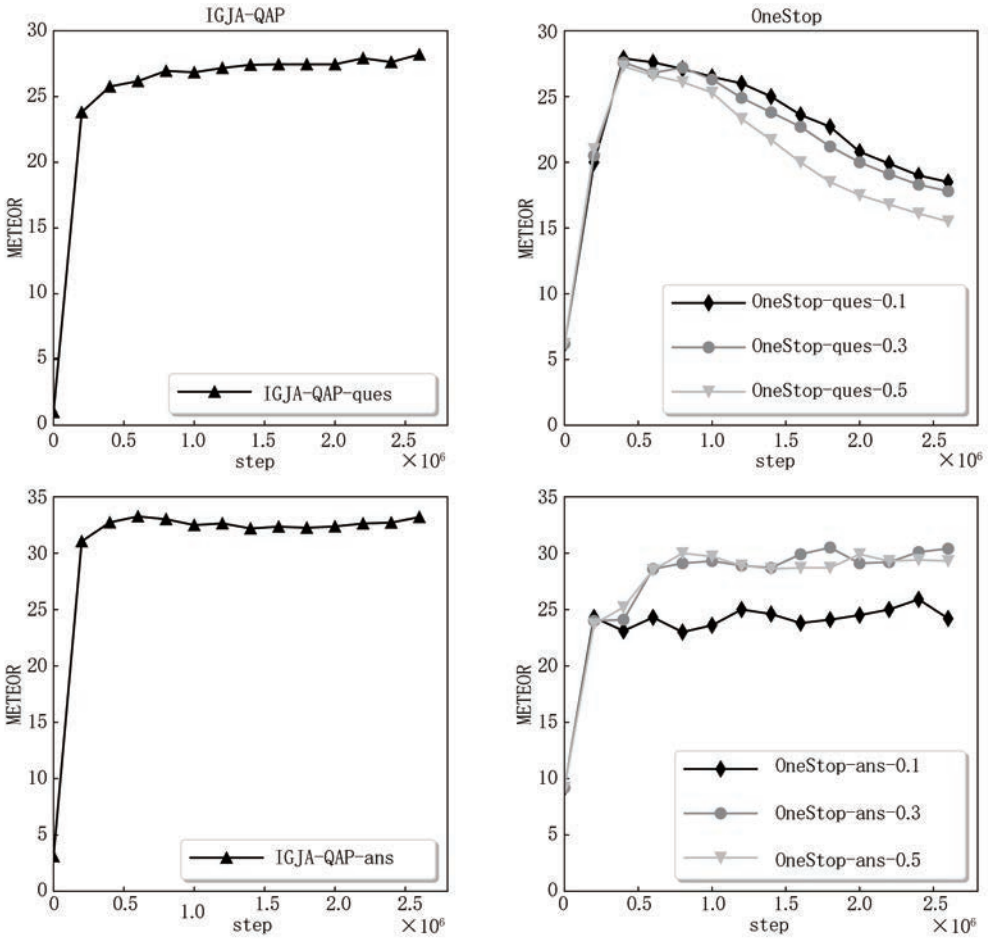


图5 IGJA-QAP和OneStop在训练过程中问题生成和答案生成的指标变化曲线(图中"ques"和"ans"分别代表问题生成和答案生成. 左栏表示在IGJA-QAP上的变化曲线,右栏表示在OneStop上的变化曲线)

能够全面地评价这种语义相似度. 因此,为了更好地评估问答对的生成质量,本文采用了人工评估的方式. 首先,我们从SQuAD[14]数据集的验证集中随机抽取200个样本,并将它们的上下文分别输入我们的模型IGJA-QAP以及OneStop模型中,记录下生成的问答对. 然后将生成的问答对分配给两个评测人员进行评价,他们需要根据问卷的每个问题对生成的问答对确定一个选择项. 如表4所示,本文通过对收集的问卷进行数据统计,以此来分析IGJA-QAP的生成质量. 与之前Liu[1]的人工评价工作类似,本文也是通过建立以下问卷,从不同的角度评价生成的质量.

表4 IGJA-QAP和OneStop模型在SQuAD数据上的人工评价结果

问题单	选项	OneStop	IGJA-QAP	Gold
语法结构是否完整	Yes	22.0%	41.5%	95.0%
	No	34.0%	12.5%	3.0%
	Understandable	44.0%	46.0%	2.0%
语义是否匹配	Yes	33.0%	69.5%	95.0%
	No	67.0%	30.5%	5%
表达是否多样	Yes	27.0%	47.5%	-
	No	73.0%	52.5%	-

注:其中Gold是指目标问答对

(1)问答对在语法结构上是否完整? 该问题是为了考查模型是否会生成语法或结构上不正确的问答对,检查模型的基本生成能力. 评测人员可以从三个选项中选择:Yes、No或Understandable. 当选择Yes时,表明生成的答案和问题很完整,没有语法或结构上的问题. 而当问题或者答案其中一个有严重的语法或结构错误,可以选择No. 当问答对其中一个在语法上不完全正确并且仍然可以推断出其含义时,评测人员可以选择Understandable.

(2)问答对在语义上是否匹配且与上下文密切相关? 生成的问答对如果在语义上缺乏相关性或者与上下文无关,表明模型缺乏理解能力,不能实现问题生成和答案生成两任务之间的跨任务信息交流,因此也无法生成具有相关性的问答对. 当评测人员

发现问题和答案之间相关性很强，并且能够紧密地贴切上下文信息，应该选择Yes. 如果上面两个条件没有同时满足，评测人员应该选择No.

(3)问答对是否同时满足语义匹配性、与上下文相关性且表述与目标问答对不同？本文在问题(2)上继续增加限定条件，从而考察模型在表达上的多样性生成，检查模型是否只进行了局部优化. 另外，这个问题也可以检查IGJA-QAP模型是否具有强大的生成和理解能力，即充分地理解上下文并生成不同于标准标签的问答对. 当生成的问答对相互匹配、与文章相关但在语法或语义与标准标签不同时，评测人员应选择Yes. 否则，应该选择No.

通过统计两位评测人员的问卷结果，得到了表4的数据结果. 从表中我们可以看到，由于数据处理过程中需要将SQuAD数据集中的段落拆分为只包含一个答案问题的句子，这样的切分方式就导致有些上下文缺失一些语义信息，目标问答对并不都是与上下文紧密相关的. 因此，表4中标准问答对并不完全匹配或者与上下文相关联. 此外，我们可以看到OneStop生成的问答对中有64.0%是完整的或可理解的，而IGJA-QAP模型得到了87.5的百分比. 表4中的比较结果表明，我们的模型在生成有意义和完整的句子方面更具优势. 当被问及问答对是否匹配时，表中两个模型之间所展现出的差距验证了本文的统一模型能够确保生成的问题答案对之间具有相同的潜在语义信息. 最后在回答生成的问答对是否多样或不同于目标问答对时，我们的模型在SQuAD数据集上达到了47.5%，生成的问答对大概一半都不同于标准问答对，而OneStop只有27.0%. 相对于OneStop，IGJA-QAP在问题(3)上接近20.5%的提升表明IGJA-QAP统一的框架相对于OneStop框架在训练时更注重整体的优化效果，增强模型的理解能力，避免陷入局部最优解. 综上对三个问题的分析，经过人工分析的结果可以表明IGJA-QAP模型可以生成高质量的、相匹配的、与上下文紧密相关且具有多样性的知识问答对.

4.6　消融实验

为了评估模型设计的合理性，研究本文提出的不同组件对整个模型性能的相对贡献，我们将本文提出的IGJA-QAP模型与以下三个变体模型进行对比实验.

(1) Ours-gate

该变体模型是从完整模型中移除答案引导的多头门机制. 即在生成问题时，经过解码器后的解码向量直接送到生成器中生成问题. 在这种情况下，没有答案引导多头门将答案信息融入到问题的生成，我们可以观察到在三个数据集上答案生成和问题生成的平均结果在METEOR指标上降低了8.4%和7.6%. 没有答案引导的多头门，就缺少问题生成和答案生成之间的信息引导. 另外，在训练时也缺少了问题对答案的约束，从而也就导致Ours-gate效果变差. 该对比实验现象可以充分说明答案的引导多头门可以改善答案生成和问题生成之间的交互.

(2) Ours-two-decoder

该变体模型是将问题生成和答案生成的解码过程分开，利用两个解码器分别解码，其他部分保持不变. 该模型由一个相同的编码器和两个单独的解码器构成，然后通过生成器和答案引导的多头门机制分别生成答案和问题. 值得注意的是，当我们的模型配备两个解码器通过共享编码器分别解码问题和答案时，答案生成任务和问题生成任务之间的交互只体现在编码阶段. 因此，如表5所示答案生成的性能下降了4.8%，问题生成的性能下降了1.1%. 消融实验结果为表明本文提出的统一框架有效地促进了两任务之间的信息交换以生成更加匹配的问答对提供了强有力的证据.

表5　消融实验结果

数据集	模型	BLEU-1		ROUGE-L		METEOR	
		问题生成	答案生成	问题生成	答案生成	问题生成	答案生成
SQuAD	Ours-gate	15.1	22.9	23.1	41.0	13.8	25.0
	Ours-two-decoder	37.8	22.9	41.1	41.2	27.1	25.4
	Ours-pointer	37.9	23.9	41.4	42.5	27.7	16.7
	IGJA-QAP	**38.4**	**25.8**	**41.6**	**56.1**	**28.2**	**33.0**
NewsQA	Ours-gate	17.5	25.3	39.5	42.6	15.2	40.5
	Ours-two-decoder	30.0	26.3	43.7	44.6	16.9	43.7
	Ours-pointer	27.5	26.5	43.8	43.9	17.3	45.4
	IGJA-QAP	**30.3**	**27.2**	**44.1**	**59.0**	**17.4**	**45.9**
CoQA	Ours-gate	9.6	17.7	38.6	47.6	10.1	17.1
	Ours-two-decoder	31.0	21.6	41.8	48.7	14.6	24.5
	Ours-pointer	11.8	21.9	38.6	47.9	12.9	23.2
	IGJA-QAP	**32.3**	**24.3**	**43.2**	**48.9**	**16.3**	**29.1**

(3) Ours-pointer

该变体在答案生成过程中去掉指针网络，从而探索其有效性. 我们可以观察到，移除指针网络会导致在三个数据集上答案生成的性能显著下降，在SQuAD数据集上METEOR值直接下降了16.3%.

这种下降是由于没有指针网络，模型无法从文档中复制单词，从而导致在答案生成任务上指标的下降.

4.7 案例研究

为了更好地说明我们模型的优越性，本文在表6中展示了本文模型IGJA-QAP和OneStop[4]两个模型的生成案例. 从案例中可以看出，我们的模型可以生成更准确、更易读、更匹配的问答对. 从第一个案例可以看出，OneStop生成的问题与原文是不相关的，即原文中没有“the largest financial endowment in Harvard?”相关联的信息. 除此之外，OneStop生成的答案也不匹配问题. 相比而言，我们的模型IGJA-QAP生成了一个匹配且与上下文相关的问答对. 在第二个案例中，我们可以观察到IGJA-QAP和OneStop都可以生成一个可读的问题. 但是OneStop模型抽取出来的答案在回答问题时显得不自然且不准确，而我们的模型IGJA-QAP模型生成的答案能够自然准确地回答生成的问题. 从上述案例中，IGJA-QAP模型可以产生匹配、与上下文密切相关且多样的问答对，生成的答案也能自然准确地回答生成的问题. 未来，我们将进一步探索如何更好地评估问答对的生成质量. 综上所述，这些案例可以表明IGJA-QAP模型具有很强的理解和生成能力.

表6 IGJA-QAP和OneStop的生成案例

上下文	Harvard's $37. 6 billion financial endowment is the largest of any academic institution	
标准问答对	问题：	What is the size of the school's endowment?
	答案：	$37. 6 billion.
OneStop	问题：	What is the largest financial endowment in Harvard?
	答案：	Billion.
IGJA-QAP	问题：	How much money is Harvard's financial endowment?
	答案：	$37. 6 billion financial endowment.
上下文	Tumor cells often have a reduced number of MHC class I molecules on their surface, thus avoiding detectionsby killer T cells	
标准问答对	问题：	What receptors do Tumor cells often have reduced concentrations of?
	答案：	MHC class I molecules.
OneStop	问题：	What is the number of MHC class I molecules on Tumor cells?
	答案：	A reduced number.
IGJA-QAP	问题：	What does Tumor cells have on their surface?
	答案：	A reduced number of MHC class I molecules on their surface.

5 总 结

本文提出了一种基于交互引导的问答对联合生成模型IGJA-QAP，用于从给定上下文中自动生成高质量的问答对. 联合式的生成框架使得问题生成和答案生成在训练时能够保持相同的上下文语义信息，从而通过联合训练实现两任务之间的信息交互. 此外，生成式的问题和答案获取方式可有效地降低在联合训练时两任务之间的难度差异，从而改善不平衡优化问题. 针对问答对生成任务所设计的答案引导的多头门机制可以使答案生成和问题生成两任务之间实现跨任务信息交流，从而增强两任务之间的相互优化. 在三个基准数据集上的大量实验表明，本文提出的模型优于先进的对比算法. 消融实验结果也说明了该模型中每个组件的有效性. 人工评估实验也证实了IGJA-QAP模型可以生成高质量且多样性的问答对.

参 考 文 献

[1] Liu B, Wei H, Niu D, et al. Asking questions the human way: scalable question-answer generation from text corpus// Proceedings of the Web Conference 2020. Taipei, China, 2020: 2032-2043

[2] Krishna K, Iyyer M. Generating question-answer hierarchies// Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy, 2019: 2321-2334

[3] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 7871-7880

[4] Du X, Cardie C. Harvesting paragraph-level question-answer pairs from wikipedia//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1907-1917

[5] Li Z, Wang W, Dong L, et al. Harvesting and refining

question-answer pairs for unsupervised QA//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online，2020：6719-6728

[6] Cui S，Bao X，Zu X，et al. OneStop QAmaker：extract question-answer pairs from text in a one-stop approach. arXiv preprint arXiv：2102.12128，2021

[7] Shinoda K，Aizawa A. Variational question-answer pair generation for machine reading comprehension. arXiv preprint arXiv：2004.03238，2020

[8] Shakeri S，dos Santos C，Zhu H，et al. End-to-end synthetic data generation for domain adaptation of question answering systems//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online，2020：5445-5460

[9] Lee D B，Lee S，Jeong W T.，et al. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online，2020：208-224

[10] Guo M，Haque A，Huang D A，et al. Dynamic task prioritization for multitask learning//Proceedings of the European Conference on Computer Vision (ECCV). Munich，Germany，2018：270-287

[11] Zhang W，Wang K，Wang Y，et al. A loss-balanced multi-task model for simultaneous detection and segmentation. Neurocomputing，2021，428：65-78

[12] See A，Liu P J，Manning C D. Get to the point：summarization with pointer-generator networks//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Vancouver，Canada，2017：1073-1083

[13] Zhao Yun，Liu De-Xi，Wan Chang-Xuan，et al. Retrieval-based automatic question answer：a literature survey. Chinese Journal of Computers，2021，44(6)：1214-1232(in Chinese)
(赵芸，刘德喜，万常选，等. 检索式自动问答研究综述. 计算机学报，2021，44(6)：1214-1232)

[14] Rajpurkar P，Zhang J，Lopyrev K，et al. Squad：100，000+ questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin，USA，2016：2383-2392

[15] Yu A W，Dohan D，Luong M，et al. Qanet：Combining local convolution with global self-attention for reading comprehension//Proceedings of the 6th International Conference on Learning Representations. Vancouver，Canada，2018：1-16

[16] Seo M J，Kembhavi A，Farhadi A，et al. Bidirectional attention flow for machine comprehension//Proceedings of the 5th International Conference on Learning Representations. Toulon，France，2017：1-13

[17] Nguyen T，Rosenberg M，Song X，et al. Msmarco：a human-generated machine reading comprehension dataset//Proceedings of the Workshop on Cognitive Computation：Integrating Neural and Symbolic Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016). Barcelona，Spain，2016

[18] Hsu C，Lind E，SoldainiL，et al. Answer generation for retrieval-based question answering systems// Proceedings of the ACL/IJCNLP 2021 Findings of the Association for Computational Linguistics. Online，2021：4276-4282

[19] Lan Y，Jiang J. Query graph generation for answering multihop complex questions from knowledge bases//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online，2020：969-974

[20] Baheti A，Ritter A，Small K. Fluent response generation for conversational question answering//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online，2020：191-207

[21] Mao Y，He P，Liu X，et al. Generation-augmented retrieval for open-domain question answering//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). Online，2021：4089-4100

[22] Dong L，Yang N，Wang W，et al.Unified Language Model Pre-training for natural language understanding and generation. arXiv preprint arXiv：1905.03197，2019

[23] Raffel C，Shazeer N，Roberts A，et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research，2020，21：1-67

[24] Menick J，Trebacz M，Mikulik V，et al. Teaching language models to support answers with verified quotes. arXiv preprint arXiv：2203.11147，2022

[25] Nakano R，Hilton J，Balaji S，et al. Webgpt：Browser assisted question-answering with human feedback. arXiv preprint arXiv：2112.09332，2021

[26] Bai Y，Jones A，Ndousse K，et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv：2204.05862，2022

[27] Ouyang L，Wu J，Jiang X，et al. Training language models to follow instructions with human feedback// Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022). OrleansNew，USA，2022：27730-27744

[28] Heilman M，Smith N A. Good question! statistical ranking for question generation// Proceedings of the Human Language Technologies：the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles，USA，2010：609- 617

[29] Heilman M. Automatic factual question generation from text[Ph. D. Thesis]. Carnegie Mellon University，USA，2011

[30] Chali Y，Hasan S A. Towards automatic topical question generation// Proceedings of the 24th International Conference on Computational Linguistics. Mumbai，India，2012：475-492

[31] Chali Y，Hasan S A. Towards topic-to-question generation. Computational Linguistics，2015，41(1)：1-20

[32] Hu W，Liu B，Ma J，et al. Aspect-based question generation// Proceedings of the 6th International Conference on Learning Representations. Vancouver，Canada，2018：1-10

[33] Serban I V，García-Durán A，Gülçehre Ç，et al. Generating factoid questions with recurrent neural networks：the 30m

factoid question-answer corpus//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 588-598

[34] Liu B, Zhao M, Niu D, et al. Learning to generate questions by learning what not to generate//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 1106-1118

[35] Zhou Q, Yang N, Wei F, et al. Neural question generation from text: a preliminary study//Proceedings of the 6th CCF International Conference on Natural Language Processing and Chinese Computing. Dalian, China, 2017: 662-671

[36] Kim Y, Lee H, Shin J, et al. Improving neural question generation using answer separation//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 6602-6609

[37] Zi K, Sun X, Cao Y, et al. Answer-focused and position-aware neural network for transfer learning in question generation//Proceedings of the 12th International Conference on Knowledge Science, Engineering and Management. Athens, Greece, 2019: 339-352

[38] Gao Y, Bing L, Chen W, et al. Difficulty controllable generation of reading comprehension questions//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 4968-4974

[39] Dong Xiao-Zheng, Hong Yu, Zhu Fen-Hong, et al. Question generation based on information features of token position. Journal of Chinese Information Processing, 2019, 33(8): 93-100(in Chinese)
(董孝政, 洪宇, 朱芬红, 等. 基于密令位置信息特征的问题生成. 中文信息学报, 2019, 33(8): 93-100)

[40] Yuan X, Wang T, Gülçehre Ç, et al. Machine comprehension by text-to-text neural question generation//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada, 2017: 15-25

[41] Du X, Cardie C. Identifying where to focus in reading comprehension for neural question generation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2067-2073

[42] Golub D, Huang P, He X, et al. Two-stage synthesis networks for transfer learning in machine comprehension//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 835-844

[43] Song L, Wang Z, Hamza W. A unified query-based generative model for question generation and question answering. arXiv preprint arXiv:1709.01058, 2017

[44] Cui S, Lian R, Jiang D, et al. Dal: Dual adversarial learning for dialogue generation. arXiv preprint arXiv:1906.09556, 2019

[45] Lyu C, Shang L, Graham Y, et al. Improving unsupervised question answering via summarization-Informed question generation//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 4134-4148

[46] Dugan L, Miltsakaki E, Upadhyay S, et al. A feasibility study of answer-agnostic question generation for education//Proceedings of the ACL 2022 Findings of the Association for Computational Linguistics. Dublin, Ireland, 2022: 1919-1926

[47] Back S, Kedia A, Chinthakindi S C, et al. Learning to generate questions by learning to recover answer-containing sentences//Proceedings of the ACL/IJCNLP 2021 Findings of the Association for Computational Linguistics. Online, 2021: 1516-1529

[48] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need// Proceedings of the Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA, 2017: 5998-6008

[49] Du X, Shao J, Cardie C. Learning to ask: Neural question generation for reading comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1342-1352

[50] Welleck S, Kulikov I, Kim J, et al. Consistency of a recurrent language model with respect to incomplete decoding//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online, 2020: 5553-5568

[51] Freitag M, Al-Onaizan Y. Beam search strategies for neural machine translation//Proceedings of the NMT@ACL 2017 First Workshop on Neural Machine Translation. Vancouver, Canada, 2017: 56-60

[52] Trischler A, Wang T, Yuan X, et al. Newsqa: A machine comprehension dataset//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada, 2017: 191-200

[53] Reddy S, Chen D, Manning C D. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7(3): 249-266

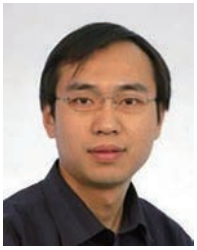
[54] Alberti C, Andor D, Pitler E, et al. Synthetic QA corpora generation with roundtrip consistency//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy, 2019: 6168-6173

[55] Devlin J, Chang M, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. Minneapolis, USA, 2019: 4171-4186

[56] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[57] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005. Ann Arbor, USA, 2005: 65-72

[58] Lin C Y. Rouge: A package for automatic evaluation of summaries// Proceedings of the Workshop on Text Summarization Branches Out, WAS 2004. Barcelona, Spain, 2004:74-81



LIU Jie, Ph.D., professor. His main research interests are natural language processing, data mining, and machine learning.

LIN Shao-Xin, M. S. candidate. His main research interest is natural language processing.

WANG Shan-Peng, B. S. His main research interest is natural language processing.

Background

Question-answer pairs (QAP) are essential for many applications, such as assisting the construction of knowledge bases, improving search engines by generating questions from documents, and training chatbots to make fluent conversations. However, obtaining question-answer pairs by human-annotation according to given documents is tedious and costly. Therefore, there is a significant need for efficient methods to automatically generate high-quality question-answer pairs from given documents.

Most existing works about generating question-answer pairs are mainly based on pipeline structure. Due to the independent learning process of question answering (QA) and question generation (QG), it is not only hard to make the best use of the semantic information, but also inclined to accumulate errors. Recently, some researchers propose an end-to-end framework that simultaneously accomplishes the subtasks of QA and QG. Despite the progress, there still exist important challenges in the QAP generation problem. Firstly, the extractive way of acquiring answers is insufficient for generating natural question-answer pairs, whether in the pipeline or end-to-end approaches. The extractive answer means truncating a span of consecutive words from the passage, which cannot comprehensively express complicated semantics in a human-like way. Secondly, the interaction between AE and QG, which are semantically related subtasks, is inadequate to mutually improve each other in the training process. The pipeline approaches directly ignore the relation between AE and QG. Thirdly, the difference in task difficulty can result in the imbalanced optimization of AE and QG.

To address these challenges, we propose a unified abstractive model (IGJA-QAP) with the multi-head answer-guided gate and the pointer network. Unlike the tranditional end-to-end method, we select an abstractive way of acquiring answers to offer better expressivity for complicated semantics. In addition, introducing the pointer network can allow the answer generation to copy words from the document, thus retaining the merit of AE. Furthermore, the same way of question-answer generation can improve the inequalities in task difficulty and reduce the imbalances between AE and QG. To ensure the common semantics to produce question-answer pairs, we propose to integrate the question-answer generation's decoding processes into the joint architecture. In this way, they can collaborate and benefit from each other to generate compatible and high-quality question-answer pairs. Therefore, the unified model can bring mutual optimization for question-answer generation, avoiding a scenario where generated answers and questions are weakly related. This setting, in turn, contributes to the improvement of imbalanced optimization. Besides, some researches propose that when a human obtains question-answer pairs based on a passage, an answer is considered a clue to guide the question generation. Accordingly, we utilize a multi-head answer-guided gate to transfer the cross-task information from the answer generation to the question generation.