计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200153

一种会话理解模型的问题生成方法

时雨涛　孙　晓
合肥工业大学计算机与信息学院　合肥 230601
合肥工业大学情感计算与先进智能机器安徽省重点实验室　合肥 230601
(2019110984@mail.hfut.edu.cn)

摘　要　会话问题生成(Conversational Question Generation,CQG)不同于根据段落和答案生成单轮问题的问题生成任务,CQG额外考虑由历史问答对构成的会话信息,生成的问题承接会话历史内容,保持较高的一致性。针对这一特性,文中提出了字级别和句级别注意力机制模块来增强对会话历史信息的提取能力,确保当前轮次的问题融合会话历史中每个词和句子的特征,从而生成连贯的、高质量的问题。疑问词的正确性较重要,生成的问题需要和数据集中原始问题对应的答案类型相互匹配,在疑问词预测模块中构造额外的损失函数作为疑问词类型的限制。综合各个模块得到会话理解模型(Conversational Comprehension Network,CCNet),实验结果表明,该模型在大部分评测指标上高于基线模型,在CoQA数据集上Bleu1和Bleu2分别达到39.70和23.76,生成的问题质量更高。在消融实验和跨数据集实验中该模型被证明是有效的,说明CCNet模型具有较强的通用能力。

关键词:问题生成;注意力机制;会话问题生成;循环神经网络;门控网络
中图法分类号　TP391

Conversational Comprehension Model for Question Generation

SHI Yu-tao and SUN Xiao
School of Computer and Information,Hefei University of Technology,Hefei 230601,China
Key Laboratory of Affective Computing and Advanced Intelligent Machines of Anhui Province,Hefei University of Technology,Hefei 230601,China

Abstract　Conversational question generation (CQG) is different from the question generation task of generating single-round questions based on paragraphs and answers. CQG additionally considers the conversational information composed of historical question and answer pairs,and the generated questions inherit the historical content of the conversation and maintain high consistency. In response to this feature,the article proposes word-level and sentence-level attention mechanism modules to enhance the ability to extract conversation history information,ensuring that the current round of questions integrates the characteristics of each word and sentence in the conversation history,thereby generating a coherent,high-quality question. The accuracy of the question word is more important. The generated question needs to match the answer type corresponding to the original question in the data set. An additional loss function is constructed in the question word prediction module as a limitation of the question word type. The conversational comprehension network (CCNet) model is obtained by synthesizing each module. Experiments show that this model is higher than the baseline model in most evaluation indicators. On the CoQA dataset,Bleu1 and Bleu2 reach 39.70 and 23.76,respectively,and the quality of the generated questions is higher. The model is proved to be effective in ablation experiments and cross-dataset experiments,indicating that the CCNet model has strong general capabilities.

Keywords　Question generation,Attention mechanism,Conversational question generation,Recurrent neural network,Gated network

1　引言

问题生成任务属于自动问答领域中的一个小分支,近些年自动问题生成任务的目标主要是在给定类型的数据源的基础上产生问题,数据源包括结构化的知识[1-2]和非结构化的文本[3-4],如长文本、问题对应的答案以及涵盖常识信息的外部知识等。在教育教学领域中,问题生成通过生成问题的方式自动评估学生的情况,有助于老师及时掌握学生的理解分析

到稿日期:2021-02-24　返修日期:2021-06-13
基金项目:国家自然科学基金(61976078)
This work was supported by the National Natural Science Foundation of China(61976078).
通信作者:孙晓(sunx@hfut.edu.cn)

能力；在对话系统中，有意义的问题可以提高人机交互的友好性和连贯性[5]。

尽管问题生成任务在实际中应用广泛，但是基于先前对话信息的问题生成在连贯性和一致性上较差。由此文献[6]提出了一个用于会话问题生成(CQG)的框架。该系统能够提出一系列以问答式对话的段落为基础的问题，在第一轮后的每一轮问题的生成都依赖于之前的历史问答信息，本文将若干个问答对组成的信息称为会话历史。为了获得训练系统的数据，需要将高质量大规模的会话式问答任务中的 CoQA[7] 和 QuAC[8] 等转换成适应 CQG 的数据集。文献[9]中提到 CQG 可以嵌入智能社交机器人或辅导系统中，衡量机器引导问答式对话的能力，提出有意义且连贯的问题来吸引用户，测试学生对某个主题的理解程度。

但是该框架存在一定的缺陷，模型输入中与答案相关的信息较少，因此我们额外考虑数据集中与当前问题相关的推理语句，从而生成与答案相关的问题。除此之外，我们还增加段落中每个单词的特征，通过 StanfordNLP 工具进行词性标注(pos_tag)和命名实体识别(NER)，丰富单词的属性特征，提高单词应用准确性。同时我们考虑到问题中疑问词的重要性，比如疑问词是“when”时就不会出现关于时间之外的答复，因此疑问词预测模块有利于精准匹配问题的类型及答案。原框架对历史问答信息的利用不充分，缺乏对该信息中问题和答案的理解，为了探究历史问答中单词和句子与问题之间的关联，我们在模型中综合词注意力机制和语句注意力机制，生成一个连贯的对话。本文提出了一个基于会话理解方法的新框架。综上所述，本文的主要贡献有以下 4 点：

(1)对段落的每个单词增加词性标注特征和命名实体识别特征，丰富单词的属性，使得单词具有更深层次的含义。

(2)模型额外加入推理语句，通过 co-attention 机制强化段落特征和会话历史特征，突出原始信息中和答案相关的部分。

(3)针对解码器输出的每个结果，将其与会话历史进行字级别和语句级别的注意力机制处理，从而强化解码器输出的特征。

(4)设计了疑问词预测模块，使得问题中生成的疑问词的准确率更高，问题和答案之间保持一致性。

2 相关工作

2.1 问题生成

根据给定的多源文本输入产生问题是问题生成(QG)任务的主要目的，早期大多数问题生成的工作都是规则化或者模板化的方法[10-11]，其规则繁琐且耗费人力。文献[12]提出了 seq2seq 网络模型架构，在文本生成、机器翻译等领域取得了很好的效果并得到广泛应用。文献[13]引入注意力机制来解决长距离文本依赖性不强的问题。文献[14]在之前的基础上提出加入注意力机制的 seq2seq 模型来解决问题生成任务。文献[15-16]将答案作为相对于段落的额外输入，以保证信息更加完整。文献[17]提出根据预设定的问题模板生成问题，但问题形式有局限性。

2.2 对话生成

多轮对话生成任务是对话领域中的小分支，其能够根据已知的对话内容生成回复。文献[18]提出了一种层次化的、递归式的序列到序列模型 HRED 来为上下文的句子建模。文献[19]在之前模型的基础上优化注意力机制的原理，其注意力权重是通过词注意力机制和句注意力机制计算得到的。

2.3 会话问题生成

通过输入段落和会话历史生成问题的任务受到很多研究者的注意，其中会话历史由历史问答对构成，该任务被称为会话式问题生成。文献[6]针对此任务首先提出了一个关于会话问题生成的框架，生成的问题能够贴近于段落信息，更加平滑地衔接上文。文献[9]提出了嵌入推理模块的 ReDR 模型，采用强化学习机制将答案的质量作为反馈微调模型。

会话问题生成将会话历史作为输入，但是之前的研究工作中对该信息的利用和挖掘有所欠缺，因此我们在受到多轮对话 HRAN 模型的启发后提出 CCNet 模型。

3 任务设定

根据 CQG 的任务设定输入信息包括会话历史 $CH_i=\{(q_1,a_1),\cdots,(q_{i-1},a_{i-1})\}$、段落 P_i 以及当前轮次问题 q_i 对应的推理片段 ST_i，推理片段表示段落中与当前轮次问答有关的文本内容。通过式(1)在当前第 i 轮次产生问题 q_i：

$$q_i=\arg\max \Pr ob(q_i\mid P_i,CH_i,ST_i) \tag{1}$$

其中，$\Pr ob(q_i\mid P_i,CH_i,ST_i)$ 表示生成当前问题的条件概率。

4 模型描述

本文的模型框架包括 3 个主要部分：1)如图 1 所示的多源输入编码器，文献[6]中针对推理片段的处理是标记其在段落中对应的单词，我们为了突出推理部分的重要性直接将推理语句作为模型的输入；2)融合 copy 机制和 coverage 机制的解码器，如图 2 所示；3)疑问词预测模块。

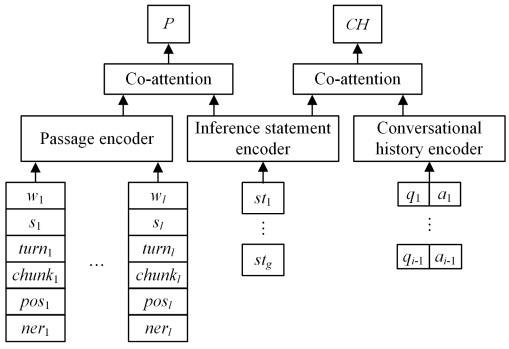

图 1 编码器

Fig. 1 Encoder

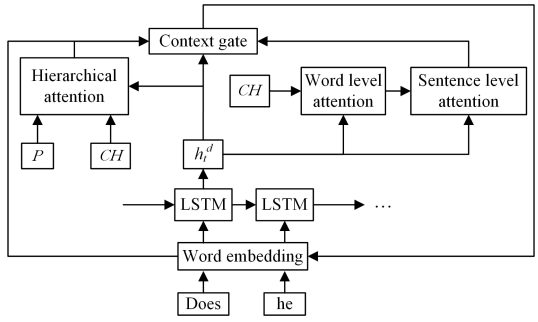

图 2 解码器

Fig. 2 Decoder

4.1 编码器

4.1.1 段落编码器

首先将段落分词，表示为$\{p_1, p_2, \cdots, p_l\}$，其中$l$表示段落信息的长度。取$p_i=[\boldsymbol{w}_i; s_i; turn_i; chunk_i; pos_i; ner_i]$，$\boldsymbol{w}_i$表示单词对应的预训练词向量；$s_i$表示对推理片段的BIO标记，使用标识符匹配其对应在段落中的位置，B_ANS和I_ANS标记开始和结束的单词，O对应其他与答案不相关的单词；$turn_i$表示当前生成问题步骤对应的轮次；段落分成若干块，每个块中的单词数量相近，$chunk_i$表示单词在段落中对应的块下标；pos_i表示单词对应的词性；ner_i表示单词对应的实体标注信息。

段落编码器采用双向LSTM(Bi-LSTM)[19]，利用隐藏状态$P=(h_1^p, \cdots, h_l^p) \in R^{d\times l}$表示段落特征，$d$表示隐藏层的维度。

4.1.2 会话历史编码器

包含前若干轮问答对的会话历史表示为$\{(q_1, a_1), \cdots, (q_{i-1}, a_{i-1})\}$。首先将每个问答对输入Bi-LSTM，生成词级别的隐藏特征$CH_{v-b}^w=(h_{v-b,1}, \cdots, h_{v-b,n}) \in R^{d\times n}$，其中$v-b$表示问答对在会话历史中的序号，$b\in[1, v)$，$n$表示问答对的长度。然后将字级别特征$CH^w$输入另一个Bi-LSTM中，得到句子级别的隐藏特征$CH^s=(h_1, \cdots, h_t) \in R^{d\times v}$。

4.1.3 推理语句编码器

推理语句表示问答过程在原段落中的依据，其作为额外信息输入模型。将预训练的词向量$\{st_1, st_2, \cdots, st_g\}$输入一个Bi-LSTM网络得到隐藏特征$ST=(h_1^{st}, \cdots, h_g^{st}) \in R^{d\times g}$，其中$g$表示推理片段的长度。我们使用$ST$对$P$和$CH^s$进行注意力机制加权，增强这两个特征中关于当前轮问答的信息。在之前的机器翻译工作中co-attention机制应用广泛[20-21]，受该方法的启发我们设计了强化模块。首先计算出ST和段落特征P的关联矩阵：

$$\boldsymbol{SP}=ST^{\mathrm{T}}(\boldsymbol{W}_{sp}P) \in \boldsymbol{R}^{g\times l} \tag{2}$$

其中，$\boldsymbol{W}_{sp} \in \boldsymbol{R}^{d\times d}$表示参数矩阵。为了深入得到推理语句与段落中每个词的相似度，我们对关联矩阵$\boldsymbol{SP}$进行列级别的归一化处理。然后将$\boldsymbol{SP}$与ST相乘并对结果进行加权处理得到强化后的段落特征。具体公式如下：

$$P=ST \cdot softmax(\boldsymbol{SP})+P \in \boldsymbol{R}^{d\times l} \tag{3}$$

同理在处理ST和CH^s两个特征时，运用相同的方法得到强化后的特征。

4.2 解码器

解码器使用层次化的LSTM迭代来获得隐藏特征$(h_1^d, \cdots, h_t^d)$，计算公式如下：

$$h_{t+1}^d=LSTM(h_t^d, w_t) \tag{4}$$

下一时间步骤的隐藏状态h_{t+1}^d通过当前状态h_t^d和词向量w_t得到。为了保证生成的问题能够综合段落和会话历史两个特征，编码器需要考虑问题中每个词与这两个特征的关联度，求解时间步t的隐藏状态与联合特征(P, CH^w, CH^s)的相似度矩阵，然后使用相似度矩阵计算段落注意力权重$\alpha=(\alpha_1, \cdots, \alpha_l)$和会话历史注意力权重$\beta=(\beta_{1,1}, \cdots, \beta_{1,m}; \beta_{i-1,1}, \cdots, \beta_{i-1,n})$，最后得到向量$\boldsymbol{c}_t$。

$$\boldsymbol{c}_t=\alpha P+\beta CH^w \tag{5}$$

上述操作的具体实现细节可从文献[6]详细了解。CQG需要强化的点主要是生成的问题，能够承接会话历史，满足逻辑性和合理性，与其语义和句式结构保持一致，产生质量较高的问题，使得会话能够顺利地进行下去。因此我们构建字级别和词级别的注意力机制，以达到平滑过渡多轮次会话的目的。

4.2.1 词层次注意力机制

考虑到每个时间步骤会话历史中每个词对生成单词的影响，在时间步骤t时，我们将h_t^d和会话历史特征CH^w进行交互处理。h_t^d与轮次k对应的$CH_k^w=(h_{k,1}, \cdots, h_{k,n})$作为词级别的注意力机制处理，公式如下：

$$wu_k=softmax(\boldsymbol{W}_1(\tanh[h_t^d; CH_k^w])) \tag{6}$$

$$histext_k=(CH_k^w)^{\mathrm{T}}wu_k \tag{7}$$

其中，$\boldsymbol{W}_1$是可学习的矩阵。综合会话历史中所有单词特征得到字级别的强化特征$histext=[histext_1, \cdots, histext_k]$，$k\in[1, v-1]$。

4.2.2 句层次的注意力机制

在进行字级别的注意力机制操作之后，我们考虑继续增强会话历史中每句话与隐藏状态h_t^d之间的关联。然后对$histext$和h_t^d进行句子级别的注意力机制处理，公式如下：

$$wp=softmax(\boldsymbol{W}_2(\tanh[h_t^d; histext])) \tag{8}$$

$$his_t=(histext)^{\mathrm{T}}wp \tag{9}$$

其中，$\boldsymbol{W}_2$是可学习的参数矩阵。最后得到在时间步骤t下隐藏状态h_t^d关于会话历史的融合特征his_t。

4.2.3 门控网络

为了能够对原始特征和强化后特征进行筛选，我们使用一个门控向量$\boldsymbol{z}$产生融合特征作为时间步骤t下的最终结果out_t，公式如下：

$$\boldsymbol{z}=sigmoid(\boldsymbol{W}_z[w_t; h_t^d; \boldsymbol{c}_t; his_t]) \tag{10}$$

$$source=\boldsymbol{W}_{source}[\boldsymbol{c}_t; his_t] \tag{11}$$

$$aim=\boldsymbol{W}_{aim}[\boldsymbol{w}_t; h_t^d] \tag{12}$$

$$out_t=\tanh((1-\boldsymbol{z})\odot aim+\boldsymbol{z}\odot source) \tag{13}$$

其中，$\boldsymbol{W}_z$，$\boldsymbol{W}_{source}$，$\boldsymbol{W}_{aim}$都是可学习的参数矩阵，$\odot$表示点乘操作。

copy机制[22]被广泛应用于自然语言处理领域中，在产生的问题中考虑了单词出现频率过低，没有出现在结果中的情况。我们同样应用这一机制，具体公式如下：

$$P(y_t \mid y_{<t}, CH, P, ST)=(1-\lambda)P_{gen}(y_t)+\lambda P_{pt}(y_t) \tag{14}$$

其中，$P_{gen}(y_t)=softmax(MLP(out_t))$，$P_{pt}(y_t)$表示单词$y_t$出现在段落$P$和会话历史$CH$的概率分布$[\alpha; \beta]$，$\lambda=sigmoid(\boldsymbol{W}_\lambda^{\mathrm{T}}out_t)$，$\boldsymbol{W}_\lambda^{\mathrm{T}}$表示可学习的矩阵。最终得到在时间步骤$t$下生成的单词分布概率，从而对应到词表中具体的单词。

4.2.4 疑问词预测模块

因为生成的问题质量高低在很大程度上取决于疑问词是否准确，所以我们提出疑问词判断机制，加入额外的损失函数来提高疑问词预测的准确性。

由于数据集中疑问词的种类繁多，因此我们根据其占有的比例进行筛选。本次统计选取11种比重值大于或等于0.01的疑问词，另外将其他所有的疑问词作为一个类别，最终得到12个类别。我们沿用之前得到的out_t特征，经过线性

层转换维度得到 $qtype$，公式如下：

$$qtype=\boldsymbol{W}_q out_1 \tag{15}$$

其中，$\boldsymbol{W}_q$ 是一个可学习的矩阵。由于疑问词一般都放在问题的首位，因此本次操作只针对解码层中第一个 LSTM 单元的输出。通过交叉熵计算出 $loss_q$ 作为对疑问词的限制。

4.3 损失函数

我们在实验中主要考虑了两个损失函数，将其汇总得到：

$$Loss=loss_{nll}+loss_q \tag{16}$$

其中，$loss_{nll}=-\log Prob(q_i \mid P_i, CH_i, ST_i)$ 表示预测出的文本序列的损失函数[12]。

5 实验

5.1 数据集

对比实验采用的数据集是包含历史问答对的 CoQA，该数据集针对一个段落有多轮的问答数据流，比较适合构成会话历史，我们使用当前问题的前若干轮问答对作为历史信息。在此数据集的基础上，实施训练和验证过程，将我们的实验结果和基线模型结果进行比较与分析。CoQA 收集 8000 多个段落，围绕每个段落有若干个问答，总共有 127000 个问答对。具体实例如图 3 所示。

My Left Foot (1989) Imagine _ , unable to make any movements except to move your left foot. The main character in My Left Foot,based on the real story of cerebral palsy sufferer Christy Brown,can barely move his mouth to speak, but by controlling his left foot, he's able to express himself as an artist and poet. For his moving performance of Brown, Daniel Lewis won his first Academy Award for best actor. Shine (1996) Do you have a talent you're afraid to share with the world? David Helfgott seemed meant from childhood to be "one of the truly great pianists," but the pressures of performing (and pleasing his father) resulted in a complete breakdown. Ten years in a mental hospital didn't weaken Helfgott's musical gift: When he was rediscovered, he was playing concertos in a bar. Shine received seven Oscar nominations , and Geoffrey Rush won best actor for his performance of Helfgott…

Q1：Who is My Left Foot based on?
A1：Christy Brown
R1：Christy Brown
Q2：What illness did he have?
A2：cerebral palsy
R2：cerebral palsy
Q3：Could he talk?
A3：barely
R3：can barely move his mouth to speak
Q4：How did he communicate?
A4：by controlling his left foot
R4：by controlling his left foot
Q5：Which pianist had a break down?
A5：David Helfgott
R5：David Helfgott
Q6：Did he recover quickly?
A6：No
R6：Ten years in a mental hospital

图 3 CoQA 数据集示例

Fig. 3 Examples of CoQA datasets

正如图 3 的数据集实例所示，围绕着一段文章有 6 个问答对，并且每个问答对都在一定程度上承接之前的问题。其中 R 表示答案在段落中相关的内容区域，一般认为是推理语句。

为了进一步验证模型的合理性和适用性，我们在另一个包含历史问答对的数据集 QuAC[8] 上进行大量实验。两个数据集的处理方法基本一致，使用 StanfordCoreNLP 工具对数据集中的文本信息进行预处理，段落分块处理，会话历史包含多轮问答对，对于文本具体的处理将在下节描述。

5.2 实现细节

本次数据处理中问答历史含有前两轮问答，文献[7]分析了该数据集中每两轮问答范围有比较大的依赖性，因此为了能够更充分地利用数据集并且得到比较好的实验结果，会话历史中问答对的数量确定为 3。

由于文章段落信息过于冗长，并且每轮都需要将整个段落输入模型。因此为了能够对段落中的每个词有所区分，我们采用分块的方法将段落分为若干个块，保证信息输入的合理性。首先统计段落的单词数量，以每个块中单词数量基本相同的标准进行分块，并标注单词当前所在的块数，后续的对比实验会比较不同块数产生的影响。标注段落中对应的推理部分，实现方法是增加段落信息的特征，标记推理语句开始和结束的单词，突出这些单词在当前轮次的重要性。

每个单词的词性标注和命名实体识别特征都是通过 StanfordNLP 工具提取得到的。词性标注用于分析当前单词的词性，命名实体识别可以识别出段落信息中特定类型的命名实体。而增加对单词的文本分析，有助于丰富单词所具有的涵义和属性。

模型参数设置方面，生成的问题最大长度设置为 15，beamsize 设置为 3。优化器采用的是 adagrad 优化方法，初始学习率设置为 0.1，编码器中 LSTM 单元的层数为 2，解码器 LSTM 单元的层数为 1，每个单元的隐藏层维度都设置为 512。所有单词都是通过 GloVe[23] 中 300 维词向量初始化的。在训练过程中词向量矩阵保持不变，不参与训练过程。

5.3 基线和消融实验

本次实验选择了 6 个基线模型：PGNet[22] 通过连接段落、对话历史和答案 3 个信息作为模型的输入；NQG[4] 与上一个模型类似，不同的地方是其将答案连接词向量后面输入模型；文献[6]中的 CFNet 是目前效果比较好的 CQG 系统，考虑了共指信息和会话流信息；MSNet 表示去除共指和会话流模块的模型；CorefNet 表示去除会话流模块的模型；FlowNet 表示去除共指模块的模型。

6 结果与分析

6.1 结果展示

对于模型结果评估的标准，我们使用问题生成任务中使用较多的指标，如 BLEU(1-4)[24]，METEOR(MET)[25]，ROUGE-L(R-L)[26] 来评测生成的问题与原问题的相关性。实验结果如表 1 所列。

表 1 CoQA 数据集上的对比实验

Table 1 Comparative experiments on CoQA dataset

Model	B1	B2	B3	B4	MET	R-L
PGNet	28.84	13.74	8.16	—	—	39.18
NQG	35.56	21.14	14.84	—	—	45.58
MSNet	36.35	21.82	15.44	11.65	16.50	46.01
CorefNet	36.67	22.26	15.78	11.93	16.66	46.4
FlowNet	36.60	22.23	15.77	11.96	16.70	46.52
CFNet	37.49	22.86	16.24	**12.26**	16.98	**46.86**
CCNet	**39.70**	**23.76**	**16.41**	12.04	**17.16**	46.79

表 1 列出了多个模型的主要结果，所有的实验结果都是针对 CoQA 数据集进行实验的，其中 CCNet 模型对应的段落

块数为5。由表1的结果可以得出，本文模型CCNet相较于CFNet模型，在B1，B2，B3，MET指标上均有所提高，而在B4和R-L指标上低于CFNet模型。具体模型中的模块效果还有待验证，在后文将单独讨论会话注意力机制模块和疑问词预测模块的重要性。

考虑到段落序列较长，单词数量较多，为了测试段落信息分块个数的影响，我们进一步改变段落的分块数量，实验结果如表2所列。

表2 段落块数分析

Table 2 Analysis on the number of paragraphs

num	B1	B2	B3	B4	MET	R-L
10	39.43	23.58	16.3	11.91	16.90	46.73
5	**39.70**	**23.76**	**16.41**	**12.04**	**17.16**	**46.79**
1	39.15	23.19	15.88	11.52	16.93	46.48

根据表2数据得出，当段落分块数为5时，模型的效果最好，所有指标都是最高的。我们从中分析得到，因为单个段落整体单词数量较多，分块处理后段落信息得到分解，段落中每个单词具有块特征，所以模型能够更好地学习到生成当前轮次的问题时应该集中在段落的哪一个块。然而块数过多的情况下，信息反而更加零散，因此最终将段落分为5块。

6.2 消融实验

在整体模型相较于基线模型有所提升的前提下，我们进一步讨论模型中的历史会话注意力机制模块和疑问词预测模块的重要性。我们设计消融实验来测试模块的有效性和可行性，BaseNet表示去除两个模块的基础模型，QTNet表示加入疑问词预测模块的模型，WSNet表示在BaseNet基础上加入字级别和句级别注意力机制的模型，实验设置中将段落的块数固定为5，实验结果如表3所列。

表3 消融实验

Table 3 Ablation experiment

Model	B1	B2	B3	B4	MET	R-L
BaseNet	38.23	22.82	15.60	11.33	16.81	46.13
QTNet	39.17	23.22	15.88	11.42	16.75	46.34
WSNet	39.22	23.25	15.91	11.54	16.89	46.4
CCNet	**39.70**	**23.76**	**16.41**	**12.04**	**17.16**	**46.79**

我们测试了QTNet和WSNet两个模型，发现它们相对于基线模型BaseNet都有不同程度的提升。

6.2.1 会话注意力机制模块的分析

WSNet模型是为了调查会话注意力机制模块的效果，其在整体模型上去除了疑问词预测模块，从表3中可以看出其在各个指标上都比BaseNet模型高。

为了进一步分析该模块的效果，如图4所示，我们通过每个词对应的注意力权重值来讨论模型的有效性，段落中的划线句子表示答案和推理语句。我们主要专注于问题的后半部分，首先得出they和look与每个历史问答对中每个单词的注意力值，然后乘以该问答对的句级别注意力值，最后将两个单词对应的权重值相加得到整体的注意力权重，结果显示于图4中。可以看到，在产生当前问题时，会话注意力机制分配更高的注意力权重值到第二、第三个问答对，“he”“wear”与“they”等与推理部分相关的单词注意力权重显著提高。这解释了为什么模型能够理解会话历史中的内容，并且生成问题“How did they look?”。

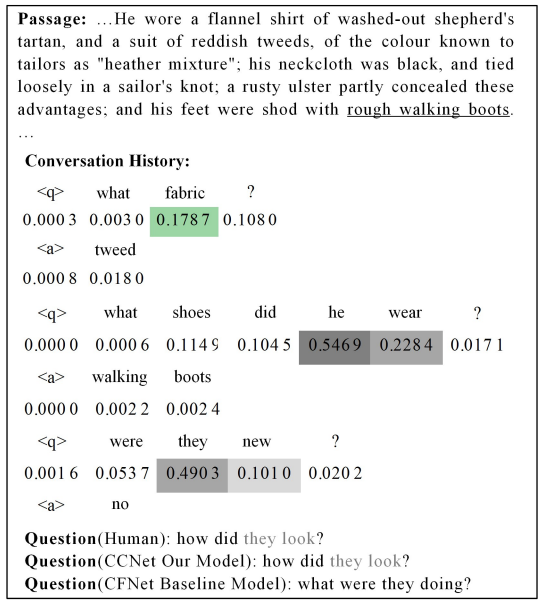

图4 注意力权重可视化

Fig. 4 Visualization of attention weight

6.2.2 疑问词预测模块的分析

表3中的QTNet模型在整体模型CCNet的基础上去除了会话注意力机制模块，并且对模型进行了微调，舍去字级别和句级别注意力机制，修改门控网络的输入和线性层的维度，以保证模型完整性。表3中的实验结果显示，QTNet模型的指标相较于BaseNet模型有所提高，由此说明，提高生成的疑问词的准确性能够产生更加贴近于原问题的问题。问题疑问词的重要性直接决定了问题的形式和内容，问题的形式限制着答案的回复，因此通过疑问词预测在一定程度上能够提高问题与给定输入信息（段落、会话历史、推理片段）的相关性。疑问词预测实验结果如表4所列。

表4 疑问词结果分析

Table 4 Result analysis of interrogative words

Model	Precision	Recall	F1	Accuracy
CFNet	40.61	39.07	38.03	56.40
CCNet	**42.55**	**39.73**	**38.79**	**58.84**

我们的评测标准选用常见的分类模型评价指标：精确率、召回率、F1、准确率。疑问词分类是多分类问题并且数据集样本数量足够，因此我们采用Macro-F1。从表4中可以看出，加入疑问词预测模块的CCNet性能优于CFNet，这说明该模块能够提高疑问词的准确性，从而生成和答案类型对应的问题，以便于后续轮次的问答。大部分疑问词都是单个词，但少部分由多个词构成疑问词的形式还有待考虑。

6.3 错误分析

经过两个模块的讨论，我们进一步分析CCNet模型的缺点，提出继续改进的方法和策略。

该模型中的字级别和句级别注意力机制能够保证问题的逻辑性和相关性，但是因为会话历史的复杂性，导致不能对其完全理解，额外的信息可能会干扰注意力权重值的分配，从而产生质量较差的问题。后续考虑增加当前轮次生成问题的

主题信息，以辅助该模型快速定位到关键单词的位置，为上下文有逻辑性地建模。如图5所示，会话历史中层层递进地讨论他女儿的失踪细节，我们的CCNet模型受到段落中划线部分推理语句的影响，专注于在2010年发生的事情，而没有很好地承接会话中对失踪信息的细节询问。

Passage: ..."Finding Aubrey" includes 11 songs written and performed by Sacco at his home studio, as well as the last three songs Aubrey herself recorded at home before the 23-year-old disappeared in April 2010 while hiking alone in Nepal.....

Conversation History:
<q>what's his daughter's name?
<a>Aubrey Sacco.
<q>what happened to her?
<a>disappeared.
<q>where?
<a>Nepal

Question(Human): what type of trip was this?
Question(CCNet Our Model): what happened in 2010?
Answer : hiking.

图5 实例1

Fig. 5 Case 1

疑问词预测模块仍然存在一些不足，如图6所示，即使疑问词是正确的，也不能保证预测出的问题与原问题的含义是相同的。很明显两个问题之间的时态不一致，虽然都是提问他在做什么事，但是Question(Human)拥有更深层的意义，更强调动作的延续性。因此只考虑疑问词来规范问题是有所欠缺的，后续需要对其进行进一步改进和提升，提高问题的多样性和复杂性。

Passage: ...One day a handsome young man called Narcissus came into the woods. He had been hunting deer and lost his way. However, the moment Echo saw him, she fell in love with him....
Conversation History:
<q>what could she say?
<a>she could only repeat the last words of those around her.
<q>where did she hide?
<a>deep in the woods.
<q>who came into the woods?
<a>Narcissus.

Question(Human): what had he been doing?
Question(CCNet Our Model): what was he doing?
Answer : hunting deer.

图6 实例2

Fig. 6 Case 2

6.4 跨数据集实验

通过上述对CoQA数据集的一系列实验，我们能够发现CCNet模型相较于基线模型来说，在大部分指标上都有所提高。其在生成新的问题时，能够综合历史问答中的单词和句子对问题中每个单词的影响，提高生成的疑问词的准确性，从而生成更贴近于对话、质量更高的问题。

为了验证CCNet模型的普适性，我们在另一个存在历史问答的数据集QuAC上进一步开展实验，以证明模型的稳定性和模型中模块的重要性。对于QuAC数据集的处理类似于CoQA数据集，将段落信息分块后，提取段落中每个单词的特征，取前两轮问答作为历史会话，将答案作为推理语句输入模型。两个数据集在疑问词预测方面有所不同，因为两个数据集对应的段落信息不同，所以对于疑问词的统计也会有差别。我们按照上述消融实验的步骤，消融会话注意力机制模块和疑问词预测模块，将其他参数做相同设置，实验结果如表5所列。

表5 QuAC数据集上的对比实验

Table 5 Comparative experiments on QuAC dataset

Model	B1	B2	B3	B4	MET	R-L
BaseNet	32.60	19.52	14.32	11.81	13.83	33.97
QTNet	32.76	19.66	14.51	11.99	14.04	34.06
WSNet	33.20	20.01	14.71	12.18	14.25	34.50
CCNet	**33.43**	**20.40**	**15.07**	**12.51**	**14.30**	**34.81**

表5的结果显示，QTNet模型和WSNet模型相对于BaseNet模型来说，指标有所提升，证明了我们提出的疑问词预测模块和会话注意力机制模块是有效的。两个数据集的生成结果有一定的差异，这是因为QuAC数据集每个批次的文本序列都相对较长，长文本情况下注意力机制的应用效果较差，所以需要继续改进注意力机制模块。

结束语 我们研究的方向是conversational question generation，这个方向是由文献[6]提出的，这个问题生成任务与问题生成主要的不同点在于，生成新问题时不仅仅考虑了段落信息，还将前若干轮次的问答对作为历史信息输入，使生成的问题能够从原文中找出线索，并且能够承接前几轮的对话内容，因而生成问题的质量更高。

本文提出了能够生成囊括历史信息并且有准确疑问词的问题CCNet模型。针对会话历史进行改进，采用字级别和词级别注意力机制，在生成问题时，达到与会话历史信息有比较高的关联性的目的。通过增加损失函数来约束疑问词的种类，从而生成与答案类型对应的问题。本文通过多种对比实验证实了本文所提方法是有效的，能够生成质量较高的问题。

在未来的工作中，关于疑问词预测除了要考虑单个疑问词外，还可以进行问题相关词推断的研究，以及考虑可能存在疑问词不同但是问题含义相同的情况。

参 考 文 献

[1] SERBAN I V, GARCIA-DURAN A, GULCEHRE C, et al. Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 588-598.

[2] GUO D, SUN Y, TANG D, et al. Question Generation from SQL Queries Improves Neural Semantic Parsing[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1597-1607.

[3] DU X, SHAO J, CARDIE C. Learning to Ask: Neural Question Generation for Reading Comprehension[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1342-1352.

[4] DU X, CARDIE C. Harvesting Paragraph-level Question-Answer Pairs from Wikipedia[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 1907-1917.

[5] WANG J, LIU J, BI W, et al. Dual Dynamic Memory Network for End-to-End Multi-turn Task-oriented Dialog Systems[C]// Proceedings of the 28th International Conference on Computa-

tional Linguistics. 2020:4100-4110.
[6] GAO Y, LI P, KING I, et al. Interconnected Question Generation with Coreference Alignment and Conversation Flow Modeling [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4853-4862.
[7] REDDY S, CHEN D, MANNINGC D. CoQA: A Conversational Question Answering Challenge[J]. Transactions of the Association for Computational Linguistics, 2019, 7:249-266.
[8] CHOI E, HE H, IYYER M, et al. QuAC: Question Answering in Context[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:2174-2184.
[9] PAN B, LI H, YAO Z, et al. Reinforced Dynamic Reasoning for Conversational Question Generation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2114-2124.
[10] VANDERWENDE L. Answering and Questioning for Machine Reading[C]// AAAI Spring Symposium: Machine Reading. 2007:91.
[11] HEILMAN M, SMITH N A. Good question! statistical ranking for question generation[C]// Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010:609-617.
[12] SUTSKEVER I, VINYALS O, LEQ V. Sequence to sequence learning with neural networks[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014:3104-3112.
[13] LUONG M T, PHAM H, MANNINGC D. Effective Approaches to Attention-based Neural Machine Translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1412-1421.
[14] ZHOU Q, YANG N, WEI F, et al. Neural question generation from text: A preliminary study[C]// National CCF Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2017:662-671.
[15] SUBRAMANIAN S, WANG T, YUAN X, et al. Neural Models for Key Phrase Extraction and Question Generation[C]// Proceedings of the Workshop on Machine Reading for Question Answering. 2018:78-88.
[16] DUAN N, TANG D, CHEN P, et al. Question generation for question answering[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017:866-874.
[17] SERBAN I V, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016:3776-3783.
[18] XING C, WU Y, WU W, et al. Hierarchical recurrent attention network for response generation[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. 2018:5610-5617.
[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[20] XIONG C, ZHONG V, SOCHER R. Dynamic coattention networks for question answering [J]. arXiv:1611.01604, 2016.
[21] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[J]. arXiv:1611.01603, 2016.
[22] SEE A, LIU P J, MANNINGC D. Get To The Point: Summarization with Pointer-Generator Networks[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1073-1083.
[23] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1532-1543.
[24] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002:311-318.
[25] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72.
[26] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Text Summarization Branches Out. 2004:74-81.

SHI Yu-tao, born in 1997, postgraduate. His main research interests include natural language processing and machine learning.

SUN Xiao, born in 1980, Ph. D, professor, is a member of China Computer Federation. His main research interests include affective computing, natural language processing, machine learning and human-machine interactions.

(责任编辑:李亚辉)