

文章编号: 1003-0077(2021)07-0001-09

问题生成研究综述

吴云芳¹, 张仰森²

(1. 北京大学 计算语言学教育部重点实验室, 北京 100871;
2. 北京信息科技大学 智能信息处理研究室, 北京 100101)

摘要: 问题生成是给定文本, 自动生成内容通顺、语义相关的自然语言问题。问题生成可应用于教育领域的阅读理解、辅助问答系统和对话系统, 因此近年来引起了研究者的广泛关注和兴趣。该文对问题生成的相关研究进行了综述。首先阐释了问题生成的研究意义与应用场景, 继而简略概述了基于规则的问题生成方法, 然后从输入文本是句子/段落、有/无答案信息等不同角度全面阐述了基于神经网络的问题生成模型。该文还介绍了问题生成的评价方法, 分析讨论了现有工作的不足, 并展望了未来可能的研究方向。

关键词: 问题生成; 机器阅读理解; 神经网络模型

中图分类号: TP391

文献标识码: A

A Survey of Question Generation

WU Yunfang¹, ZHANG Yangsen²

(1. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China;
2. Institute of Intelligent Information Processing, Beijing Information Science
and Technology University, Beijing 100101, China)

Abstract: Question generation (QG) aims to automatically generate fluent and semantically related questions for a given text. QG can be applied to generate questions for reading comprehension tests in the education field, and to enhance question answering and dialog systems. This paper presents a comprehensive survey of related researches on QG. We first describe the significance of QG and its applications, especially in the education field. Then we outline the traditional rule-based methods on QG, and make a detailed description on the neural network based models from different views. We also introduce the evaluation metrics of generated questions. Finally, we discuss the limitations of previous studies and suggest future works.

Keywords: question generation; machine reading comprehension; neural network model

0 引言

问题生成(question generation, QG)是与机器阅读理解(machine reading comprehension, MRC)相关联的一个自然语言处理任务。其定义为: 给定一个自然语言的文本, 自动生成内容相关、语法通顺的问题。为了清晰地呈现任务形式, 表 1 给出了一个汉语问题生成的例子, 摘选自 NLPCC-2017 问答评测数据集^[1]; 表 2 给出了一个英语例子, 摘选自阅读理解数据集 RACE^[2], 其中, 波浪线标示了问题所

对应的关键句。

问题生成可形式化定义如下: 给定一个文本 $D = \{x_1, \dots, x_n\}$, 目标是生成自然语言问题 $Q = \{y_1, \dots, y_{|y|}\}$, 即:

$$\begin{aligned}\bar{Q} &= \arg \max P(Q | D) \\ &= \arg \max_y \sum_{i=1}^{|y|} \log P(y_i | D, y_{<i})\end{aligned}\quad (1)$$

其中, x_i 表示输入文本中的一个词语, n 表示输入文本的长度; y_i 表示生成问题中的一个词语, $|y|$ 表示生成问题的句子长度。

近年来, 问题生成引起了研究者越来越广泛的

收稿日期: 2020-10-04

定稿日期: 2020-11-29

基金项目: 国家自然科学基金(62076008); 科技创新 2030—“新一代人工智能”重大项目(2020AAA0106600)

关注和兴趣,涌现出了大量高水平研究论文。本文将对问题生成的前沿研究进行综述,组织结构安排如下:第1节阐述了问题生成的研究意义与应用;第2节概括描述了基于规则方法的问题生成研究;第3节对问题生成的神经网络模型方法进行了详细的综述,从输入句子/段落、有/无答案信息等不同角度进行了全面描述;第4节介绍了问题生成的评价方法;第5节对当前问题生成研究的不足进行了思考,对未来的研究方向进行了展望;第6节为结语。

表1 汉语阅读理解的问题生成例子

<p>段落: 香港中学生联盟是不属于任何政党的香港学生组织。当初由一群倾向泛民主派的香港中学生于2003年香港七一游行后8月21日发起及成立。该联盟由不同中学的学生组成。<u>该联盟的成立目的是“以学生的角度了解社会及政治事务”、“就影响着中学的政府政策发表意见”及“提升青少年对社会及政治事务的关注及责任心”。</u>该联盟的活动包括每年举办一次“青年圆桌会议”及参与政治事务的意见表达。</p> <p>问题: 为什么成立香港中学生联盟?</p>
--

表2 英语阅读理解的问题生成例子

<p>Passage: In some countries, people eat rice every day. Sometimes they eat it twice or three times a day for breakfast, lunch and supper. Some people do not eat some kinds of meat. Muslims, for example, do not eat pork. <u>Japanese eat lots of fish. They are near the sea, so it is easy for them to get fish.</u> <u>In the west, such as England and the USA, the most important food is potatoes.</u> People there can cook potatoes in many different ways. Some people eat only fruit and vegetables. They do not eat meat or fish or anything else from animals. They eat food only from plants. They say the food from plants is better for us than meat.</p> <p>Question1: Why do Japanese eat lots of fish?</p> <p>Question2: What is the most important food in some western countries?</p>
--

1 问题生成的研究意义

问题生成研究最广泛的应用当在教育领域。在教育场景下,问题被认为是进行学习和教学评估的重要工具。阅读是人类汲取知识的重要途径,而通过问题可以测试读者对于关键概念的理解、对于文本信息的提取归纳、对于核心思想的演绎推理等。同时,高质量的问题又能指导学习者进行有效的学习,提升其阅读质量。然而,构建一定数量且拥有较

高质量的问题是一件十分耗时且极其艰难的工作,因此,多年前研究者就尝试将问题生成这样一件费时费力的事情交给计算机去完成。早在2003年,Mitkov和Ha就利用计算机技术来自动生成多项选择题^[3],他们的研究表明,计算机技术比人力构造更为高效,构造1000个问题借助计算机系统的帮助仅需30个小时,而纯人力构造则需要115个小时。后来,Heilman和Smith^[4]面向教育评测、Lindberg等人^[5]面向在线学习分别进行了问题生成研究。此外,Rus等人^[6]还多次组织了问题生成的竞赛评测。

近期以来,基于神经网络模型的、教育领域的问题生成研究逐步进入了研究者的视野。Lai等人发布了真实英语考试的阅读理解数据集RACE^[2],基于此数据集,Gao等人^[7]、Zhou等人^[8]开展了问题干扰项自动生成的研究。2018年Guanliang Chen等人发布了专门针对教育领域的问题生成数据集LearningQ^[9]。斯坦福大学Willis等人^[10]面向段落级问题生成,聘请教育领域专家标注了文档中值得提问的关键短语,与众包平台的标注结果进行了详细的比较,分析表明教育专家关注的问题焦点与普通标注者存在一定差异。

此外,问题生成对其他自然语言处理任务也有巨大的促进作用:

其一,辅助问答系统。自动问题生成可以协助构建问题语料库,缓解问答系统语料不足的难题。例如,Tang等人^[11]、Duan等人^[12]和Zhang等人^[13]的研究均表明,利用自动生成的问题作为数据补充,或者将答案抽取与问题生成作为孪生任务同时训练,可以显著提升问答系统的性能。

其二,辅助对话系统。在对话系统中嵌入问题生成模型,可以加强系统与人的交互,帮助系统更准确地判断人的意图,从而使人机对话更为深入流畅。例如,Wang等人^[14]实验结果表明自动生成的问题可以触发更深入的对话;Aliannejadi等人^[15]的研究表明,提出一个高质量的澄清式问题,可以使信息检索的p@1准确率提升170%。

2 基于规则方法的问题生成研究

和其他自然语言处理任务一样,早期学者采用基于规则的方法来自动生成问题。Mitkov等人^[3]基于转换规则生成问题,利用WordNet^[16]知识生成多项选择题中的干扰项。Heilman和Smith^[4]实现了一种过度生成然后排序的方法(overgenerate-

and-rank)用于事实性问题生成: 首先编写句法和词汇规则将陈述句转换为疑问句, 然后用逻辑回归模型对生成的问题重新排序。Lindberg 等人^[5]使用模板方法融入语义角色来进行问题生成。早期基于规则的方法取得了一定的效果, 初步证明了自然语言处理技术可以帮助生成问题从而减少人工劳动。但是, 由于语言本身的复杂性, 人工发现和归纳出所有的问题规则几乎是不可能的; 而且规则方法难以扩展, 为某一个领域制定的规则通常很难在其他领域快速移植。

最新的一项研究中, Dhole 和 Manning^[17]也采用基于规则方法来自动生成问题, 使用了依存句法树、语义角色标注等自动分析技术, 调用了 WordNet^[16]、VerbNet^[18]中的语义知识, 获得了比神经网络模型更优的效果。虽然他们的规则方法在英语中取得了满意的结果, 但是成熟的自动句法语义分析技术、优良的第三方语义资源在其他语言中是很难获得的, 也因此 Dhole 和 Manning^[17]的方法不易在其他语言中推广。

3 基于神经网络模型的问题生成研究

近年来, 基于神经网络模型的问题生成研究层出不穷。大多数工作都采用了序列编码-解码框架, 使用带注意力机制的长短期记忆网络(long short-term memory, LSTM), 而采用转换网络(transformer network)的研究并不多见。由于缺乏专门用于问题生成的数据集, 已有的研究大多是在阅读理解数据集上进行训练和测试, 例如, SQuAD^[19]和 MS-MARCO^[20]。先进的深度学习模型不仅使得自动问题生成得以端到端地执行, 省去了人工设计规则的繁琐过程, 同时还取得了更为理想的效果, 生成了质量更高的问题。林林总总关于问题生成的研究由于其输入内容不尽相同, 也就不再笼统地陈述。根据输入文本是句子/段落、有/无答案信息, 本节将从以下几个方面来缕析基于神经网络模型的问题生成的研究进展。

3.1 问题生成的神经网络开端研究

在 2017 年, 研究者开始尝试使用神经网络模型来自动生成问题。Du 等人^[21]采用双向 LSTM 编码器分别对段落文本和关键句进行编码, 再使用一个 LSTM 网络结合注意力机制进行解码。他们在 SQuAD 数据集上进行了问题生成的实验, 由于

SQuAD 的测试数据没有完全公开, 于是将公开部分的 SQuAD 数据按照 80%:10%:10%的比例划分为训练集、开发集和测试集, 后被称为 Du split。实验结果的 BLEU-4 值达到了 12.28; 而采用 Heilman 和 Smith^[4]的基于规则的方法, BLEU-4 仅为 11.18, 这表明神经网络模型比基于规则的方法能生成更高质量的问题。同年, Zhou 等人^[22]也提出了神经网络模型的问题生成框架, 不同的是, 他们对输入文本进行编码时融入了答案短语对应的 BIO 位置信息(B: 答案的首词; I: 答案内部的词; O: 答案外部的词), 解码器采用基于注意力的门控循环单元(gate recurrent unit, GRU)来生成问题, 并应用拷贝机制从原文中拷贝一些词语。他们同样在 SQuAD 数据集上进行了问题生成的实验, 将原先的开发集以 1:1 的比例随机划分为开发集和测试集, 后被称为 Zhou split。同样在 2017 年, Yuan 等人^[23]在 LSTM 编码-解码框架下, 融入了强化学习策略, 用预训练的 QA 模型来测试自动生成的问题能否被正确回答, 在 SQuAD 上评测, 最好结果 BLEU-4 为 10.5。

3.2 有答案的句子级问题生成

有答案的句子级问题生成(answer-aware sentence-level question generation): 给定一个句子及对应的答案信息, 自动生成问题。Zhou 等人^[22]最早建模了这个任务, 后期很多工作都在探索如何有效地利用答案信息。Song 等人^[24]采取了三种策略来匹配答案和句子: 完全匹配、注意力匹配和最大注意力匹配。Sun 等人在指针生成模型(pointer-generator model)^[25]的基础上, 利用答案信息来预测疑问词的概率分布, 并加入相对位置编码使模型更多关注答案词周边的语境信息^[26]。董孝政等人^[27]将“潜在提问对象”的位置信息与全句信息相融合。Kim 等人^[28]将原文中对应的答案词语替换为一个特殊标记, 同时增加了一个关键词网络(keyword-net)去更好地捕捉答案中的关键信息。Liu 等人^[29]提出了一个问题生成的拷贝网络(copy network for question generation)用以预测问题中的词语应该从原文中拷贝还是从词表中生成, 他们在依存树上用图卷积网络(graph convolutional networks, GCN)来预测线索词(即需要拷贝的词), 然后将线索词作为特征融入原文的编码。Zhou 等人^[30]采用层级结构的联合学习框架, 嵌入语言模型来帮助编码器更好地捕捉输入句子的句法信息, 从

而促进解码器生成更高质量的问题。Zhou 等人^[31]将问题类型预测作为一个子任务,在联合学习框架下进行问题生成。Li 等人^[32]认为,前人的假设“距离答案词越近的词越重要”不完全准确,他们利用第三方关系抽取工具(OpenIE)抽取结构化关系作为枢纽语境(to the point context),采用门控注意力机制(gated attention mechanism)去动态注意文本表征和关系表征。Ma 等人^[33]提出了一个多任务学习框架:用分类器判断生成的问题与答案在语义上是否匹配,利用输入句子和生成的问题来预测答案的位置,将多个损失函数置于一起来联合学习。Jia 等人^[34]引入复述知识来协助问题生成,他们先用回译(back-translation)方法自动获得句子、问题的复述语句,而后将复述生成作为一个子任务来辅助问题生成。

可以看到,自 2017 年基于神经网络模型的问题生成任务提出后,至 2020 年已引起了学界的快速关注并取得了长足进展,表 3 列出了上述不同研究的性能指标。表中 Zhou split 和 Du split 是在 SQuAD 数据上对开发集和测试集两种不同的数据划分方式。

表 3 有答案句子级问题生成的性能进展

研究工作 会议-年份	Zhou split		Du split	
	BLEU-4	METEOR	BLEU-4	METEOR
Zhou et al. ^[22] NLPCC-2017	13.29	—	—	—
Song et al. ^[24] NAACL-2018	13.91	—	13.98	18.77
Sun et al. ^[26] EMNLP-2018	15.64	—	—	—
Kim et al. ^[28] AAAI-2019	16.17	—	16.20	19.92
Liu et al. ^[29] WWW-2019	17.55	21.24	—	—
Zhou et al. ^[30] EMNLP-2019	16.23	—	—	—
Zhou et al. ^[31] EMNLP-2019	16.31	—	—	—
Li et al. ^[32] EMNLP-2019	16.37	20.68	16.27	20.36
Ma et al. ^[33] AAAI-2020	16.32	20.84	—	—
Jia et al. ^[34] ACL-2020	16.93	20.58	17.21	20.96

3.3 有答案的段落级问题生成

有答案的段落级问题生成(answer-aware paragraph-level question generation):给定一个段落(由多个句子组成)及对应的答案信息,自动生成问题。在基于句子生成问题的基础上,研究者开始思考如何利用更大范围的段落语境来帮助问题生成。Zhao 等人^[35]利用门控自注意力网络(gated self-attention network)来编码段落信息,在解码时应用了最大指针(maxout pointer)来解决重复生成问题,此工作成为了后来众多段落级问题生成的基础。为了解决语义漂移问题,Zhang 和 Bansal^[13]在强化学习框架下,引入了两个新的奖励因子(reward):问题复述概率(QPP)来评估生成的问题和标准问题是复述的概率;问答概率(QAP)来计算生成的问题和正确答案匹配的概率。Nema 等人^[36]应用了双解码器,第一个解码器生成一个草稿,第二个解码器采用双重注意力机制,同时关注到原文编码和第一个解码器生成的问题。Chen 等人^[37]采用双向门控图神经网络来编码段落信息,在强化学习框架下融合了交叉熵和评价结果作为奖励因子。为了更好地利用篇章信息,Tuan 等人^[38]提出了一个多步注意力机制来捕捉段落中多个句子之间的语义关联。

由于应用了更多的语境信息,段落级问题生成普遍比句子级问题生成取得了更好的性能,具体如表 4 所示。

表 4 有答案段落级问题生成的性能进展

研究工作 会议-年份	Zhou split		Du split	
	BLEU-4	METEOR	BLEU-4	METEOR
Zhao et al. ^[35] EMNLP-2018	16.85	20.62	16.38	20.25
Zhang et al. ^[13] EMNLP-2019	—	—	18.37	22.65
Nema et al. ^[36] EMNLP-2019	—	—	16.99	21.10
Chen et al. ^[37] NeurIPS 2019	18.30	—	17.94	21.76
Tuan et al. ^[38] AAAI 2020	17.76	21.56	17.09	21.25
Wang et al. ^[39]	24.32	26.10	22.78	25.49
Bao et al. ^[40]	26.30	27.09	24.70	26.33
Yan et al. ^[41]	26.72	27.64	25.01	26.83
Xiao et al. ^[42]	26.95	27.57	25.40	26.92

基于大规模语料的预训练模型在自然语言处理的各项理解和生成任务中取得了令人瞩目的成绩,在 2020 年的最新研究中,研究者纷纷将预训练模型用于问题生成任务。Wang 等人^[39]提出了一个自注意力蒸馏模型(self-attention distillation)来压缩庞大的转换预训练模型。Bao 等人^[40]利用遮罩策略学习词项之间的关系,利用伪遮罩去学习语段内部的关系。Yan 等人^[41]设计的预训练模型的目标是同时预测 n 元词项,应用了 n 支(n -stream)注意力机制。Xiao 等人^[42]提出了一个多阶预训练和精调框架,其训练目标是预测语义完整的连续语段而不是单个词语。这些预训练模型都在有答案句子级问题生成任务上进行了实验,表 4 列出了不同预训练模型的实验结果。

3.4 无答案的句子级问题生成

无答案的句子级问题生成(answer-agnostic sentence-level question generation): 仅给定一个句子而无答案信息,自动生成问题。沿袭传统的规则方法,Du 等人^[21]建模的问题生成就没有答案信息,其论文中没有特别关照问题焦点。Ma 等人^[43]基于细粒度的层面(aspect-based)来进行问题生成,但实验结果表明,自动预测需要提问的层面时,生成问题的性能并没有得到提升。Wang 等人^[44]提出了一个多模块互动框架(multi-agent communication framework),用一个局部抽取模块自动识别出值得提问的短语,用其帮助问题生成,进而用生成的问题反馈来提升关键短语的识别性能。Scialom 等人^[45]采用了基于注意力机制的转换网络,并针对性地增加了拷贝机制、命名实体泛化策略(placeholding strategy)和语境相关词向量,将 BLEU-4 值从 Du 等人^[21]的 12.28 提升至 13.23。

3.5 无答案的段落级问题生成

无答案的段落级问题生成(answer-agnostic paragraph-level question generation): 仅给定一个段落(由多个句子组成)而无答案信息,自动生成问题,这方面的研究比较有限。Du 等人^[46]提出了一个流水线模式(pipeline)来处理这个任务: 先使用层级编码获得段落表征来预测值得提问的关键句子(question-worthy sentences),而后再应用句子级别的问题生成模型。Subramanian 等人^[47]同样采用两阶段的处理方式: 先用一个指针网络(pointer

network)识别出关键短语,然后将关键短语作为答案线索来生成问题。Willis 等人^[10]也是两步策略: 首先利用一个包含丰富特征的网络来识别关键短语,继而基于关键短语来生成问题。

3.6 汉语问题生成研究

相比英语蓬勃发展、硕果累累的问题生成研究,汉语问题生成的基础数据、模型探索都非常有限。仅有的研究是,Kumar 等人^[48]提出用跨语言训练的方式来进行问题生成,在汉语的 DuReader^[49]数据上进行了实验,但结果不理想。Chi 等人^[50]提出了一个通用型的面向跨语言生成的预训练模型,在汉语的事实性问答数据 WebQA^[51]上进行了问题生成实验。

4 问题生成的评价方法

问题生成属于文本生成的范畴,在评价方法上借鉴了机器翻译、自动文摘等任务,一般采用自动评测和人工评测两种方式评价问题生成的质量。

自动评测采用 BLEU^[52]、METEOR^[53]和 ROUGE^[54],其中,以 BLEU-4 为主要指标。BLEU 主要计算 n 元词语的匹配准确度;METEOR 在计算相似度时考虑了同义关系、复述关系、词根联系;ROUGE 着重考察 n 元词语的召回程度。但自动评价方法机械地计算生成问题和标准问题的词语匹配程度,不能完全反映问题生成的质量,因此,一般问题生成研究中都会同时加入人工评测。从自动生成的问题中随机选择一些样例(例如,200 个),交由多名标注者(多为 3 名)从不同角度进行人工评价。常包括以下三个方面: ①流利度: 是否是合乎语法的和流畅的;②相关度: 是否与输入文本相关;③可回答性: 生成的问题是否可以被正确答案所回答(在有答案输入的情形下)。

研究者也对问题生成的评价方法进行了探讨。Nema 和 Khapra^[55]指出,传统的评价方法不能反映出问题的可回答性(answerability),他们提出了 QBLEU 评价方法,将相关实词、问题类型、命名实体、功能词语的考评融入 BLEU 计算。后来的有些工作也加入了 QBLEU 指标,例如 Nema 等人^[36]的研究。Sultan 等人^[56]指出,传统指标与多样性表达是反比相关的,进而提出了多样性感知(diversity-aware)的评价指标,不过这个指标还需要进一步验证。

5 讨论与分析

5.1 前人研究的不足

问题生成最初的研究主要是面向教育场景、为智能教育服务的 (Mitkov 和 Ha^[3], Heilman 和 Smith^[4], Lindberg 等人^[5], Rus 等人^[6]), 而神经网络时代大多数研究将问题生成当作问答系统的一个对偶任务, 在问答系统的数据上进行, 并没有考虑教育场景下问题生成的特殊需求。

面向教育的阅读理解问题与机器阅读理解问题存在差别。RACE^[2] 是一个真实的教育考试数据, 收集了大量的中国学生英语学习的阅读考试问题。前期研究中, 我们对 RACE 数据做了一些清洗, 删除了宽泛的万能问题, 如 “what’s the best title of this passage?”, 然后比较了 RACE 和 SQuAD 数据的分布差异, 如表 5 所示。

表 5 RACE 和 SQuAD 数据的分布差异

数据集	段落数	段落长度	问题数	问题长度	答案长度
SQuAD	20 958	134.8	97 888	11.31	2.91
RACE	12 734	263.3	20 486	10.87	6.36

可以看到: ①考试数据 RACE 是由教育专家构建的, 投入成本大, 耗时长, 数据规模 (段落数、问题数) 远小于众包平台构建的 SQuAD; ②RACE 的段落长度几近是 SQuAD 的 2 倍, 会完整地叙述一个故事或议论一个话题; ③SQuAD 数据中大多是事实性问题, 答案多是短小的命名实体, 而 RACE 提供的答案通常会完整的长句子。这样的数据分布使得面向教育领域的问题生成异常困难。我们重新训练了 Zhao 等人^[35] 的模型, 在 SQuAD 数据上 BLEU-4 可以达到 16.31, 而同样的模型在 RACE 数据上精细调参后 BLEU-4 仅为 8.83。

概括而言, 前人研究存在的不足体现在以下四方面, 而这也是目前问题自动生成的难点所在:

(1) 前人研究大多是给定答案来生成问题, 然而在真实的应用场景下, 是仅给定一段文档、无答案地生成问题。在一段文档中人工划定需要提问的关键短语 (或句子) 是一件耗时耗力、艰难的工作。前人对无答案的段落级问题生成研究非常有限而且性能不高, Du 和 Cardie^[46] 在 SQuAD 数据上评测的 BLEU-4 仅为 11.50, Subramanian 等人^[47] 的结果仅为 10.4, 远不能满足实际应用的需求。

(2) 前人工作投入了很多精力在输入文本中匹配答案信息, 但却很少关注输入文本自身的信息捕捉, 在问题生成上还未深入探究段落内容表示、篇章结构信息建模等问题, 而长文档的内容表示是阅读考试问题生成 (如 RACE) 面临的一个难题。

(3) 前人在自动生成问题时, 还未能充分利用人类知识, 例如语言复述、经验常识等, 而这些知识对于生成具有一定难度的、有一定推演力的问题非常重要。

(4) 前人的问题生成研究大多是基于英语数据进行的, 很少有工作投入到汉语问题生成的基础数据建设和模型研究上。

5.2 未来研究的展望

问题自动生成的研究日新月异, 涌现出了很多新颖的工作, 未来的研究将可能围绕以下方向展开。

(1) 面向更多样化的文本。之前的研究大多是将问题生成作为阅读理解的对偶任务, 主要基于阅读理解的数据集 SQuAD^[19], 其文本来源于维基百科, 答案多是简单的语块。未来的问题生成研究将面向更广泛领域的文本, 处理更多现实生活的任务。例如, 基于真实的考试数据 RACE, Gao 等人^[7]、Zhou 等人^[8] 对问题干扰项生成进行了研究。Yu 等人^[57] 面向在线产品和服务评价来自动生成问题, 以帮助用户快速地攫取信息。

(2) 服务于更深入的阅读理解。已有的研究大多是基于简短的段落级文本和语段答案来生成问题, 对文本内容的理解是有限的。未来的问题生成研究将面向更长的真实完整的文档, 需要对文本进行篇章级的深入理解和分析。例如, Pan 等人^[58] 提出了深度问题生成 (deep QA) 任务, 目标是对文本信息进行一定的推理和整合来生成复杂的问题。Ko 等人^[59] 构建了一个新的问题生成数据集 INQUISITIVE, 探讨人类在阅读一篇文档时可能发问的知识性问题, 寻求更高层次的语义和篇章理解。

(3) 充分利用预训练语言模型。基于大规模语料库的预训练语言模型在自然语言处理的各项任务中取得了令人瞩目的进展, 在问题生成研究中亦取得了骄人的成绩, 例如 Wang 等人^[39], Bao 等人^[40], Yan 等人^[41], Xiao 等人^[42] 的研究。面向普适任务的预训练模型如何更好地应用于问题生成任务, 还有很大的探索空间。

(4) 在更多语言中开展研究。目前的问题生成研究主要是基于英语文本数据, 在可预见的未来, 借

鉴英语的研究成果,问题自动生成研究将在更多语言中展开,并发布公开评测数据,例如汉语、日语等语言。

6 结语

问题生成研究近年以来逐渐成为关注热点,高水平的论文工作层出不穷。本文综述了问题生成的前沿研究进展,从不同角度叙述了神经网络模型的问题生成方法,并对前人研究的不足作了思考,对未来研究趋势作了展望。虽然英语的问题生成已经取得了令人瞩目的成绩,但在汉语方面,还缺乏优良的公开评测数据集,也没有有效的针对汉语问题特点的神经网络模型,汉语的问题生成研究还任重而道远。

参考文献

- [1] Nan Duan, Duyu Tang. Overview of the NLPCC-2017 shared task: Open domain Chinese question answering [C]//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, 2017.
- [2] Guokun Lai, Qizhe Xie, Hanxiao Liu, et al. Race: Large-scale reading comprehension dataset from examinations[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 785-794.
- [3] Mitkov Ruslan, Ha Le An. Computer-aided generation of multiple-choice tests[C]//Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing, 2003.
- [4] Michael Heilman, Noah A Smith. Good question! Statistical ranking for question generation [C]//Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010: 609-617.
- [5] David Lindberg, Fred Popowich, John Nesbit, et al. Generating natural language questions to support learning on-line[C]//Proceedings of the 14th European Workshop on Natural Language Generation, 2013: 105-114.
- [6] Vasile Rus, Brendan Wyse, Paul Piwek, et al. A detailed account of the first question generation shared task evaluation challenge [J]. Dialogue and Discourse, 2012, 3(2):177-204.
- [7] Yifan Gao, Lidong Bing, Piji Li, et al. Generating distractors for reading comprehension questions from real examinations[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019: 6423-6430.
- [8] Xiaorui Zhou, Senlin Luo, Yunfang Wu. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9725-9732.
- [9] Guanliang Chen, Jie Yang, Claudia Hauff, et al. LearningQ: A Large-scale dataset for educational question generation[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.
- [10] Angelica Willis, Glenn Davis, Sherry Ruan, et al. Key phrase extraction for generating educational question-answer pairs [C]//Proceedings of the 6th ACM Conference on Learning@scale, 2019.
- [11] Duyu Tang, Nan Duan, Tao Qin, et al. Question answering and question generation as dual tasks [EB/OL]. CoRR, abs/1706.02027, 2017.
- [12] Nan Duan, Duyu Tang, Peng Chen, et al. Question generation for question answering [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 866-874.
- [13] Shiyue Zhang and Mohit Bansal. Addressing semantic drift in question generation for semi-supervised question answering [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019: 2495-2509.
- [14] Yansen Wang, Chenyi Liu, Minlie Huang, et al. Learning to ask questions in open domain conversational systems with typed decoders[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018:2193-2203.
- [15] Aliannejadi Mohammad, Zamani Hamed, Crestani Fabio, et al. Asking clarifying questions in open-domain information-seeking conversations [C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
- [16] George A Miller. WordNet: An electronic lexical database[M]. MIT Press, 1998.
- [17] Dhole K D, Manning C D. Syn-QG: Syntactic and shallow semantic rules for question generation[C]//Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics, 2020.
- [18] Karin Kipper Schuler. VerbNet: A broad coverage, comprehensive verb lexicon [D]. Ph D. Thesis, University of Pennsylvania, 2005.
- [19] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, et al. Squad: 100,000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Con-

- ference on Empirical Methods in Natural Language Processing, 2016.
- [20] Tri Nguyen, Mir Rosenberg, Xia Song, et al. Ms marco: A human generated machine reading comprehension dataset[J/OL]. ArXiv, preprint arXiv: 1611.09268, 2016.
- [21] Xinya Du, Junru Shao, Claire Cardie. 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.
- [22] Qingyu Zhou, Nan Yang, Furu Wei, et al. Neural question generation from text: A preliminary study [C]//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, 2017:662-671.
- [23] Xingdi Yuan, Tong Wang, Caglar Gulcehre, Aless. Machine comprehension by text-to-text neural question generation [C]//Proceedings of the 2nd Workshop on Representation Learning for NLP, 2017: 15-25.
- [24] Linfeng Song, Zhiguo Wang, Wael Hamza, et al. Leveraging context information for natural question generation[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018:569-574.
- [25] Abigail See, Peter J Liu, Christopher D Manning. 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.
- [26] Xingwu Sun, Jing Liu, Yajuan Lyu, et al. Answer-focused and position-aware neural question generation [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3930- 3939.
- [27] 董孝政, 洪宇, 朱芬红, 等. 基于密令位置信息特征的问题生成[J]. 中文信息学报 2019, 33(8): 93-100.
- [28] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, et al. Improving neural question generation using answer separation[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019:6602-6609.
- [29] Bang Liu, Mingjun Zhao, Di Niu, et al. Learning to generate questions by learning what not to generate [C]//Proceedings of the Web Conference, 2019: 1106-1118.
- [30] Wenjie Zhou, Minghua Zhang, Yunfang Wu. Multi-task learning with language modeling for question generation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019:3392-3397.
- [31] Wenjie Zhou, Minghua Zhang, Yunfang Wu. Question-type driven question generation [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019:6031-6036.
- [32] Jingjing Li, Yifan Gao, Lidong Bing, et al. Improving question generation with to the point context [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019: 3216-3226.
- [33] Xiyao Ma, Qile Zhu, Yanlin Zhou, et al. Improving question generation with sentence-level semantic matching and answer position inferring[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020.
- [34] Xin Jia, Wenjie Zhou, Xu Sun, et al. How to ask good questions? Try to leverage paraphrases [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6130-6140.
- [35] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, et al. Paragraph-level neural question generation with maxout pointer and gated self-attention networks [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3901-3910.
- [36] Preksha Nema, Akash Kumar Mohankumar, Mitesh M Khapra, et al. Let's ask again: Refine network for automatic question generation [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019:3314-3323.
- [37] Yu Chen, Lingfei Wu, Mohammed J Zaki. Natural question generation with reinforcement learning based graph-to-sequence model [C]//Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019.
- [38] Luu Anh Tuan, Darsh J Shah, Regina Barzilay. Capturing greater context for question generation [C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020.
- [39] Wang W, Wei F, Dong L, et al. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers [EB/OL]. ArXiv abs/2002.10957, 2020.
- [40] Bao H, Dong L, Wei F, et al. UniLMv2: Pseudo-masked language models for unified language model pre-training [J/OL]. ArXiv preprint arXiv: 2002.12804. 2020.
- [41] Yan Y, Qi W, Gong Y, et al. ProphetNet: predicting future n-gram for sequence-to-sequence pre-training [C]//Proceedings of the Association for Computational Linguistics; EMNLP 2020:2401-2410.

- [42] Xiao D, Zhang H, Li Y, et al. ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020:3997-4003.
- [43] Jinwen Ma, Wenpeng Hu, Bing Liu, et al. Aspect-based question generation[C]//Proceedings of the International Conference on Learning Representations, 2018.
- [44] Siyuan Wang, Zhongyu Wei, Zhihao Fan, et al. A multi-agent communication framework for question-worthy phrase extraction and question generation [C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019:7168-7175.
- [45] Thomas Scialom, Benjamin Piwowarski, Jacopo Staiano. Self-attention architectures for answer-agnostic neural question generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2019:6027-6032.
- [46] Xinya Du, Claire Cardie. Identifying where to focus in reading comprehension for neural question generation [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2067-2073.
- [47] Sandeep Subramanian, Wang Tong, Xingdi Yuan, et al. Neural models for key phrase detection and question generation[C]//Proceedings of the Workshop on Machine Reading for Question Answering, 2017:78-88.
- [48] Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, et al. Cross-lingual training for automatic question generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2019:4863-4872.
- [49] Wei He, Kai Liu, Jing Liu, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering, 2018: 37-46.
- [50] Zewen Chi, Li Dong, Furu Wei, et al. Cross-lingual natural language generation via pre-training[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020.
- [51] Peng Li, Wei Li, Zhengyan He, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering [EB/OL]. arXiv: 1607.06275. 2016.
- [52] Kishore Papineni, Salim Roukos, Todd Ward, 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.
- [53] Michael Denkowski, Alon Lavie. Meteor universal: Language specific translation evaluation for any target language[C]//Proceedings of the 9th Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2014: 376-380.
- [54] Chin Yew Lin. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of Text Summarization Workshop of the 42th Annual Meeting of the Association for Computational Linguistics, 2004: 74-81.
- [55] Preksha Nema, Mitesh M Khapra. Towards a better metric for evaluating question generation systems [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
- [56] Md Arafat Sultan, Shubham Chandel, Ramón F As-tudillo, et al. On the importance of diversity in question generation for QA[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [57] Qian Yu, Lidong Bing, Qiong Zhang, et al. Review-based question generation with adaptive instance transfer and augmentation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:280-290.
- [58] Pan L, Xie Y, Feng Y, et al. Semantic graphs for generating deep questions [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [59] Ko W J, Chen T Y, Huang Y, et al. Inquisitive question generation for high level text comprehension [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 6544-6555.



吴云芳(1973—),博士,副教授,主要研究领域为文本生成,智能问答,分级阅读。
E-mail: wuyunf@pku.edu.cn



张仰森(1962—),博士,教授,主要研究领域为自然语言处理,网络内容安全。
E-mail: zhangyangsen@163.com