计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700162

融合检索与生成的复合对话模型

杨慧敏　马廷淮
南京信息工程大学计算机与软件学院 南京 210044
(2432640905@qq.com)

摘　要　对话模型是自然语言处理的重要方向之一。现如今的对话模型主要分为基于检索的方式和基于生成的方式。然而,检索方式无法回应语料库中未出现的问句,而生成方式容易出现安全回复的问题。鉴于此,提出融合检索与生成的复合对话模型,通过将检索方式与生成方式相结合来弥补各自的缺点。首先通过检索模块得到 K 个检索上下文以及所对应的 K 个检索候选回应。在多回应生成模块中进一步结合检索上下文得到若干生成候选回应。最后的候选回应排序模块分为预筛选与后排序两个步骤。预筛选部分通过计算输入问题与候选回应的相似度得到最优检索回应与最优生成回应,后排序部分进一步选出对于输入问题最合适的回答。实验结果显示,相对于传统模型,复合对话模型在 BLUE 指标上提升了 6%,在多样性指标上提升了 12%。

关键词:对话系统;检索模型;生成模型;Transformer;后排序
中图法分类号　TP319.1

Compound Conversation Model Combining Retrieval and Generation

YANG Hui-min and MA Ting-huai
College of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract　Conversation model is one of the important directions of natural language processing. Today's dialogue models are mainly divided into retrieval-based methods and generation-based methods. However, the retrieval method cannot respond to questions that do not appear in the corpus, and the generation method is prone to problems with safe responses. In view of this, a compound conversation model that combines retrieval and generation is proposed, and the retrieval method and generation method are combined to make up for their shortcomings. First, K retrieval contexts and corresponding K retrieval candidate responses are obtained through the retrieval module. In the multi-response generation module, retrieval contexts are further combined to obtain several generation candidate responses. The candidate response ranking module is divided into two steps: pre-screening and post-reranking. The pre-screening part obtains the optimal retrieval response and the optimal generated response by calculating the similarity between the input question and candidate responses, and the post-reranking part further selects the most suitable answer to the input question. Experimental results show that the BLUE index increased by 6%, and the diversity index increased by 12%.

Keywords　Conversation system, Retrieval model, Generation model, Transformer, Post-reranking

1　引言

对话模型(conversation model)是自然语言处理的重要研究领域之一,其任务是对用户输入的问题返回合适的语句作为回应。使用自然语言进行对话,对话主题多变且语义丰富,因此传统的基于规则的方法和基于模板的方法很难应用到开放域的对话模型中。近年来,构建对话模型的技术主要分为两种:基于检索的方式[1]和基于生成的方式[2]。检索式对话模型根据输入问题在预定义语料库中选择合适的回复[3-4]。然而对于语料库中未出现的句子,检索式对话模型很难给出合适的回复。如果依赖人工来收集语料库,将花费大量的时间、人力和物力,并且这几乎是不可能完成的任务。而生成式对话模型可以根据预先训练的模型合成语料库中所没有的新句子[5-6]。在接收到用户输入后,生成模型采用其他技术生成回复,作为对话模型的输出[7]。因为不依赖于预定义的回复库,所以生成模型并不要求非常精准的语料库,但同时它也存在着许多亟待解决的问题,如生成的回复可能会出现上下文不一致、语法错误或语句不通顺等[8]。基于以上问题,本文提出了融合检索与生成的复合对话模型,将检索模型与生成模型进行融合,提高了回应句子的贴切性。

到稿日期:2020-07-26　返修日期:2020-09-17
基金项目:国家自然科学基金(U1736105)
This work was supported by the National Natural Science Foundation of China(U1736105).
通信作者:马廷淮(thma@nuist.edu.cn)

2 融合检索与生成的复合对话模型

将检索模型与生成模型应用于对话领域中各有优缺点。检索模型直接使用语料库中已有的语句进行匹配并做出回应,因此回应语句质量较高。而生成模型不需要依赖规则,可以自动从对话语料中学习如何生成文本。本文将检索模型与生成模型相结合,提出融合检索与生成的复合对话模型,其主要分为 3 个模块:检索模块、多回应生成模块和候选回应排序模块。

2.1 检索模块

检索模型是对话模型中一种常用的方式。给定语料库 $D=\{(c_n,r_n)\}_n^N$,c 表示上下文,r 表示上下文所对应的回应。对于输入问题 x,检索模型计算其与语料库的每一个上下文的相似度。在得到输入问题与每个上下文的匹配得分后,根据得分进行排序,将得分最高的上下文所配对的回应作为输入问题的回答。检索模型的示意图如图 1 所示。

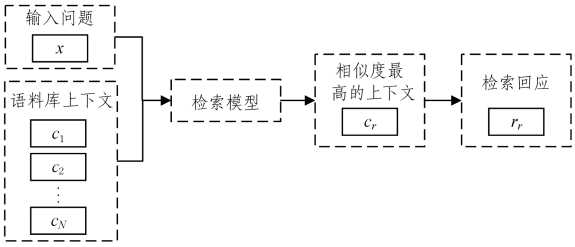


图 1 检索模型示意图

Fig. 1 Schematic diagram of retrieval model

检索模型使用深度学习模型进行检索,通过深度神经网络将输入数据进行编码,获得输入数据的隐藏表示,并根据隐藏表示进一步计算相似度。基于深度学习的方法可以学习出很多人类难以发现的非线性特征。在自然语言处理中,通常使用孪生网络来计算输入问题与语料库中上下文的语义相似度[9]。

在判断输入问题与语料库上下文是否相似时,孪生网络通过定义两个网络结构对输入语句分别进行表征,并通过余弦相似度来度量输入语句之间的相似程度。孪生网络的两个网络结构相同且参数共享,当输入的句子结构比较相似且来自同一领域时,孪生网络更为适用。孪生网络的另一种形式是伪孪生网络[10]。伪孪生网络的两个网络结构可以不同,也可以相同,但不共享参数。在计算句子和图片相似度时,或者输入的两个句子来自不同领域时,可以使用伪孪生网络来计算相似度。在对话模型中,对输入问题 x 和上下文 $c_n(1\leqslant n\leqslant N)$ 进行相似性计算时,一般两句话的领域相同且结构相似。因此,孪生网络更适合进行对话文本的相似度计算。

循环神经网络(Recurrent Neural Network,RNN)可以处理时间序列数据,使用时序反向传播算法进行学习。根据链式求导法则,RNN 存在梯度消失和梯度爆炸的问题,而长短期记忆网络(Long Short-Term Memory,LSTM)通过引入门控机制,缓解了这一问题。训练正向和反向两个 RNN,最后连接每个时间步的正向 LSTM 和反向 LSTM 的隐藏状态作为每个时刻的最终隐藏状态,达到同时记录过去的信息并且"看到"未来的信息的目的,从而获得更多的语句信息。在标准的 NLP 任务中(如文本匹配、命名实体识别等),双向 LSTM 相比标准 LSTM 有很大的提高。而 LSTM 的层数会极大地影响模型的训练,超过 3 层 LSTM 的模型往往很难训练。经过综合考虑,本文在孪生网络中使用两层双向 LSTM 来计算输入问题与上下文之间的相似性。

通过孪生网络分别计算出输入问题与语料库中每一个上下文的相似性。根据相似性进行排序,将相似度最高的前 K 个上下文作为检索上下文,并将其所对应的回应作为检索候选回应。在检索模块中,共得到 K 个检索上下文以及 K 个检索候选回应。

2.2 多回应生成模块

生成式对话模型在接收到用户输入的问题后,自主地组织词语生成一句回复作为对话模型的输出。生成式对话模型可以不断地从已有的对话数据中学习对话规律,并在每次收到用户输入的问题时,组织词语回答问题。生成模型的示意图如图 2 所示。

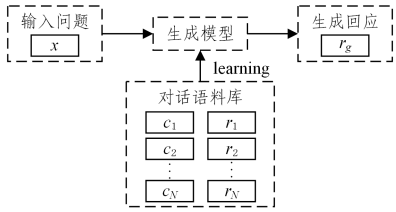


图 2 生成模型示意图

Fig. 2 Schematic diagram of generation model

在生成模型中,训练数据集以及输入问题中的有用信息需要转化为模型参数才能用于生成。然而模型的参数数量有一定限制,并且模型特别依赖于输入问题的质量,如果输入问题无法提供有效信息会使得模型生成"安全回复"的概率大大增加。为了解决这个问题,除了输入问题以外,检索模块中得到的 K 个检索上下文同样作为生成模型的输入并进行生成。这 K 个检索上下文与输入问题相似,可以给生成模型提供更丰富多样的信息,以减小出现"安全回复"的概率。

利用 Transformer[11] 代替 Seq2seq 进行生成,可以完全基于注意力机制来学习输入序列的相关信息,避免了 RNN 无法并行化的特性,使得模型可以在更短的训练时间内获得更好的效果。文献[11]提出了基于多头注意力机制的编码器解码器框架并引入位置编码来处理序列位置信息。

Transformer 模型的编码器部分由 N 个相同层组成,每层分为两个部分:多头注意力机制和前馈神经网络。每一部分都利用残差连接和层正则化进行连接,即每一部分的输出为 $LayerNorm(x+f(x))$,$f(x)$ 是这一部分本身的实现函数。与编码器相似,解码器也由 N 个相同层组成,每一层除了多头注意力机制和前馈神经网络两部分外,还有一个额外的掩盖注意力机制,这一额外部分是希望在对每一个位置进行预测时,只考虑这个位置以前的输出。同样,这一部分也使用了残差连接和层正则化进行连接。

每一个多头注意力机制包含多个自注意力机制,每一个自注意力机制称为一个头。自注意力机制的输入包含 d_k 维的查询和键,以及 d_v 维的值。查询与所有键进行点乘以计算相似性,并且除以 $\sqrt{d_k}$ 以防止结果过大。自注意力机制通过

softmax 函数来获得每一个值的权重，最后加权求和得到最终的输出。在实际运算中，一系列查询可以转化为矩阵 $\boldsymbol{Q}$，键和值也可以转化为矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。自注意力机制的输出如式(1)、式(2)所示：

$$\boldsymbol{S}=\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} \tag{1}$$

$$A=Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{S}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{S}$ 是 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似度矩阵，A 是注意力函数。为了获得更多信息，每一个自注意力机制在进行操作之前进行了线性变化，使得不同的注意力机制可以学习不同的表示。对于每一个头，即对于每一个注意力机制，其注意力函数如式(3)所示：

$$A_i=Attention(Q_i,K_i,V_i) \tag{3}$$

多头注意力机制最后的输出是由每一个自注意力机制相连接并经进一步变换而得到的，具体计算公式如式(4)所示：

$$multiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(A_1,A_2,\cdots,A_H)\boldsymbol{W}^o \tag{4}$$

其中，H 是多头注意力机制中头的数量，$\boldsymbol{W}^o\in R^{Hd_v\times d_f}$ 是训练得到的权重矩阵。考虑到计算成本，每一个头的维度减少到 $d_k=d_v=d_f/h$。

前馈神经网络主要是由两个线性全连接层组成，第一个线性全连接层是 Relu 激活函数，第二个线性全连接层的具体计算公式如式(5)所示：

$$FFN(x)=\max(0,x\boldsymbol{W}_1+b_1)\cdot\boldsymbol{W}_2+b_2 \tag{5}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$ 为权重矩阵，b_1，b_2 为偏差。前馈神经网络的作用就是将多头注意力机制得到的向量投影到更大的空间内，以便于提取重要信息，最后再投影回向量原来的空间。

由于没有使用 RNN 结构，注意力机制无法捕捉输入序列的顺序信息，因此我们使用位置编码来弥补这一缺陷。位置编码用于捕捉输入问题序列的相对位置信息或绝对位置信息。我们使用正弦和余弦函数对位置信息进行编码，位置编码的维度与输入编码维度相同，通过将位置编码与带有语义信息的输入编码相加来得到最终向量。位置编码的计算公式如式(6)和式(7)所示：

$$PE_{(pos,2m)}=\sin(pos/1\,000^{2m/d_f}) \tag{6}$$

$$PE_{(pos,2m+1)}=\cos(pos/1\,000^{2m/d_f}) \tag{7}$$

其中，pos 表示输入问题的词位置，m 表示位置编码维度。

在检索模块中，可以得到与输入问题最相关的 K 个相似上下文。为了进一步利用检索模型的相关信息，除了输入问题以外，检索模型检索出的 K 个相似上下文也将作为生成模型的输入，给生成模型提供更多信息。在多回应生成模块中，对于输入问题和检索上下文，共得到 $K+1$ 个生成候选回应。这 $K+1$ 个生成候选回应是生成模型从已有的对话数据中学习对话规律并组织词语生成的全新语句。

2.3 候选回应排序模块

检索候选回应无法处理语料库中未出现的问题，而生成候选回应可能会出现语法错误、语义不通等问题。对于输入问题，只能选择一条语句作为最终回应，因此，需要对检索候选回应与生成候选回应进行筛选，将与输入问题最贴切的回应作为输出。

利用检索模块共获得 K 个检索候选回应，利用多回应生成模块共获得 $K+1$ 个生成候选回应。由于这些候选回应数量较多，需要将其分别通过预先过滤与后排序这两个步骤进行处理。

2.3.1 预先过滤

最终回应是否贴切与输入问题直接相关。因此，与检索模块中计算输入问题与上下文的相似性不同，为了找出对输入问题最合适的回应，直接计算输入问题与候选回应的相似性针对性更强。对于每一个输入上下文，共得到 K 个检索候选回应与 $K+1$ 个生成候选回应。

针对这些候选回应，首先在检索回应和生成回应内部进行预先筛选，过滤掉相似度较低的回应，从而得到最优检索回应和最优生成回应以便于进一步的决策。

TF-IDF(Term Frequency-Inverse Document Frequency)是最为经典的检索算法[12]。TF-IDF 算法是一种基于统计的计算方法，用以评估一个词语对于一个文件集或一个语料库中的其中一份文件的重要程度。某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见词语，保留重要词语。在计算出所有句子的 TF-IDF 值后，计算两个句子 TF-IDF 值的余弦相似度，两个句子越相似，则余弦相似度越大。相比其他检索算法，TF-IDF 更易于理解，计算更简单，并且对语料库中的所有元素进行了综合考量。

本文使用 TF-IDF 进行相似度计算，对于输入问题和所有的候选回应分别计算对应的语句向量。得到语句向量后，将输入问题向量与检索回应向量和生成回应向量分别通过点积进行相似度计算。再通过对相似度进行排序，将检索候选回应中相似度最高的回应作为最优检索回应，最后通过同样的方式得到最优生成回答。

2.3.2 后排序

最优检索回应和最优生成回应是分别在检索模型和生成模型中筛选出的最优回应，并且经过上一部分的计算，可以得到这两个回应与输入问题的相似度。如果仅使用相似度作为选择最终回应的唯一标准，则无法进一步避免这两种不同模型各自的缺陷。检索模型只能从语料库中选择句子作为回应，使得回应固定不灵活；而生成模型的缺陷是容易生成无意义、通用的回应，即“安全回应”。我们发现安全回复如“我也是”“嗯嗯”“不知道”等通常不超过 3 个字[13]。基于此考虑，最优检索回应默认作为对于输入问题的回答。当最优生成回应的句子长度大于 3 且与输入问题的相似度大于最优检索回应时，使用最优生成回应作为最终回答。后排序模块的选择方式如式(8)所示：

$$r=\begin{cases}r_+, & \mathrm{sim}(c,r_+)>\mathrm{sim}(c,r_l)\ \text{or}\ \mathrm{len}(r_+)>3\\ r_l, & \text{otherwise}\end{cases} \tag{8}$$

其中，r_l 为最优检索回应，r_+ 为最优生成候选回应。通过这样的方式，最终回应既有可能来自检索模型，也有可能来自生成模型，这在一定程度上结合了两种模型。后排序模块进一步对候选回应进行决策，避免了生成模型带来的安全回应和语法混乱问题，也减小了检索模型出现对输入问题不相关回应的概率。

通过这两部分处理，本文得到了与输入问题最贴切的回

应,也就是模型的最终结果。该方法在一定程度上增加了回应句子的贴切度和延续性。融合检索与生成的复合对话模型如图 3 所示。

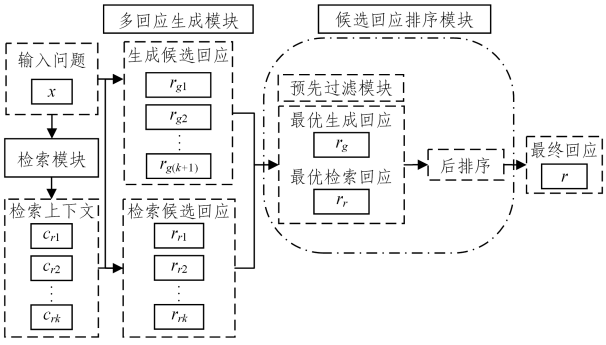


图 3 复合模型示意图

Fig. 3 Schematic diagram of compound model

3 实验结果与分析

3.1 数据集

生成对话模型适用于开放域对话过程,在训练模型时需要使用贴近生活的自然语言。由于在中文会话领域中没有标准的、权威的数据集用于实验,本文从新浪微博中收集对话文本。本文通过爬虫软件对用户发表的微博进行收集,将微博用户发表的微博作为上下文,微博下的第一条评论作为上下文对应的回应。由于微博 API 对数据传输量有所限制,我们最终收集了 400 000 条微博数据并对收集到的微博数据集进行过滤,以去除一些无意义回答,如"。。。""emmm"等。

3.2 参数设置

为了处理复杂的中文文本数据,对于数据集中的所有句子,本文首先使用"jieba"分词工具对文本进行分词操作,之后使用哈工大中文停用词表去除停用词。为了统一,对于这两个数据集,我们各随机抽取 1 000 个上下文-回应对作为测试集,将剩下的数据作为训练集。本文使用 word2vec 对词向量进行预训练,词向量大小为 300。对于每一个输入问题,从语料库中检索出 5 个与输入问题最相关的上下文,即 $k=5$[14]。Transformer 生成模型的编码器和解码器相同层的数量 N 为 6,隐藏单元为 512,多头注意力头数为 8,并且本文使用了 0.1 的丢弃率以预防过拟合。我们使用学习率为 0.000 1 的 Adam 优化器[15]来训练模型,并且设置批大小为 128。为了防止过拟合,我们使用测试集上对数似然的 Early Stopping[16]作为停止标准,并且应用宽度为 5 的波束搜索进行文本生成。本文的实验环境如表 1 所列。

表 1 实验环境

Table 1 Experimental environment

CPU	E5-2680
GPU	4 * Tesla K80
Memory	128 GB
Environment	Python3.6,Tensorflow1.4

3.3 对比分析

为了评估本文提出的融合检索与生成的复合对话模型,我们将其与带有注意力机制的序列到序列模型(Seq2seq Model with Attention,S2SA)[6]、基于 TF-IDF 的检索模型、Transformer 生成模型[10]、基于 TF-IDF 与 Seq2seq 的混合模型[17]进行了比较。这些基线模型包含本文所提模型的一些组成部分,以及一些最新模型。带有注意力机制的序列到序列模型(S2SA)是经典的生成模型,其编码器和解码器都使用隐藏层为 1024 的两层 LSTM,并且加入了注意力机制来进一步学习信息。其次,基于 TF-IDF 的检索模型是复合对话模型的一个重要部分,将其作为基线模型可以显示出检索模型与生成模型结合后生成模型是否起到作用。同理,与 Transformer 生成模型进行对比可以明显地观察出检索模型所起的作用。基于 TF-IDF 检索模型+Seq2seq 与本文模型相似,是经典的将检索模型与生成模型相结合的方法。但是,它没有对检索模型和生成模型进行创新,并且结合方式比较简单,没有进一步对两者进行融合。

在评估指标中,本文使用 BLEU[18]和 Distinct[19]来对回应进行检测。BLEU 的核心思想是通过比较预测回应和参考回应之间的 n-gram 匹配来计算匹配数,这些匹配与位置无关[20]。匹配越多,回应的得分就越高,质量就越好。BLEU 的计算公式如式(9)和式(10)所示:

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N}W_n\log P_n\right) \tag{9}$$

$$BP=\begin{cases}1, & q>p\\ e^{(1-p/q)}, & q\leqslant p\end{cases} \tag{10}$$

其中,q 表示预测回应长度,p 表示标准回应长度,BP 是简短惩罚,W_n 为权重,P_n 表示 n-gram 的精度。本节中,n-gram 取值为 1,2,3,即使用 BLEU-1,BLEU-2 和 BLEU-3 作为评估指标。除了 BLEU 指标,针对对话模型中的安全回复问题,本文通过 Distinct 来衡量回应的多样性。Distinct-1,Distinct-2 分别由不同的一元词和二元词数量与生成单词总数相除得到。

表 2 列出了各模型在数据集中的评估结果。基于 Transformer 的生成模型在各种指标上的性能优于 Seq2seq 模型,这表示 Transformer 算法仅依赖于注意力机制也可以很好地表示句子的语义,生成的回应信息更加丰富且更多样。这进一步表明了基于 Transformer 的模型对安全回复问题有一定的缓解作用。本文提出的复合模型在 BLEU 指标上的性能优于其他基线模型,这表明复合模型能很好地将检索模型与生成模型相结合,模型中的候选回应排序模块通过预选和决策两个步骤成功地选出了与输入问题最贴切的回应。

表 2 实验结果的对比

Table 2 Comparison of experimental results

Model	BLEU-1	BLEU-2	BLEU-3	Distinct-1	Distinct-2
Seq2seq	8.457	4.668	3.589	0.167	0.417
Retrieval model	11.53	7.50	6.24	0.28	0.654
Transformer	9.71	5.74	4.34	0.22	0.630
TF-IDF+Seq2seq	11.92	6.756	5.16	0.204	0.499
Our Method	12.71	8.13	6.41	0.23	0.660

在多样性方面,本文提出的复合模型在 Distinct 指标中略差于检索模型。这是因为检索模型直接选择语料库中的句

子作为回应，因此相比生成模型获得了较高的多样性。然而复合模型进一步在检索模型中融入了生成模型，使得多样性略有欠缺。尽管如此，本文提出的复合模型仍然获得了相对较高的值。在表 2 中，本文模型的 Distinct-2 指标达到了最高值，这进一步证明了候选回应排序的有效性。我们也尝试将检索模型与 Seq2seq 相结合，尽管同样利用了排序机制，但相比本文提出的模型而言，效果仍有所欠缺，这进一步证明了候选回应排序模块的有效性。

值得注意的是，利用 Transformer 生成回应，在生成的语句更通顺的同时，可以并行化计算，模型训练时间也显著缩短。如表 3 所列，在数据集中，每一回合 Transformer 的训练时间仅仅是 Seq2seq 的 1/6 左右。这表示去除 RNN 完全基于注意力机制进行生成，即使用 Transformer 代替传统的 Seq2seq 作为生成模块可以极大地提高模型的效率。

表 3 模型训练时间的比较

Table 3 Comparison of training time

Model	Dataset
Seq2seq	30 min\epoch
Transformer	5 min\epoch

本文使用微博数据集测试集中的 1 000 条对话数据来展示候选回应排序中后排序模块选择检索回应和生成回应的数量。如表 4 所列，在将检索模型与 Seq2seq 相结合时，检索回应与生成回应的数量相差不明显。然而将检索模型与 Transformer 相比较时，选择生成回应的数量比检索回应多了两倍，这个变化表明从后排序模块的角度来说，Transformer 的性能优于 Seq2seq。尽管本文后排序的规则偏向于生成模型，检索模型仍然占据了相当一部分，这进一步证明了无论是检索模型还是生成模型都为最后得到的回应做出了相当的贡献。

表 4 后排序模块中选择检索回应和生成回应的数量

Table 4 Selection of number of retrieved responses and generated responses in post-reranking module

模型	检索数量	生成数量
检索＋Seq2seq	437	563
检索＋Transformer	226	724

表 5 列出了本文模型和基线模型的两个例子。可以发现，由于检索回应从语料库中直接选择，所以检索模型的回应语义丰富、逻辑完整，不会出现语法错乱的问题。但是从第一个例子中可以发现，对于语料库中未出现的问题，检索模型不能很好地进行处理，出现回应句子与输入问题毫不相关的情况。使用 Transformer 生成回应可以对语义进行更加充分的分析，比起传统的 Seq2seq 生成模型来说，更少受到安全回复的影响，因此 Transformer 可以生成语义更加丰富的回应。从表 5 中也可以观察到，后排序模块起到了至关重要的作用，对检索回应和生成回应进行决策，过滤掉不合适的回应，为输入问题选出最为贴切的回应作为最终结果。通过以上分析可知，本文提出的复合模型将检索模型与生成模型相结合，经过预筛选和决策来得到最终回应，获得对输入问题更加贴切的回应，保证对话得以延续。

表 5 对话结果展示

Table 5 Display of conversation results

输入问题	你在哪过年
Seq2seq	对不起我没明白你的意思
检索	手伴抱枕海贼王黑执事
Transformer	在梅州啊一个山清水秀的好地方
复合模型	在梅州啊一个山清水秀的好地方
输入问题	好马上就回来
Seq2seq	哦哦
检索	你别急淡定
Transformer	慢慢来不用急的
复合模型	慢慢来不用急的

结束语 本文提出了融合检索与生成的复合对话模型，考虑到检索模型回应的语句流畅且无语法错误，而生成模型可以合成语料库中未出现的新句子，因此将检索模型与生成模型相结合来融合两者的优点。本文利用检索模块得到若干检索候选回应。使用基于自注意力机制的 Transformer 代替 Seq2seq 进行语句生成，在获得更好效果的同时大幅减少了训练时间。利用多回应生成模块可以得到若干生成候选回应。在候选回应排序模块中，通过预先筛选和后排序两部分得到最终回应。实验结果表明，复合模型得到的回应更贴切，语义更丰富。然而本文提出的模型主要针对单轮对话，在未来的工作中，可以加入层次结构，同时捕获句子的语义和多轮上下文的语义，以提高模型的回复准确性和话题一致性。

参 考 文 献

[1] WANG Y, HE Q T. Research on Intelligent Question Answering System[J]. Electronic Technology and Software Engineering, 2019(5): 174-175.

[2] VINYALS O, LE Q. A neural conversational model[J]. arXiv: 1506.05869, 2015.

[3] SHEN Y, HE X, GAO J, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C]// Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. Shanghai, China: ACM, 2014: 101-110.

[4] WAN S, LAN Y, XU J, et al. Match-srnn: Modeling the recursive matching structure with spatial rnn[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, USA: Margan Kaufmann, 2016: 2922-2928.

[5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. Montreal, Quebec, Canada: MIT PRESS, 2014: 3104-3112.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015: 1-9.

[7] ZHAO Y Y, WANG Z Y, WANG P, et al. A review of task-based dialogue systems[J]. Chinese Journal of Computers, 2020, 43(10), 1862-1896.

[8] HORI T, WANG W, KOJI Y, et al. Adversarial training and de-

coding strategies for end-to-end neural conversation models[J]. Computer Speech & Language,2019,54:122-139.

[9] BROMLEY J,GUYON I,LECUN Y,et al. Signature verification using a "siamese" time delay neural network[C]// Advances in Neural Information Processing Systems. 1994:737-744.

[10] CHI Z,ZHANG B. A sentence similarity estimation method based on improved siamese network[J]. Journal of Intelligent Learning Systems and Applications,2018,10(4):121-134.

[11] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. Long Beach,USA:MIT PRESS,2017:5998-6008.

[12] ZHU Z,LIANG J,LI D,et al. Hot topic detection based on a refined TF-IDF algorithm[J]. IEEE Access,2019,7:26996-27007.

[13] GU Y J,GUI X L,LI D F,et al. A Survey of Machine Reading Comprehension Based on Neural Networks[J]. Journal of Software,2020,31(7):2095-2126.

[14] PANDEY G,CONTRACTOR D,KUMAR V,et al. Exemplar encoder-decoder for neural conversation generation[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics,2018:1329-1338.

[15] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

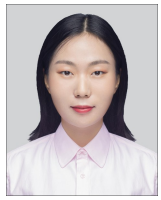
[16] PRECHELT L. Automatic early stopping using cross validation:quantifying the criteria[J]. Neural Networks,1998,11(4):761-767.

[17] WU Y,WEI F,HUANG S,et al. Response generation by context-aware prototype editing [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI,2019:7281-7288.

[18] PAPINENI K,ROUKOS S,WARD T,et al. BLEU:a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia,USA:Association for Computational Linguistics,2002:311-318.

[19] LI J,GALLEY M,BROCKETT C,et al. A diversity-promoting objective function for neural conversation models[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. San Diego,USA: Association for Computational Linguistics,2016:110-119.

[20] ZHOU Q A,LI Z J. Improved model and tuning method for natural language understanding of task-oriented dialogue system based on BERT[J]. Journal of Chinese Information Processing,2020,34(5):82-90.



YANG Hui-min, born in 1997, postgraduate. Her main research interests include data mining and data sharing.

MA Ting-huai, born in 1974, Ph.D, professor, is a member of China Computer Federation. His main research interests include data mining, data sharing and privacy protection.