文章编号：2095-2163(2020)08-0010-05 中图分类号：TP391.1 文献标志码：A

基于深度神经网络的文本问题生成技术综述

周青宇，周 明
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：问题生成作为自然语言生成任务的一种，近年来受到了研究人员的广泛重视。该任务的目标是在给定输入文本的情况下，生成一个关于该输入的合理的问题。问题生成有很多应用场景，例如在线教育、搜索引擎提示和问答系统等等。近年来，基于深度神经网络的方法使文本生成技术摆脱了基于模板的生成方式，开始采用基于编码器-解码器框架的文本生成方法。本文介绍了现有基于深度神经网络的文本问题生成技术的研究背景和国内外的研究现状，概述了目前该任务的相关数据集，并对各种方法进行了分类：基于循环神经网络的文本问题生成、基于 Transformer 的文本问题生成和基于预训练技术的文本问题生成。同时，总结了当前该任务面临的挑战和难点。

关键词：自然语言生成；神经网络；问题生成

Research on text question generation based on deep neural networks: a literature review

ZHOU Qingyu, ZHOU Ming
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 As one of the natural language generation tasks, question generation has been a research trend in recent years. The goal of this task is to generate a reasonable text question given a specific input text. There are many application scenarios for this task, such as online education, search engine promotion, question answering, etc. In recent years, language generation has evolved to a new era from template-based method to neural-network-based method. In this paper, we introduce the current research progress of text question generation based on deep neural networks, including datasets and models. The current methods can be summarized to three categories: RNN-based method, Transformer-based method and pre-training-based method. Finally, we introduce the current challenges and difficulties for question generation task.

【Key words】 natural language generation; neural network; question generation

0 引 言

随着数据量的增长和图形处理器(GPU)计算能力的提升，基于深度神经网络的自然语言生成技术为计算机处理并产生人类语言带来了新的发展前景。

问题生成作为人工智能和自然语言处理领域的一个任务，可以自动地从一段给定文本当中生成一个相关联的问题，且该问题可以由该句子或该句子中的某一部分进行回答。该技术有着很多实际应用，例如在文本问答或搜索引擎当中生成候选问题；在英文教育领域中进行自动的阅读理解问题生成。因此，利用计算机自动生成有意义且质量更高的自然语言问题，目前吸引了大量研究人员的目光。

从文本问题生成的历史发展可以分为如下阶段：基于规则的文本问题生成、基于传统统计机器学习的文本问题生成和基于深度神经网络的文本问题生成。本文聚焦于基于深度神经网络的文本问题生成，并将其近年来的发展情况分为 3 个类别：基于循环神经网络(RNN)的文本问题生成、基于 Transformer 的文本问题生成和基于预训练技术的文本问题生成。

1 国内外研究现状

通过对国内外基于深度神经网络的文本问题生成技术的调研，可以将解决该任务的方法分为 3 类。

(1) 基于循环神经网络(RNN)的文本问题生成。这类方法是基于循环神经网络组成的编码器-解码器结构进行文本问题生成。同时，该方法还可以结合拷贝机制，从而可以将给定文本中的内容拷贝到输出问题当中[1]。该方法可以生成答案无关的问题，即在输入文本时不指定具体的答案。同时，通过与答案特征结合，该任务的输入可以指定一个具体的答案位置或答案单词，从而将基于 RNN 的模型扩展为答案相关的问题生成模型。除了答案的位置，这类模型还可以与问答系统协同训练，从而在同一个模型下进行问题生成和自动问答[2]。

作者简介：周青宇(1992-)，男，博士研究生，主要研究方向：文本自动摘要、自然语言生成。

收稿日期：2020-03-28

（2）基于 Transformer 的文本问题生成。Transformer 模型是继循环神经网络之后的新一代序列到序列模型，其完全基于自注意力（Self-Attention）结构，而不采用对时间序列的循环展开。Transformer 结构的提出使神经网络机器翻译任务取得了重大突破，性能获得了很大提升。研究人员将 Transformer 结构同样应用到了问题生成任务当中，仍然可以与拷贝机制结合，从而提高文本问题生成任务的性能。

（3）基于预训练技术的文本问题生成。预训练技术是近年来开始获得突破性进步的自然语言处理处理技术。该技术从大量的语料中学习自然语言知识，从而提高下游任务的表现。其中较为常用的是预训练模型 BERT，该模型提出了掩码训练的方式训练语言模型。在下游任务中，通过对语言模型的微调即可以获得较好的表现。后续的研究工作探索了多种不同的预训练模型，并将其用于问题生成任务中，使该任务的性能获得了进一步的提升[3]。这类方法的模式为“预训练-微调”，且一次预训练就能够利用大量的语料，微调过程时间更短，其在文本问题生成领域具有十分广阔的应用前景。

2　数据集

深度神经网络技术的兴起基于两个必要因素，即计算机性能的提升和大规模训练数据的获取。在文本问题生成领域，近年来也出现了许多大规模的数据集，使得基于深度神经网络的方法获得实际应用。现有的文本问题生成数据集多是基于自动问答数据集改造而来，例如 SQuAD 和 NewsQA 等等。当前流行的文本问题生成数据集的多种统计信息，见表 1。

表 1　文本问题生成数据集统计信息

Tab. 1　Statistics of text question generation datasets

数据集	数据条数/K	输入粒度	是否有答案	答案形式
SQuAD 1.1	86.6	句子	有	抽取式
SQuAD 1.1	86.6	段落	无	抽取式
SQuAD 2.0	150	段落	部分无答案	抽取式
MSMARCO	1 000	段落	有	生成式
NewsQA	120	段落	部分无答案	生成式

SQuAD1.1 数据集是由斯坦福大学发布的机器阅读理解数据集。该数据集中的每一条是一个三元组（P，Q，A），其中 P 为段落，Q 为问题，A 为对应的答案，三者均为文本形式。该语料构建自英文维基百科。研究人员首先从维基百科随机选出 536 篇文章并抽取出段落；接下来则使用了众包方式标注出这些段落中值得提问的答案并写出对应的问题。而在文本问题生成任务中，研究人员则从另一个角度使用这个数据集，即从 P 或（P，A）生成 Q。由于该数据集的 P 为段落，而在文本问题生成任务中，研究人员选择了不同的输入级别。Zhou 等人首先使用了句子作为输入。由于 SQuAD 数据集中的输入 P 为段落，因此将 P 中包含答案 A 的句子作为输入。Du 等人则使用了整个 P 作为输入，且不使用答案 A。后续工作中对该数据集的使用也有所不同。有些工作使用了带答案的版本，即答案相关的问题生成；而其他工作则只使用了段落，即答案无关的问题生成。

斯坦福大学之后又发布了 SQuAD2.0 数据集，该数据集的特点是其中有些问题在原文中无法找到答案。有研究人员也基于 SQuAD2.0 进行无法回答的问题生成。

微软发布了阅读理解数据集 MSMARCO，该数据集与 SQuAD 的不同点在于其答案 A 不是 P 中的原文。因此，回答的过程需要进行生成而非简单的抽取。该数据集规模十分庞大，包含一百万条问答数据。其数据来源为 Bing 搜索引擎的搜索记录，因此其问题的语言风格与正规文本有所不同。

NewsQA 是由微软发布的问答数据集，该数据集包含了 12 万个问答对。该数据集的文本来源为 CNN 新闻语料，与前面两个数据集不同，NewsQA 中的答案需要从多个句子中整合信息才能够回答，因此其问答难度相对前面两个数据集更高。此外，NewsQA 当中也存在没有答案或单一答案的问答对。

3　文本问题生成方法

3.1　基于循环神经网络（RNN）的文本问题生成

（1）答案相关的文本问题生成

在任务设定下，其定义由二元组（P，A）生成对应的 Q，且生成的 Q 与 P 组成的问答对的答案应当为给定的答案 A。

基于循环神经网络的答案相关的问题生成模型 NQG++扩展了基于编码器-解码器结构的序列到序列模型，模型的结构如图 1 所示。

该模型中的编码器扩展了答案位置和词汇级别的特征。答案位置由 BIO 模式编码，表示句子中的一个片段是答案 A。具体地，B 代表答案的开始位置，I 代表答案内部，而 O 代表不是答案的其它单词。对于词汇级别的特征，该工作使用了词性、命名实体和大小写单词等特征。在解码过程中，该模型使用了带注意力机制和拷贝机制的 RNN 解码器。

其解码过程可以如公式(1)-(5)表示:

$$s_t = GRU(w_{t-1}, c_{t-1}, s_{t-1}), \quad (1)$$

$$s_0 = \tanh(W_d h_1 + b), \quad (2)$$

$$e_{t,i} = v_a^T \tanh(W_a s_{t-1} + U_a h_i), \quad (3)$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i=1}^{n} \exp(e_{t,i})}, \quad (4)$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i. \quad (5)$$

其中,GRU 表示解码器中的循环神经网络单元; W_d 为初始化 GRU 的 MLP 中的可训练参数; W_a 和 U_a 为注意力机制中的参数[4]。

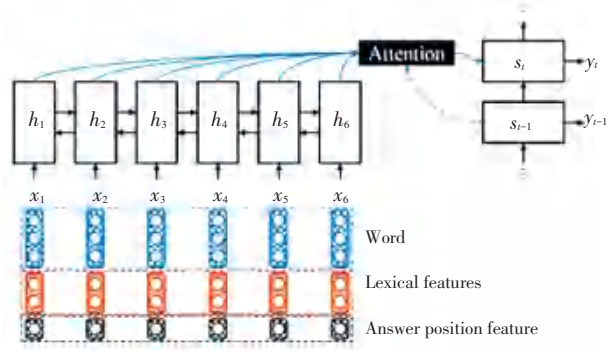

图 1 基于循环神经网络的答案相关的问题生成模型

Fig. 1 Answer-aware question generation based on recurrent neural networks

(2)答案无关的文本问题生成

在该任务设定下,其定义为由段落 P 生成对应一个问题 Q,且 Q 的答案应该包含在 P 中。但由于任务并未显示指定答案,因此有多种可能的问题 Q 存在。

Du 等人提出了基于循环神经网络的答案无关的问题生成[5]。其模型结构也是基于编码器-解码器结构的序列到序列模型,并使用了注意力机制。模型同时使用了句子和段落级别的输入。在解码过程中,模型将句子级别的隐含状态和段落级别的隐含状态拼接起来,以初始化解码器状态。但在解码过程中,注意力机制只应用于句子级别的隐含状态。模型在后处理阶段还进行了未登陆词(UNK)替换,将 UNK 替换为解码器注意力分数最高的输入单词。

基于以上两种模型,研究人员提出了大量的后续工作以提高文本问题生成任务。与生成对抗网络(GAN)结合可以进行问题类型的预测或领域自适应。通过对答案相关文本问题生成模型中输入答案方法的改进也可以提高生成问题类型的准确度。将强化学习技术应用在生成问题质量的评估过程[6]当中,也可以提升该任务的效果。此外,多种网络结构也被相继提出,例如关键词抽取和问题生成两段式模型、最大指针和门控注意力模型、上下文信息感知模型等等[7-10]。

将问题生成与自动问答两个对偶任务进行结合[2,11]并互相增益也成为一个研究热点。进行跨语言的生成[12]可以解决一部分数据缺失的问题。此外,问题生成技术还在在线教育和开放领域对话任务中得到了和具体应用场景相关的更进一步研究。

3.2 基于 Transformer 的文本问题生成

Transformer 模型是基于多头自注意力(Multi-Head Attention,MHA)结构的一种序列到序列模型,如图 2 所示。其优点是不需要在时间维度上展开,因此提高了模型的并行性。

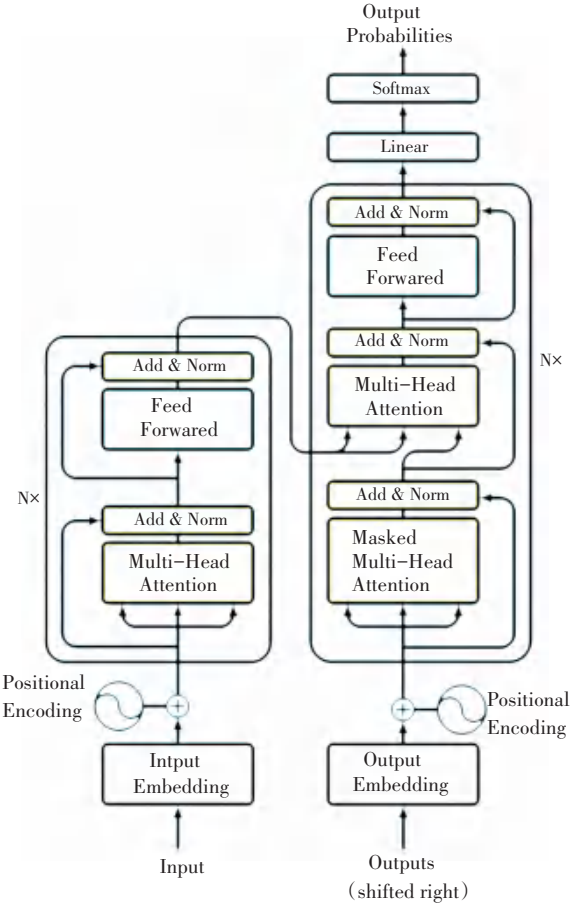

图 2 基于多头自注意力的 Transformer 模型

Fig. 2 Transformer based on Multi-Head Attention

Scialom 等人提出了使用 Transformer 模型进行答案无关的文本问题生成任务[13]。此外,作为一种序列到序列模型,该模型同样可以与拷贝机制结合。通过大量的实验发现,在 Transformer 结构下,使用了拷贝机制、ELMo 词向量和后处理 UNK 的方法在 SQuAD 数据集上取得了最佳的效果。

3.3　基于预训练技术的文本问题生成

3.3.1　基于预训练语言模型的文本问题生成

BERT 是一种基于双向 Transformer 的预训练语言模型。该模型使用掩码机制学习掩码处的上下文信息，从而得到一个基于掩码的语言模型（Masked Language Model）。在下游任务上进行微调后即可获得很好的效果。

Chan 等人提出了基于 BERT 的答案相关的文本问题生成模型 BERT-SQG[14]。模型首先将段落、答案和当前时刻 i 已生成问题的单词按照 BERT 的格式进行拼接作为输入：

$$X_i = ([CLS], P, [SEP], A, [SEP], \hat{q}_1, \cdots, \hat{q}_i, [MASK])$$

之后，按照 BERT 的掩码语言模型，即可以预测[MASK]位置的单词，也就是下一个时刻 $i+1$ 需要生成的问题的单词。这个过程不断迭代，直到生成[SEP]为止。最后，将所有生成的 $\hat{q}_i$ 拼接起来即获得了生成的问题 Q'。

3.3.2　基于预训练序列到序列模型的文本问题生成

Dong 等人提出了一种统一预训练语言模型 UniLM[8]，该模型是掩码语言模型和序列到序列模型的结合，因此即可以完成语言模型任务，也可以进行有条件的语言生成任务。UniLM 中包含了 3 种预训练任务，分别为双向语言模型、从左至右语言模型和序列到序列模型。该模型通过简单地控制掩码的方式完成这 3 种预训练任务。

在微调过程中将问题和答案进行拼接作为模型的输入：$X = ([SOS], P, [EOS], A, [EOS])$。接下来，该模型采用其序列到序列方式微调模型并进行问题的解码。除了进行问题的生成，UniLM 生成的问题还可以提高问答系统在 SQuAD 问答任务上的性能。

4　文本问题生成的挑战和难点

文本问题生成经历了基于规则、统计机器学习和深度神经网络 3 个阶段的发展，已经取得了突破性的进步。尤其是深度神经网络的引入使得该任务摆脱了基于规则模板的束缚，进入了新的研究时代。然而，现有工作中仍然存在一些亟待解决的问题，在本文中总结如下：

（1）生成的问题无法用指定的答案回答。在答案相关的文本问题生成任务中，生成的问题 Q 理应可以使用给定的答案 A 进行回答。然而，现有工作中普遍存在问题 Q 与答案 A 不对应的情况。这大大限制了该技术在实际场景中的应用。因此，解决该问题是目前文本问题生成任务的一个重点研究方向。

（2）领域自适应的问题。绝大多数现有工作都是基于 SQuAD 数据集进行文本问题生成。然而，该数据集基于维基百科进行标注，且其答案部分多是事实类问题（如命名实体）。因此在其他的语料，例如口语和聊天语料中，训练好的模型存在领域适应的问题。因此如何将在问答数据上训练的问题生成模型迁移到其他领域中是一个十分有应用价值的研究方向。

（3）缺乏大规模语料的问题。现有的问题生成语料均是英文语料，对于其他语种存在语料缺乏的问题。近年来神经网络机器翻译正朝着无监督与弱监督方向深入研究。然而，其中通用的反向翻译（Back-Translation）方法难以直接用于文本问题生成任务当中。因此，对于该任务，在缺乏语料的情况下无监督与弱监督学习是一个尚待解决的难题。

5　结束语

作为人工智能和自然语言处理领域的一个任务，文本问题生成成为近年来的一个研究热点。本文简述了基于深度神经网络的文本问题生成技术：基于循环神经网络的文本问题生成、基于 Transformer 的文本问题生成和基于预训练技术的文本问题生成。利用神经网络技术，文本问题生成任务近年来取得重大突破。针对该任务面对的挑战，未来可以从生成问题与答案的相关性、模型的领域自适应和无监督与弱监督几个角度进行更深入的研究。

参考文献

[1] GU J, LU Z, LI H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence Learning[J]. In Proceedings of the 54th ACL (Volume 1: Long Papers), 2016:1631-1640.

[2] SONG L, WANG Z, HAMZA, W. A unified query-based generative model for question generation and question answering [J]. arXiv preprint arXiv:1709.01058. 2017.

[3] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation [C]// In Advances in Neural Information Processing Systems, 2019:13042-13054.

[4] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473. 2014.

[5] DU X, SHAO J, CARDIE C. Learning to ask: Neural question generation for reading comprehension[J]. arXiv preprint arXiv: 1705.00106. 2017.

[6] YUAN X, WANG T, GULCEHRE C, et al. Machine comprehension by text-to-text neural question generation[J]. arXiv preprint arXiv:1705.02012, 2017.

（下转第 18 页）