DOI:10.11830/ISSN.1000-5013.202304010

基于模板学习的智能侨情问句生成方法

方昱龙，王泽锦，王华珍，何霆

(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘要： 为解决侨情问句甚少导致训练的侨情语料较少的问题,提出一种基于模板学习的智能侨情问句生成方法。首先,对侨情篇章文本进行包含主题、关系、对象的三元组抽取;其次,构建训练数据集,输入数据由主题和关系构成,输出数据为问句模板;随后,采用以 BERT+LSTM+Attention 为核心算法的 seq2seq 框架,实现问句模板生成;最后,对模板问句进行主题文本替换,从而得到最终的实例化问句。采用 BLEU,ROUGE-*N*,公开问答系统评测及人工评价方式对文中方法进行评价。结果表明:BLEU,ROUGE-*N*,公开问答系统评测及人工评价方式对文中方法的评测结果分别为 0.77,0.67,81%,88%,较基线模型有较大的提升。

关键词： 侨情;问句生成;模板学习;seq2seq;注意力机制

中图分类号： TP 394.4　　**文献标志码：** A　　**文章编号：** 1000-5013(2023)06-0735-08

Intelligent Question Generation Method Based on Template Learning for Overseas Chinese Situation

FANG Yulong, WANG Zejin, WANG Huazhen, HE Ting

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: To address the issue of limited training for overseas Chinese language materials due to the scarcity of overseas Chinese question sentences, a template learning based intelligent overseas Chinese situation generation method is proposed. Firstly, the text of overseas Chinese situation is extracted by triplet including theme, relationship and object. Secondly, the training data set is constructed, its input data are composed of themes and relationships, and its output data is question template. Then, BERT+LSTM+Attention as the core algorithm of seq2seq framework is applied to generate question template. Finally, the template question is replaced by the theme text to get the final instantiated question. BLEU, ROUGE-*N*, public question answering system evaluation and human evaluation method were used to evaluate the proposed method. The results show that the evaluation results of the BLEU, ROUGE-*N*, public question answering system evaluation and human evaluation method are 0.77, 0.67, 81% and 88%, respectively, with significant improvements compared to the baseline model.

Keywords: overseas Chinese situation; question generation; template learning; seq2seq; attention mechanism

侨情是指一个国家或地区的华侨华人、归侨侨眷的情况。当前人们在认知侨情时,常使用百度、谷歌等大众通用问答引擎来进行侨情问答,这是一种促进信息精确获取、提升知识输入的高效率模式。通

收稿日期： 2023-04-18

通信作者： 王华珍(1975-),女,副教授,博士,主要从事人工智能、机器学习、增强现实、虚拟现实等的研究。E-mail:wanghuazhen@hqu.edu.cn。

基金项目： 国家重点研发计划项目(2018YFB1402500);教育部中外语言交流合作中心国际中文教育研究课题(21YH30B);福建省社会科学基金资助项目(FJ2021B110);中央高校基本科研业务费自主项目(TZYB-202005);华侨大学"华侨华人研究"专项经费资助一般项目(HQHRYB2019-01)

过使用侨情问答,人们无需阅读长篇章文本就可以获取到侨情信息,这样一方面满足了现代人快节奏的生活方式,另一方面也促进了人们想了解侨情的动力。然而,通用问答引擎采用的技术是通用型问答模型,而非针对某个特定领域,其垂直性能不足,不能很好地反映该特定领域的认知信息。因此,研发特定侨情问答引擎具有现实意义。而当前侨情问句甚少,用来训练的侨情语料较少,导致如今还未出现高性能的侨情问答模型。因此,利用海量的侨情篇章文本生成问答对以构建问答模型训练集,从而研究问句生成(question generation,QG)的模型,是实现高性能侨情问答的关键前提。

问句生成是人工智能领域中一项重要的研究分支,其研究内容可概括为由各种输入形式的文本中的知识、语义、语法信息等自动生成问句[1]。近年来,问句生成已成为自然语言处理领域中炙手可热的研究点之一。传统研究主要基于规则操作句法树或知识库生成问题,如基于语法的问句生成方法[2-3]、基于语义的问句生成方法[4-6]、基于模板的问句生成方法[7-10]。随着人工智能技术的发展,深度学习模型为问句生成提供了一个由数据驱动、端到端可训练的框架[11]。与传统的基于规则的问句生成方法相比,基于深度学习生成的问句在流畅性和多样性方面都有了很大的提升。在运用深度学习方法进行问题生成时,大多数研究者采用带注意力机制的 seq2seq 方法来实现[12-14]。此外,Liu 等[15]将传统模板法与 seq2seq 结合生成问句;Tuan 等[16]融入上下文关联信息对问句生成进行改进。由此可见,问句生成的主流技术是基于大规模问答对语料的深度学习模型。但在侨情知识问答领域,由于侨情知识具有多国别的地理分散性及主题多样性,因此,无法针对各个侨情主题分别获取海量问答语料对,进而无法满足大数据深度学习模型训练的需求,说明侨情问答具有小规模学习的特性。

综上可知,注意力机制、模板法、seq2seq 模型等技术虽然在问句生成领域具有重要的价值,但上述研究仍存在许多不足,如模板法生成的问句往往句式单一,难以达到提问角度的多元化和复杂化;注意力机制和 seq2seq 模型在没有大量语料支持的前提下,训练出的模型不能满足精度需求等。因此,本文将模板法与 seq2seq 模型相结合,引入基于 Transformer 的双向编码器表达(BERT)预训练模型,并嵌入注意力机制,提出一种基于模板学习的智能侨情问句生成方法。

1 研究设计

1.1 研究框架

针对侨情问句较少导致无法构建大规模问答对语料、进而无法训练高性能侨情问答模型的问题,提出一种将模板法与 seq2seq 模型相结合,引入 BERT[17]预训练模型,并嵌入注意力机制的智能侨情问句生成方法(QGTL-OCS),其研究框架,如图 1 所示。

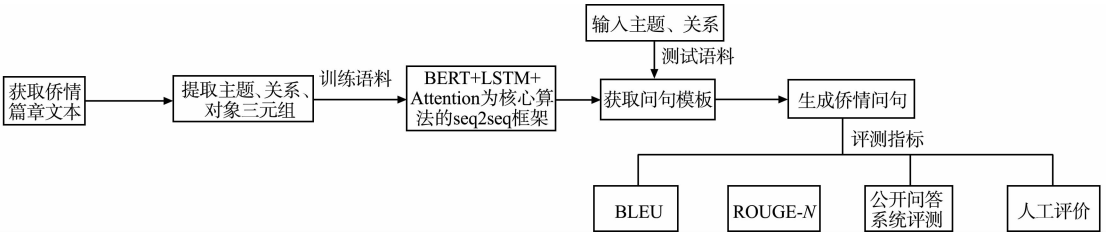

图 1 智能侨情问句生成方法研究框架

Fig. 1 Research framework of intelligent question generation methods for overseas Chinese situation

首先,对侨情篇章文本进行包含主题、关系、对象的三元组抽取;其次,构建训练数据集,输入数据由主题和关系构成,输出数据为问句模板;随后,采用以 BERT+长短时记忆网络(LSTM)+Attention 为核心算法的 seq2seq 框架实现问句模板生成;最后,对模板问句进行主题文本替换,从而得到最终的实例化问句。采用双语评估辅助工具 BLEU、ROUGE-N、公开问答系统评测和人工评价方式对实例化问句进行评测,从而衡量实例化问句的质量。

1.2 侨情问答对三元组抽取

1.2.1 *提取主题、关系和对象三元组* 为了能围绕"侨情"生成一系列紧扣主题的问答对,首先需要获取侨情相关的初始问答对 $B_{QA}=\{Q,A\}$,其中,Q 为问句集合,A 为答句集合,进而从传统的侨情问答篇章文本(问句和答句都是非结构化文本段)中提取核心信息,包括主题、关系和对象三元组,以构成问答

对三元组 $F=(P,R,O)$，其任一问答对三元组定义为 $f=(p,r,o)$。主题和关系来源于问句，对象来源于答案。问句中的主谓关系一般对应主题和关系元组，对象是答句中回应问句的核心实体信息。获得的问答对三元组代表了问答对语料的核心关键信息，去除了问答对语料中冗长的自然语言表达，实现问答对语料的信息聚合。

1.2.2　*构建问句模板*　在问答对中，关系元组代表问句的模式信息，其数量是有限的。问答对中主题、对象元组代表问句的具体信息，具有多元性、广域性的特点，其数量是无限的。因此，每种关系对应一种特定类型的问句，可相应地构建出特定类型的问句模板。结合具体的主题和对象元组，可以将问句模板实例化为问句。问句模板的生成流程包括最优问句筛选和主题抽象化。

由传统的问答对中问句所组成的语料库 Q 中特定关系 $r_i\in R=\{r_1,r_2,\cdots,r_n\}$（$n$ 为关系的总数）所对应的问句语料库 Q'（$Q'\subset Q$）数量庞大且句式多样，因此，应从 Q' 中筛选出一个最优问句 q_i。首先，筛选包含疑问词并且主语、谓语、宾语齐全的问句集 Q''；再从 Q'' 中选择谓语词是关系词 r_i，且长度最短、字数最少的问句作为最优问句 q_i。最优问句 q_i 中的主题和关系组成了问句两元组 $g_i=(p_i^{\text{best}},r_i)$。

最优问句 q_i 中包含具体的主题，即主题是问题模板实例化的数据。为了生成问句模板，进一步将 q_i 中的实例化主题数据抽象成主题概念，即用〈SUB〉标签代替实例化主题，获得问句模板 s_i。抽取最优问句并形成模板问句的过程示意图，如图 2 所示。

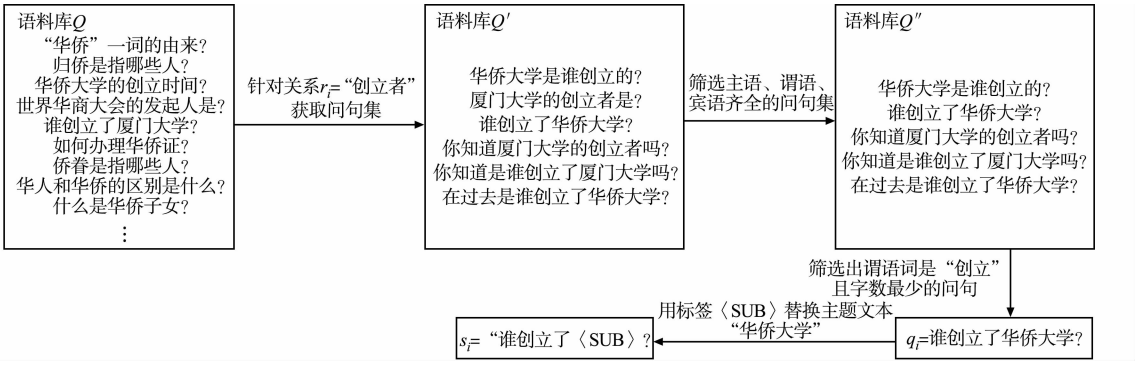

图 2　抽取最优问句并形成模板问句的过程示意图

Fig. 2　Schematic diagram of process for extracting optimal question and forming template question

首先，从语料库 Q 中针对关系“创立者”获取问句集 Q'，包含“华侨大学是谁创立的?”“厦门大学的创立者是?”“谁创立了华侨大学”“你知道厦门大学的创立者吗?”等多条问句；其次，采用疑问词词典法和依存句法分析技术筛选主语、谓语、宾语成分齐全的句子组成问句集 Q''，如问句“厦门大学的创立者是?”缺少宾语成分，无法入选 Q''；再次，从问句集 Q'' 中筛选出谓语词是“创立”且字数最少的问句，即“谁创立了华侨大学?”作为最优问句 q_i；最后，用〈SUB〉标签代替实例化主题“华侨大学”，获得问句模板“谁创立了〈SUB〉?”。每个关系 $r_i\in\{r_1,r_2,\cdots,r_n\}$ 都可以生成一个对应的问句模板，最终可以获得 n 个对应的问句模板。

1.3　问句模板生成算法设计

1.3.1　*算法思想*　从语料到模板的过程是将传统问答对语料转换为三元组，在同关系的问句中挑选最优问句，抽象化后得到问句模板。从问答对三元组到问句模板的映射过程采用深度学习模型实现。因此，训练集 $D_{\text{train}}=\{D_{\text{in}},D_{\text{out}}\}$，其中，$D_{\text{in}}=\{g_i\mid i=1,\cdots,n\}$，$D_{\text{out}}=\{s_i\mid i=1,\cdots,n\}$。采用 seq2seq 模型实现深度学习，是一个序列经过映射得到另一个序列的过程，可看作是一种 Encoder-Decoder[18] 结构。从问答对三元组到问句模板生成算法的模型框架，如图 3 所示。当输入一个主题与关系元组后，seq2seq 模型会生成一个带有 SUB 标记的问句模板。

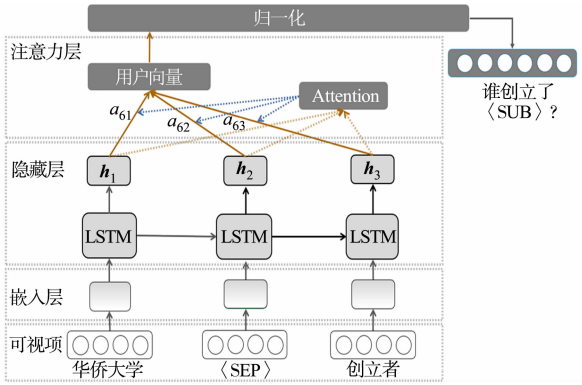

图 3　算法模型框架

Fig. 3　Framework of algorithm model

1.3.2 *算法原理* 已知输入数据 $D_{\text{in}}=\{g_i=(p_i^{\text{best}},r_i)\mid i=1,\cdots,n\}$,设计 seq2seq 模型 Φ 实现问句模板的生成,即

$$s_i=\Phi(g_i)。$$

式中:$s_i\in D_{\text{out}}=\{s_i\mid i=1,\cdots,n\}$。

采用 BERT+LSTM+Attention 实现算法模型 Φ,算法原理如下。

1) 文本的向量表示。获取 D_{in},D_{out} 和标识符〈SUB〉,〈SEP〉所组成的集合的字典大小为 $|V|$。对问句二元组 g_i 进行 BERT 编码获得嵌入矩阵 $\boldsymbol{X}_i\in\mathbf{R}^{|V|\times J}$($J$ 为嵌入词的长度)。此时,$J=3$,$\boldsymbol{X}_{i,1}$ 为 p_i^{best} 的 BERT 编码向量,$\boldsymbol{X}_{i,2}$ 为插入标识〈SEP〉的 BERT 编码向量,$\boldsymbol{X}_{i,3}$ 为 r_i 的 BERT 编码向量。在 $\boldsymbol{X}_{i,1}$ 和 $\boldsymbol{X}_{i,3}$ 之间插入标识〈SEP〉,便于在 Encoder 结构中分离主题和关系。对问句模板 s_i 进行 BERT 编码,获得嵌入矩阵 $\boldsymbol{Y}_i\in\mathbf{R}^{|V|\times K}$($K$ 为问句模板的长度),用于 Decoder 模块中。

2) 算法优化过程。根据 LSTM 算法对输入数据 $\boldsymbol{x}_t\in\boldsymbol{X}_i$ 进行运算,即

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\text{f}}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\text{f}}),\tag{1}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\text{t}}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\text{i}}),\tag{2}$$

$$\hat{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_{\text{C}}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\text{C}}),\tag{3}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\odot\hat{\boldsymbol{C}}_t,\tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{\text{o}}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\text{o}}),\tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{C}_t)。\tag{6}$$

式(1)~(6)中:$\boldsymbol{h}_t$ 是在时刻 t 的隐藏状态;$\boldsymbol{f}_t$,$\boldsymbol{i}_t$,$\boldsymbol{o}_t$ 分别为 LSTM 单元的遗忘门、输入门、输出门;$\hat{\boldsymbol{C}}_t$,$\boldsymbol{C}_t$ 分别为 LSTM 单元在时刻 t 的历史状态和当前状态;$\boldsymbol{W}_{\text{f}}$,$\boldsymbol{W}_{\text{t}}$,$\boldsymbol{W}_{\text{C}}$,$\boldsymbol{W}_{\text{o}}$,$\boldsymbol{b}_{\text{f}}$,$\boldsymbol{b}_{\text{i}}$,$\boldsymbol{b}_{\text{C}}$,$\boldsymbol{b}_{\text{o}}$ 均为超参数。

为了进一步提升 seq2seq 的模型性能,使用一个带有 Attention 机制的 LSTM 算法实现 Decoder。Decoder 在时刻 t 输出的词汇概率向量 $\boldsymbol{P}(\boldsymbol{y}_t\mid\boldsymbol{H},\boldsymbol{o}_t)$ 是由 t 时刻 LSTM 的输出向量 $\boldsymbol{o}_t$(式(5))和 Encoder 的 t 时刻前的所有隐藏状态 $\boldsymbol{H}=\{\boldsymbol{h}_t\}_{t=1}^{t-1}$,可以表示为

$$\boldsymbol{P}(\boldsymbol{y}_t\mid\boldsymbol{H},\boldsymbol{o}_t)=\text{softmax}(\boldsymbol{W}_{\text{s}}\odot\tanh(\boldsymbol{W}_{\text{t}}[\boldsymbol{o}_t,\boldsymbol{a}_t]))。\tag{7}$$

式(7)中:$\boldsymbol{a}_t$ 是由 Encoder 每一时刻隐藏状态加权表示的 Attention 向量,反映了 Encoder 任意时刻隐藏层值对 Decoder 输出的影响程度;$\boldsymbol{W}_{\text{s}}$,$\boldsymbol{W}_{\text{t}}$ 为待学习的参数。

采用 align 函数计算 $\boldsymbol{o}_t$ 和 $\boldsymbol{H}$ 的相关系数 α_t,其公式分别为

$$\alpha_t=\text{align}(\boldsymbol{o}_t,\boldsymbol{H}),\tag{8}$$

$$\alpha_{t_i}=\frac{\exp(z(\boldsymbol{o}_t,\boldsymbol{h}_i))}{\sum_{j=1}^{J}\exp(z(\boldsymbol{o}_t,\boldsymbol{h}_j))}。\tag{9}$$

式(9)中:$z(\boldsymbol{o}_t,\boldsymbol{h}_i)$是用来计算 $\boldsymbol{o}_t$ 和任意时刻隐藏层状态 $\boldsymbol{h}_i$ 的数量积,其公式为

$$z(\boldsymbol{o}_t,\boldsymbol{h}_i)=\tanh(\boldsymbol{W}_{\text{p}}\boldsymbol{h}_i)\odot\tanh(\boldsymbol{W}_{\text{q}}\boldsymbol{o}_t)。\tag{10}$$

式(10)中:$\boldsymbol{W}_{\text{p}}$,$\boldsymbol{W}_{\text{q}}$ 为待学习的参数。

1.3.3 *算法伪代码* 基于模板学习的智能侨情问句生成方法的伪代码如下。

输入训练集 $D_{\text{train}}=\{D_{\text{in}},D_{\text{out}}\}$,其中,$D_{\text{in}}=\{g_i\mid i=1,\cdots,n\}$,$D_{\text{out}}=\{s_i\mid i=1,\cdots,n\}$,以及一条待测试的问句二元组 $g_{\text{test}}=(t_{\text{test}},r_{\text{test}})$。

输出一个问句 $q_{g_{\text{test}}}$。

1) 文本向量化表达:获取 D_{in},D_{out} 和标识符〈SUB〉,〈SEP〉所组成的集合的字典大小为 $|V|$。对问句二元组 g_i 进行 BERT 编码获得嵌入矩阵 $\boldsymbol{X}_i\in\mathbf{R}^{|V|\times J}$($J$ 为嵌入词的长度)。对问句模板 s_i 进行 BERT 编码获得嵌入矩阵 $\boldsymbol{Y}_i\in\mathbf{R}^{|V|\times K}$($K$ 为问句模板的长度);

2) 初始化 seq2seq 模型;

3) while 不满足终止条件

选择一个 batch 训练语料;

运行 seq2seq 模型优化算法;

4）保存侨情问句生成模型 Φ；

5）将待测试的问句二元组 g_{test} 输入到模型 Φ 中，生成问句模板 s_{test}；

6）将 s_{test} 中主题标记〈SUB〉用 t_{test} 进行替换，获得实例化的侨情问句 $q_{g_{\text{test}}}$。

2　实证分析

2.1　侨情篇章文本收集与处理

网络爬虫是一种按照开发者所制定的规则，自动嗅探、发现网络上存在的资源并进行获取的程序。利用网络爬虫技术，通过设定搜索条件，就能高效地获取所需内容。以“华侨”“华人”“侨胞”或“涉侨”等词汇为搜索关键字，爬取百度知道平台网页内容，获取侨情相关的问答篇章文本。最终得到初始侨情问答篇章文本规模为 17 656 条。

由于爬虫获得的问答篇章文本篇幅冗长、噪声过多，因此，对侨情问答篇章文本使用人工编辑方式进行预处理。如问答篇章文本中的问句：“一般认为，1729 年荷印巴达维亚华侨成立的明诚书院是华侨教育的开端。”经人工处理为：“华侨教育从什么时候开始?”经过预处理后，获得高质量精简问答对 17 656个。进一步通过手工筛选与编辑处理方式进行主题、关系、对象三元组的抽取，得到预处理后的 17 656个问答对三元组，结果如表 1 所示。

表 1　预处理后的问答对及其三元组

Tab. 1　Preprocessed question and answer pairs and triples

问句	答案	问答对三元组
华裔和华侨的区别是什么?	定居在国外的华人是否具有中国国籍	华裔和华侨 ‖ 区别 ‖ 定居在国外的华人是否具有中国国籍
华侨大学创办时间是什么时候?	1960 年 11 月 1 日	华侨大学 ‖ 创办时间 ‖ 1960 年 11 月 1 日
华侨大学是什么类别的大学?	公立大学	华侨大学 ‖ 类别 ‖ 公立大学
华侨大学的校训是什么?	会通中外、并育德才	华侨大学 ‖ 校训 ‖ 会通中外、并育德才

2.2　算法实验设置

对 17 656 个问答对及其对应的三元组采用依存句法分析等自然语言处理技术获得 5 438 个问句模板，从而获得 5 438 条学习语料。进一步使用 seq2seq 模型来实现模板学习。模型超参数设置参考了 Liu 等[15]的工作，具体如下：隐藏层状态维数为 500，batch 大小为 32，epoch 大小为 50，l_{r} 为 0.000 5。针对 5 438 条学习语料进行训练集和测试集的划分，分别为 3 807 条与 1 631 条。

2.3　算法实验结果与分析

采用双语评估辅助工具 BLEU、ROUGE-N 指标、公开问答系统评测及人工评价方式对 QGTL-OCS 模型模型进行性能评价。

BLEU[19]的核心在于比较生成问句文本和参考问句文本之间 n-gram 的重合程度，重合程度越高，则认为生成问句文本的质量越高，其计算式为

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} W_n \times \lg P_n\right), \tag{11}$$

$$\text{BP} = \begin{cases} 1, & l_{\text{c}} > l_{\text{r}}, \\ \exp\left(1 - \dfrac{l_{\text{r}}}{l_{\text{c}}}\right), & l_{\text{c}} \leqslant l_{\text{r}}。 \end{cases} \tag{12}$$

式(11)，(12)中：l_{c} 为生成问句文本长度；l_{r} 为参考问句的文本长度；P_n 为 n-gram 的精度；N 为 n-gram 取值；W_n 为 n-gram 的权重，$W_n = N/4$，文中 N 取 4；BP 为惩罚因子。

此外，通过 ROUGE-N[20]指标基于召回率对生成问句进行评估，其计算式为

$$\text{ROUGE-}N = \frac{\sum_{\text{word}\{\text{References}\}} \sum_{\text{gram}_N \text{word}} \text{Count}_{\text{matc}}(\text{gram}_N)}{\sum_{\text{word}\{\text{References}\}} \text{Count}(\text{gram}_N)}。 \tag{13}$$

式(13)中：$\text{Count}_{\text{matc}}(\text{gram}_N)$ 表示生成问句文本 word 和参考问句文本 References 重合的的 n-gram 个数；$\text{Count}(\text{gram}_N)$ 表示参考问句文本的 n-gram 个数。

公开问答系统评测是一种面向实践应用的评测策略,采用百度人工智能(AI)开放平台推出的通用闲聊机器人(https://ai.baidu.com/unit/v2/static/socialbot)进行效果评估。通过将测试集中的主题和关系数据送入模型生成问句,把所得问句作为通用闲聊机器人的输入,得到对应的答案。再将答案与测试集中三元组 $F=(T,R,O)$ 的对象 O 进行对比,若对象 O 被包含于所得答案中,则判定模型产生的问句为正确。最终评测得分为测试正确的结果与测试总问句数的比值。由于通用闲聊机器人可以覆盖最大会话空间,因此,公开问答系统评测方法具有全面性。但该方法依赖于公开问答系统的性能,评测稳定性较差。

人工评价是指侨情专业人员对算法最终输出的问句进行语法正确性、语义完整性和问句所属类型等进行评测的方式。该评价方法虽然会耗费极高的人力,但评判结果准确,且真实可信。

根据上述4种评测方法,得到模型评价结果,如表2所示。由表2可知:文中所提算法的BLEU指标评测结果为0.77,表明该算法产生的问句模板比较贴合相同关系中的最优问句抽象所得的模板;该算法的ROUGE-N指标评测结果为0.67,表明该算法生成的问句文本与参考的问句文本存在一定的差距,但处于可接受范围内;公开问答系统评测得到的准确率为81%,体现了该算法生成的问句的准确性;人工评价的准确率达到了88%,体现问答系统反馈的有效性。

表2 模型评测结果

Tab.2 Model evaluation results

评测方法	评测结果
BLEU	0.77
ROUGE-N	0.67
公开问答系统评测	81%
人工评价	88%

综上可知,基于模板学习的智能侨情问句生成方法继承了模板法的特性,并且在输入指向准确、针对性强的关系的前提下,可以生成易于理解、符合语言习惯、切合主题的问句。

2.4 对比实验与消融实验

通过对比实验与消融实验进一步评测QGTL-OCS模型的有效性。主要模型如下。

1)循环神经网络(RNN)[21]和BERT-RNN模型。RNN主要用于处理序列数据,具有记忆能力和对时间依赖关系的建模能力。BERT-RNN是以BERT为文本嵌入模型的RNN架构。

2)LSTM[22]和BERT-LSTM模型。LSTM是长短期记忆网络,用于解决传统RNN在长序列训练中的梯度消失和梯度爆炸问题;LSTM引入了门控机制,通过控制信息的流动和遗忘,有效地捕捉序列中的长期依赖关系。BERT-LSTM是以BERT为文本嵌入模型的LSTM模型。

3)双向长短期记忆网络(BiLSTM)[23]和BERT-BiLSTM模型。BiLSTM是一种递归神经网络的变体,用于处理序列数据。与传统的LSTM不同,BiLSTM在处理序列同时考虑了过去和未来的上下文信息,以更全面地捕捉序列的特征。BERT-BiLSTM是以BERT为文本嵌入模型的BiLSTM模型。

4)QGTL-OCS$_{DE}$消融模型。删除BERT-QGTL-OCS模型中的三元组抽取模块,直接将文本语料作为输入,并用LSTM+Attention来实现问句生成。

5)QGTL-OCS$_{DA}$消融模型。删除QGTL-OCS模型中的Attention模块,仅用三元组抽取+LSTM来实现问句生成。

6)QGTL-OCS$_{DB}$消融模型。删除QGTL-OCS模型中的BERT模块,仅用三元组抽取+LSTM+Attention来实现问句生成。

7)QGTL-OCS模型。采用三元组抽取+BERT+LSTM+Attention来实现问句生成。

不同模型的对比结果,如表3所示。由表3可知:相比单一的可用于处理序列数据的RNN,LSTM,BiLSTM模型,QGTL-OCS模型采用三元组抽取并结合BERT与Attention机制生成问句模板,进一步生成问句的这种技术路线更加有效,得到的问句质量有较大的提升。究其原因可能是三元组抽取获得了文本语料精确的主题和关系信息,BERT更准确地获得了文本的向量表达,而Attention能对不

表3 不同模型的对比结果

Tab.3 Comparison results of different models

模型	BLEU	ROUGE-N
RNN	0.40	0.24
BERT-RNN	0.44	0.27
LSTM	0.57	0.39
BERT-LSTM	0.59	0.43
BiLSTM	0.60	0.44
BERT-BiLSTM	0.62	0.47
QGTL-OCS$_{DE}$	0.59	0.45
QGTL-OCS$_{DA}$	0.60	0.44
QGTL-OCS$_{DB}$	0.58	0.40
QGTL-OCS	0.77	0.67

同的信息赋予不同的权重，使模型在生成问句模板时更加精确。对比消融模型可知：三元组抽取模块、BERT 模块与 Attention 模块都不同程度地增强了生成问句的质量。

2.5　样例分析

根据所提算法原理，输入待测试的主题和关系，则输出一个侨情问句。模型测试实例结果，如表 4 所示。由表 4 可知：文中算法可针对侨情实体信息生成语法正确、语义完整、类型丰富的侨情问句。

表 4　模型测试实例结果

Tab. 4　Example results of model testing

输入主题	输入关系	输出问句
陈嘉庚	籍贯	陈嘉庚是哪里人？
侨眷	定义	侨眷定义是什么？
李光耀	出生	李光耀出生是什么时候？
华侨联合会	创立者	谁创立了华侨联合会？
美国留学生数量最多	原因	美国留学生数量最多原因是什么？
新加坡	人口	新加坡人口有多少？

3　结论

由于侨情知识具有多国别的地理分散性及主题多样性，无法针对各个侨情主题分别获取海量问答语料对，导致当前侨情问句甚少，用来训练的侨情语料较少，还没有出现高性能的侨情问答模型。因此，提出一种基于模板学习的智能侨情问句生成方法（QGTL-OCS），弥补了传统模板法生成问句的不足。将模板法与 seq2seq 模型相结合，并嵌入注意力机制的小规模学习的技术，实现智能高效的侨情问句生成。该方法能从小规模实例训练集中学习出通用问句模板，进而实例化数量多、类型丰富的侨情问句，从而构建具有推广和应用价值的侨情问答系统。

该研究大大促进了人们对侨情的认知和把握，也有利于政府对侨情信息的管理和侨务工作的实施。研究还进一步提升了问答对三元组抽取的自动化程度，可构建更合理的生成问句评估指标。另外，QGTL-OCS 模型的输出结果为问句，这些问句可服务于侨情问答系统。后续工作将结合侨情的领域特点，开发定制化的问答系统。

参考文献：

[1] RUS V，WYSE B，PIWEK P，*et al*. The first question generation shared task evaluation challenge[C]// Proceedings of the 6th International Natural Language Generation Conference. Trim Castle：School of Computer Science and Statistics，2010：7-9. DOI：10.5555/2187681.2187740.

[2] HEILMAN M，SMITH N A. Extracting simplified statements for factual question generation[C]// Proceedings of QG 2010：The Third Workshop on Question Generation. Pittsburgh：Springer-Verlag，2010：11-20.

[3] ALI H，CHALI Y，HASAN S A. Automatic question generation from sentences[C]// In Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles. Montréal：Université de Montréal，2010：213-218.

[4] MANNEM P，PRASAD R，JOSHI A. Question generation from paragraphs at UPenn：QGSTEC system description [C]// Proceedings of QG 2010：The Third Workshop on Question Generation. Pittsburgh：Springer-Verlag，2010：84-91.

[5] SHANTHI BALA P，AGHILA G. Q-genesis：Question generation system based on semantic relationships[C]// Proceedings of ICBDCC18. Coimbatore：Springer，2019：509-517. DOI：10.1007/978-981-13-1882-5_44.

[6] KWATE DASSI L. Semantic-based self-critical training for question generation[EB/OL]. (2021-08-26)[2023-04-10]. https：//doi.org/10.48550/arXiv.2108.12026.

[7] DIVATE M，SALGAONKAR A. Automatic question generation approaches and evaluation techniques[J]. Current Science，2017，113(9)：1683-1691. DOI：10.18520/cs/v113/i09/1683-1691.

[8] CURTO S，MENDES A C，COHEUR L. Question generation based on lexico-syntactic patterns learned from the web[J]. Dialogue and Discourse，2012，3(2)：147-175. DOI：10.5087/dad.2012.207.

[9] LINDBERG D，POPOWICH F，NESBIT J，*et al*. Generating natural language questions to support learning on-line

[C]// Proceedings of the 14th European Workshop on Natural Language Generation. Sofia: Association for Computational Linguistics, 2013: 105-114.

[10] CHALI Y, GOLESTANIRAD S. Ranking automatically generated questions using common human queries[C]// International Natural Language Generation Conference. Edinburgh: Association for Computational Linguistics, 2016: 217-221. DOI: 10.18653/v1/W16-6635.

[11] WANG Liuyin, XU Zihan, LIN Zibo, *et al*. Answer-driven deep question generation based on reinforcement learning [C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: Association for Computational Linguistics, 2020: 5159-5170. DOI: 10.18653/v1/2020.coling-main.452.

[12] DU Xinya, SHAO Junru, CARDIE C. Learning to ask: Neural question generation for reading comprehension[EB/OL]. (2017-04-29)[2023-03-19]. https://doi.org/10.48550/arXiv.1705.00106.

[13] ZHOU Qingyu, YANG Nan, WEI Furu, *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. DOI: 10.1007/978-3-319-73618-1_56.

[14] YUAN Xingdi, WANG Tong, GULCEHRE C, *et al*. Machine comprehension by text-to-text neural question generation[C]// Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver: Association for Computational Linguistics, 2017: 15-25. DOI: 10.18653/v1/W17-2603.

[15] LIU Tianyu, WEI Bingzhen, CHANG Baobao, *et al*. Large-scale simple question generation by template-based seq2seq learning[C]// National CCF Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2017: 75-87. DOI: 10.1007/978-3-319-73618-1_7.

[16] TUAN L A, SHAH D, BARZILAY R. Capturing greater context for question generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 9065-9072. DOI: 10.1609/aaai.v34i05.6440.

[17] DEVLIN J, CHANG Mingwei, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2023-03-19]. https://doi.org/10.48550/arXiv.1810.04805.

[18] CHO K, MERRIENBOER B V, GULCEHRE C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014: 1724-1734. DOI: 10.3115/v1/D14-1179.

[19] 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: ACL, 2002: 311-318. DOI: 10.3115/1073083.1073135.

[20] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona: ACL, 2004: 74-81.

[21] WERBOS P J. Generalization of backpropagation with application to a recurrent gas market mode[J]. Neural Networks, 1988, 1(4): 339-356. DOI: 10.1016/0893-6080(88)90007-X.

[22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[23] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681. DOI: 10.1109/78.650093.

(**责任编辑**: 黄晓楠 **英文审校**: 吴逢铁)