软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(11):4239–4250 [doi: 10.13328/j.cnki.jos.006340]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于对话结构的多轮对话生成模型*

姜晓彤, 王中卿, 李寿山, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 王中卿, E-mail: wangzq@suda.edu.cn

摘 要: 目前, 多轮对话生成研究大多使用基于 RNN 或 Transformer 的编码器-解码器架构. 但这些序列模型都未能很好地考虑到对话结构对于下一轮对话生成的影响. 针对此问题, 在传统的编码器-解码器模型的基础上, 使用图神经网络结构对对话结构信息进行建模, 从而有效地刻画对话的上下文中的关联逻辑. 针对对话设计了基于文本相似度的关联结构、基于话轮转换的关联结构和基于说话人的关联结构, 利用图神经网络进行建模, 从而实现对话上下文内的信息传递及迭代. 基于 DailyDialog 数据集的实验结果表明, 与其他基线模型相比, 该模型在多个指标上有一定的提升. 这说明使用图神经网络建立的模型能够有效地刻画对话中的多种关联结构, 从而有利于神经网络生成高质量的对话回复.

关键词: 图神经网络; 对话生成; 人机对话; 对话结构

中图法分类号: TP391

中文引用格式: 姜晓彤, 王中卿, 李寿山, 周国栋. 基于对话结构的多轮对话生成模型. 软件学报, 2022, 33(11): 4239–4250. http://www.jos.org.cn/1000-9825/6340.htm

英文引用格式: Jiang XT, Wang ZQ, Li SS, Zhou GD. Multi-turn Dialogue Generation Model with Dialogue Structure. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4239–4250 (in Chinese). http://www.jos.org.cn/1000-9825/6340.htm

Multi-turn Dialogue Generation Model with Dialogue Structure

JIANG Xiao-Tong, WANG Zhong-Qing, LI Shou-Shan, ZHOU Guo-Dong

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Recent research on multi-turn dialogue generation has focused on RNN or Transformer-based encoder-decoder architecture. However, most of these models ignore the influence of dialogue structure on dialogue generation. To solve this problem, this study proposes to use graph neural network structure to model the dialogue structure information, thus effectively describing the complex logic within a dialogue. Text-based similarity relation structure, turn-switching-based relation structure, and speaker-based relation structure are proposed for dialogue generation, and graph neural network is employed to realize information transmission and iteration in dialogue context. Extensive experiments on the DailyDialog dataset show that the proposed model consistently outperforms other baseline models in many indexes, which indicates that the proposed model with graph neural network can effectively describe various correlation structures in dialogue, thus contributing to the high-quality dialogue response generation.

Key words: graph neural network; dialogue generation; human-machine dialogue; dialogue structure

对话生成是自然语言处理中的关键任务之一, 迄今已有许多相关应用进入了人类的日常生活中, 如智能电商客服“阿里小蜜”, 手机等终端设备中的智能生活助理 Siri、Cortana 以及跨平台的智能机器人微软小冰等. 因此, 对话生成的相关研究成为近期的一个研究热点.

近年来, 许多对话系统的相关研究采用了生成式模型. 在生成式模型中, 较为传统的研究方法是使用基于层次性的循环神经网络(RNN)的序列到序列结构作为模型架构[1,2], 然而受限于 RNN 的梯度衰减问题, 此类

* 基金项目: 国家自然科学基金(61806137, 61702149)
收稿时间: 2020-11-12; 修改时间: 2021-01-10, 2021-02-28; 采用时间: 2021-03-23; jos 在线出版时间: 2021-04-20

模型难以记忆对话历史中的长期关键内容，倾向于生成“我不知道”“抱歉”等无意义的通用单一回复. 为了解决这个问题，许多研究在此基础上进行了改进，如加入变分推断[3]、融合记忆网络[4]、修改变分推断中的先验分布[5,6]等. 还有一些方法考虑到了强调历史信息中的关键内容，如在上述方法的基础上加入注意力机制[7]、使用 Transformer 来代替序列到序列中的传统 RNN 结构[8]等.

但从实际效果上来看，利用前述方法生成的对话回复仍然或多或少地存在着一些缺陷，难以保障生成回复的多样性和流畅性. 一部分原因是：目前的大部分生成式模型都是扁平的序列到序列结构，不易刻画对话文本中的内部结构，难以把握对话内在的逻辑关系. 而对话的内部结构作为对话文本内的潜在信息，对于对话生成有着极大的帮助. 为了有效利用对话的内部结构，本文设计了以下 3 种对话内部的关联结构.

- 基于文本相似度的关联结构. 话轮是在对话过程中，某一说话人连续说出的一番话. 这种结构是以对话文本中的每一条话轮作为基本元素、以时间为序的线性结构，它描述了对话历史文本内容之间的逻辑关联信息. 以图 1 中的对话为例，前 4 个话轮中存在多个重复出现的词语，如“book”“law”等，这些重复出现的关键字可以帮助模型感知到对话内部的词汇衔接现象，并确定对话的主要内容与法律书籍相关. 对话历史文本承载了较为全面且整体的信息，我们有必要使用基于文本相似度的关联结构来帮助对话回复的生成.
- 基于话轮转换的关联结构. 这种结构是以每个话轮作为基本元素、以话轮转换过程作为逻辑关系的线性结构. 话轮转换是说话人更替的过程，话轮转换前后的两个话轮往往是互相依赖的，比如“问与答”“道歉与接受”“祝贺与感谢”等[9]. 在示例的两个话轮转换过程中，每两个转换过程前后的话轮都构成了独立的问答结构. 第 1 组话轮转换前后的话轮是关于来访原因的问答，第 2 组话轮转换前后的话轮是关于书籍信息的问答. 转换前后两个话轮的互相依赖性，使得该种结构能够有效地辅助对话生成.
- 基于说话人的关联结构. 该结构以每个话轮作为基本元素，在来自同一说话人的话轮之间建立逻辑关系. 在图 1 的示例中，说话人 *A* 一直在试图获取书籍信息，承担一个发问的角色；而说话人 *B* 一直在提供说话人 *A* 所需求的书籍信息，承担一个回应的角色. 说话人的身份、性格信息不同，使得他们各自具有独特的话语风格、特点. 因此，基于说话人的关联结构可以将说话人的因素纳入到对话回复生成时的考虑中.

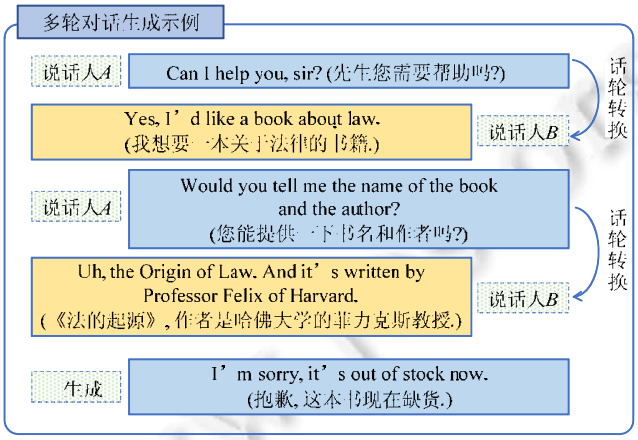


图 1　多轮对话生成示例图

本文提出利用上述 3 种关联结构进行对话建模. 图是一种刻画事物之间关联关系的数据结构，可以利用节点和边来表征复杂的关联结构. 相比扁平的序列化的建模方式，图结构可以更自然地表示上述的对话关联结构，以便更好地理解对话内在逻辑. 因此，我们选择利用图神经网络对对话内部结构进行建模，以期提升对

话回复生成的质量与效果. 模型的主要结构为: 第 1 步, 编码, 使用双向 LSTM 编码器分别将对话历史中的每一条话轮进行编码, 从而构成对话上下文表示; 第 2 步, 构建对话内部的关联关系, 本模型使用图神经网络将对话按照不同的关联结构进行建模. 具体地, 将对话抽象为图结构, 每个图都代表一个对话, 图中的某一顶点代表对话中的某一话轮, 边代表话轮之间的关系. 针对不同的关联结构, 模型采用了不同的建边方法, 建立了 3 种关系图, 分别是文本相似度关系图、话轮转换关系图、说话人关系图. 最后, 图卷积和图池化层对图结构中的信息进行读出, 再把读出后的表示送入解码器中. 之后, 通过集束搜索得到解码语句, 并最终将其作为模型生成的对话回复内容.

综上, 本文提出了一个基于对话结构的多轮对话生成模型, 模型主要由编码器、图神经网络层和解码器构成. 我们在开放域数据集 DailyDialog 上进行了多次重复实验, 并且与其他生成式模型的实验结果进行了比较. 同时, 本文也对各种构图方法及其组合做了分别实验. 评测结果表明, 本模型在多个指标上能够优于其他基线模型.

1 相关工作

对话生成任务通常可以根据是否考虑历史对话信息而分为单轮对话生成任务和多轮对话生成任务. 近年来, 对话生成任务的热点逐步从单轮对话任务转向为多轮对话任务. 这是因为多轮对话更贴合现实对话的特征, 符合智能客服、智能家居机器人等商业落地项目的需求; 并且, 具有“多内容”“多限制”特征的多轮对话能够为研究人员带来更大的挑战[10].

关于多轮对话生成任务的研究可以大致分为生成式模型和检索式模型两种.

- 检索式模型主要基于文本匹配模型, 试图在预先构造的会话历史存储库中找到最相关的上下文-响应对. 这种模型具有可解释性强、可控性强的特点, 但检索性能受响应存储库大小的限制, 同时也难以生成新的内容[11,12].
- 而生成式模型无须会话历史存储库, 可以根据对话历史生成全新的内容, 但生成内容的可解释性和可控性仍有待提高[13].

多轮对话的生成式模型主要基于端到端的序列到序列模型[14]. Vinyals 等人[1]首次提出了使用序列到序列模型来进行多轮对话生成任务, 将对话生成任务转变为翻译任务. Serban 等人[2]提出使用层次序列到序列模型(hierarchical recurrent encoder decoder, HRED), 在话语编码之上增加了对上下文的编码. Serban 等人[15]还提出利用多尺度分析循环神经网络, 将 HRED 原有的上下文编码与粗粒度内容编码结合, 更好地捕捉整个对话历史的粗粒度特征. 但是 HRED 相关模型受限于循环神经网络的梯度衰减问题, 模型倾向于生成通用的“我不知道”“抱歉”等无意义的安全回复.

为了解决这个问题, Serban 等人[3]结合变分自编码器[16]的思想, 将以高斯分布为先验分布的潜变量引入到 HRED 的解码器中, 提出了潜变量层次序列到序列模型(latent variable hierarchical recurrent encoder-decoder, VHRED), 从而增强了模型中语义层面的随机因素, 有助于生成多样性的、高质量的回复. 随后, 有许多基于 VHRED 模型的改进[17], Zhao 等人[18]针对对话生成中的一对多映射问题, 提出了利用条件变分自编码器来帮助生成多样性的回复. Chen 等人[4]进一步考虑了回复内容与历史内容的远距离依赖问题, 提出了层次变分记忆网络模型(hierarchical variational memory network, HVMN), 其基本思路是: 将记忆网络(memory network)引入到了 VHRED 中, 通过记忆单元与潜变量共同实现读取和记忆相关上下文内容. Shen 等人[19]的对话语义关系模型(CSRR)从对话的实际场景出发, 强调了对话历史中的询问部分(query)与回复是最紧密相关的. 为此, CSRR 模型在 VHRED 模型的基础上加入了多个潜变量, 用以分别捕捉询问、回复和询问回复对(query-response)的内在逻辑. 以上的 VHRED 模型都以高斯分布作为潜变量分布, Zeng 等人[5]认为: 对称的高斯分布虽然方便分析, 但无法有效地表达潜变量的复杂性, 提出了使用灵活结构的狄利克雷分布来代替传统的高斯分布假设. 类似地, Gu 等人[6]使用条件瓦瑟斯坦自编码器, 提出了面向多模态的对话回复生成模型 DialogWAE.

前面所述的模型主要以循环神经网络(recurrent neural network, RNN)为关键模块, Xing 等人[7]首先将注意力机制列入了模型组件范围, 他们认为, 注意力机制可以更为显式地考虑到历史话语、字词对于回复话语的重要性, 于是提出了加入注意力机制[20]的层次循环注意力网络 HRAN. Transformer 模型[21]的出现, 鼓励了自注意力网络在生成式模型中的进一步应用[22], Zhang 等人[8]提出了借助 Transformer 结构的 ReCoSa 模型, 该模型利用 Transformer 结构来解决远距离的依赖性, 并且可以显式地筛选对话历史中的重要部分. Bao 等人[23]还结合了 Bert[24]等预训练模型的思想, 以对话回复生成和对话动作识别为预训练任务, 设计了对话生成的预训练模型 PLATO, 来支持多种对话类型的生成任务. 同时, 微软提出了基于 GPT-2 的 DialoGPT[25], 使用 48 层 Transformer 来捕捉文本数据中的长程依赖性.

层次性的序列到序列模型能够在一定程度上有效地刻画对话的层次结构, 以 Transformer 为结构的模型更好地将对话结构进行了层次化的建模, 但仅依靠序列到序列模型和 Transformer 的相对位置编码, 仍然较难刻画本文所提出的 3 种关联结构. 为了刻画多个话轮之间的复杂关联关系, 本文按关联关系将对话建模成不同的图结构, 在序列到序列模型的基础上加入了图神经网络层, 从而提出了基于对话结构的多轮对话生成模型.

2 基于对话结构与图神经网络的多轮对话生成模型

对于对话 $c=\{u_1,u_2,...,u_{n-1},u_n\}$, 其中, u_i 代表第 i 个话轮, 本模型的任务是: 根据已有的前 $n-1$ 轮对话历史信息, 其中包括文本与已标注的其他辅助信息等, 来合理地预测出第 n 个话轮, 并将其作为对话回复.

如图 2 所示, 本模型以多输入的序列到序列模型为基础框架, 模型构成可以大致分为 3 个部分: 双向多输入 LSTM 编码器、图神经网络层、单向 LSTM 解码器. 第 1 步, 将每个话轮都作为一个独立输入, 经由 LSTM 编码器生成 $n-1$ 个中间语义向量 $X_i(i=1,2,\ldots,n-1)$, 便可得到整个对话历史的中间语义向量 $C=Average(X_1,X_2,\ldots,X_{n-2},X_{n-1})$; 第 2 步, 将 n 个中间语义向量 $n-1$ 作为 n 个节点的特征, 按照本文设计的 3 种关联结构组建节点之间的边特征, 这些节点特征和边特征共同构成了图, 之后进行图卷积和图池化计算, 即可从图中得到 n 个节点的隐状态 $H_i(i=1,2,\ldots,n)$,整个对话历史的隐状态可表示为 $C'=Average(H_1,H_2,\ldots,H_{n-2},H_{n-1})$; 第 3 步, 解码, 使用一个单向 LSTM 作为解码器, 结合整个对话历史的中间语义向量 C 与隐状态 C'共同解码. 测试时, 通过集束搜索算法, 最终得到模型生成的对话回复.

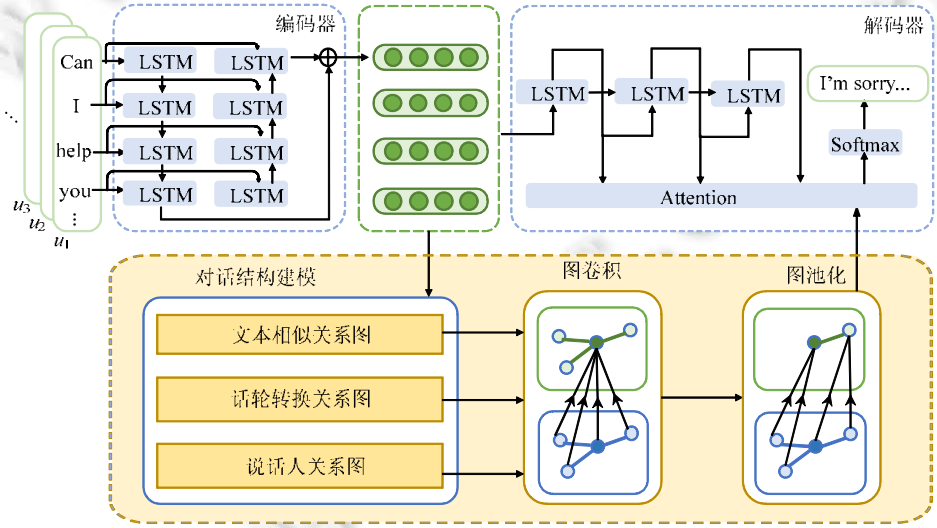


图 2 模型示意图

2.1 双向多输入LSTM编码器

长短期记忆网络(long short term memory networks, LSTM)是在循环神经网络(recurrent neural networks,

RNN)的基础上加以改进的网络模型, 旨在处理传统 RNN 难以解决的长距离依赖的问题. LSTM 的组成部分主要有记忆单元(memory cell)和 3 个门控制单元, 即输入门(input gate)、遗忘门(forget gate)、输出门(output gate). 其中: 记忆单元负责记忆长期历史信息, 遗忘门负责决定记忆单元中历史信息的去留, 输入门负责决定当前时刻的网络输入有多少保存到记忆单元中, 输出门负责决定当前的记忆单元状态有多少进入当前的输出. 记忆单元中存储长期记忆, 当前输出中则可以体现短期记忆.

本文使用了词级别的双向多输入 LSTM 来作为编码器. 不同于单向编码器, 双向编码器可以从两个方向对序列进行刻画, 可以表达出更多的语义信息. Transformer 编码器在数据量较大、单句长度较长的场景下, 能够显著优于双向 LSTM 编码器, 但本实验使用的 DailyDialog 数据集具有数据规模较小、单句长度较短的特点, 因此双向 LSTM 编码器也可以对本实验使用的文本数据进行有效的编码, 并在本实验中取得了更好的效果(见后文表 2).

给定一组话语构成的对话信息 $c_i=\{u_1,u_2,...,u_{n-2},u_{n-1}\}$, 其中每条话语由多个字词单位构成, 即 $u_i=\{x_1,x_2,...,x_{n-2},x_{n-1}\}$. 多输入编码器将每个话语视为一个独立输入, 从而缓解长距离依赖问题. 在具体实现方面, 首先通过词嵌入矩阵将话语中的字词 x_i 转换为实值向量 w_i, 即 $u_i=\{w_1,w_2,...,w_{n-2},w_{n-1}\}$; 然后, LSTM 将每条话语作为独立输入, 分别按照正向、反向的顺序计算出话语的前向隐藏状态 $\{\vec{h}_1,\vec{h}_2,...,\vec{h}_{n-2},\vec{h}_{n-1}\}$ 和反向隐藏状态 $\{\overleftarrow{h}_1,\overleftarrow{h}_2,...,\overleftarrow{h}_{n-2},\overleftarrow{h}_{n-1}\}$; 再将两个末时刻的隐藏状态 $\vec{h}_{n-1}$ 和 $\overleftarrow{h}_{n-1}$ 进行拼接, 从而得到话语的编码向量 X_i, 即 $X_i=[\vec{h}_{n-1};\overleftarrow{h}_{n-1}]$. 整个对话历史的中间语义向量可表示为 $C=Average(X_1,X_2,...,X_{n-2},X_{n-1})$.

2.2 图神经网络层

卷积神经网络(convolution neural networks, CNN)在自然语言处理、图像处理等多个领域展现了优秀的建模效果. 传统的卷积神经网络模型实质是在数据结构平移不变性的前提下进行特征提取, 因此相关应用只能局限在规则的欧氏空间中. 但现实中的许多数据是以不规则的非欧式空间结构呈现的, 比如分子结构数据、社交网络结构以及对话数据结构等.

Bruna 等人[26]首次提出了将卷积神经网络模型泛化至非欧式空间的方法. 之后的相关研究基本沿袭了卷积神经网络的网络局部连接和卷积核参数共享的核心思想, 通过在不规则的图结构上定义卷积算子、池化算子来提取空间特征. 许多文献广泛使用了切比雪夫多项式来拟合卷积核, 在此基础上, Kipf 和 Welling[27]一同提出了图卷积网络模型(graph convolution networks, GCN), 借助图的拉普拉斯矩阵的特征值和特征向量来定义卷积算子的方法.

对于一个多轮对话, 以 $n-1$ 个顶点来表示前 $n-1$ 个话轮, 以话轮的中间语义编码 X_i 作为第 i 个顶点 $vertex_i$ 的特征, 以设定好的规则建立邻接矩阵来表示顶点 i 与顶点 i 之间的边特征信息 $edge(i,j)$, 从而构建了对话的图结构 $G=(vertex,edge)$. 如图 3 所示(其中, A, B 代表不同的说话人. 文本相似度关系图-1 为无权图, 文本相似度关系图-2 的权重为前后话语之间的余弦相似度), 参照之前设计的 3 种对话关联结构, 本文设定了如下 3 种具体的关系图结构.

- 文本相似度关系图(text-based similarity relation graph)

文本相似度关系图负责刻画基于文本相似度的关联结构, 即按照话轮发生的时间顺序, 将话轮顶点依次相连. 以图 1 所示的对话内容为例, 对话历史中的 4 句话被抽象为图中的 4 个顶点, 每一句话的顶点与下一句话的顶点相连. 图中边的权重设计有两种方案: 一种是权重默认设置为 1, 即无权图; 另一种是两条话语之间的余弦相似度. 即给定一组对话 $c=\{u_1,u_2,...,u_{n-1},u_n\}$, 可以将边设计为

$$edge(u_i,u_j)=\begin{cases}1, & \text{if } i,j\in[1,n-1] \text{ and } j-1=1\\ 0, & \text{else}\end{cases} \tag{1}$$

或

$$edge(u_i,u_j)=\begin{cases}\cos(u_i,u_j), & \text{if } i,j\in[1,n-1] \text{ and } j-1=1\\ 0, & \text{else}\end{cases} \tag{2}$$

- 话轮转换关系图(turn switching based relation graph)

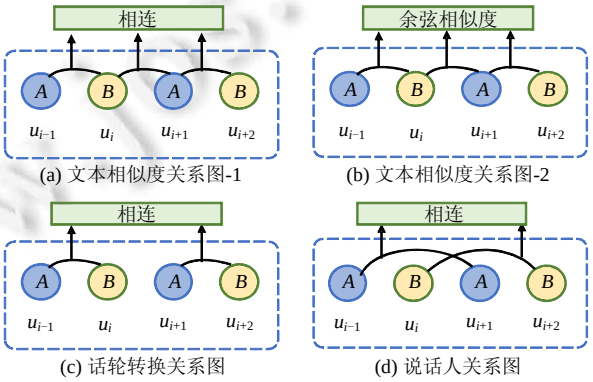
话轮转换关系图负责刻画基于话轮转换的关联结构，即将话轮转换前和转换后的两个话语顶点相连. 该图考虑了潜在于话轮转换过程中的一些交互逻辑，如问答、请求与回复等，从而帮助生成内容能够完成与前文的交互. 以图 1 所示的对话内容为例，对话历史中的 4 句话被抽象为图中的 4 个顶点，前两句话构成了一个话轮转换的过程，于是，第 1 句话的顶点与第 2 句话的顶点之间可建立边关系；后两句话构成了第 2 个话轮转换过程，于是，后两句话的顶点之间可建立边关系. 即给定一组对话 $c=\{u_1,u_2,...,u_{n-1},u_n\}$，可以将边设计为

$$edge(u_i,u_j)=\begin{cases}1, & \text{if } i,j\in[1,n-1] \text{ and } j-1=1 \text{ and } i \text{ is } odd\\ 0, & \text{else}\end{cases} \tag{3}$$

- 说话人关系图(speaker-based relation graph)

说话人关系图对应基于说话人的关联结构，即在图中连接来自同一说话人的话语顶点. 这种图考虑了说话人变量在对话中的影响，如说话人在对话中承担的角色、说话人的语言倾向等，从而使得说话人因素能够辅助对话回复生成. 以图 1 所示的对话内容为例，对话历史中的 4 句话被抽象为图中的 4 个顶点，其中，第 1 句话和第 3 句话来自说话人 *A*，第 2 句话和第 4 句话来自说话人 *B*. 于是，可将第 1 句话和第 3 句话的顶点相连、第 2 句话与第 4 句话的顶点相连. 即给定一组对话 $c=\{u_1,u_2,...,u_{n-1},u_n\}$，边结构应被设计为

$$edge(u_i,u_j)=\begin{cases}1, & \text{if } i,j\in[1,n-1] \text{ and } j-1=2\\ 0, & \text{else}\end{cases} \tag{4}$$



(a) 文本相似度关系图-1　(b) 文本相似度关系图-2

(c) 话轮转换关系图　(d) 说话人关系图

图 3　4 种具体的关系图结构

本模型使用了 Kipf 和 Welling 提出的图卷积算子[27]，对以上不同的图结构进行单层图卷积操作. 卷积后，图中节点的特征值可作为节点经过图卷积后的中间向量. 之后，通过图全局平均池化操作提取对话图中的隐藏特征，具体公式如下:

$$V=GCN(vertex,edge) \tag{5}$$

$$H_i=\frac{E_i^T v_i}{\sum E_i} \tag{6}$$

其中，$\sum E_i$ 代表经过图卷积后的节点集合，v_i 代表经过图卷积后的第 i 个节点的特征值. H_i 为池化后的节点隐状态表达向量，E_i^T 为第 i 条话语对应的边特征信息的转置矩阵，X_i 为第 i 条话语经过图卷积操作后对应的中间向量. 于是，通过求取 n 条话语对应表达向量的平均值，可以得到关于每个对话的隐状态语义表示:

$$C'=Average(H_1,H_2,...,H_{n-2},H_{n-1}).$$

2.3 结合注意力机制的单向LSTM解码器

在序列到序列模型中，解码器主要负责将中间语义向量转换为目标语言，即自然语言. 而对话回复的生成必然伴随着对上下文中的部分关键信息的聚焦过程，因此有必要在解码器中加入注意力机制. 本模型选用

了结合注意力机制的单向 LSTM 解码器来进行解码. 在 t 时间步时, 首先计算对话隐状态语义表示 C'对上一步的隐藏状态 s_{t-1}的注意力权重向量 a_{t-1}, 然后拼接该权重向量 a_{t-1}与上一步隐藏状态 s_{t-1}, 从而得到这一时间步的隐藏状态 s_t. 最终可以得到在 $D=\{u_1,u_2,...,u_{n-1}\}$的对话历史下, 该时间步的预测概率值. 具体公式如下:

$$a_{t-1}=Attention(s_{t-1},C',C') \tag{7}$$

$$s_t=Concatenate(s_{t-1};a_{t-1}) \tag{8}$$

$$p(y_t|y_1,y_2,...,y_{t-1},D)=softmax(y_{t-1},s_t,C') \tag{9}$$

2.4 模型训练与测试

给定对话历史 $c_i=\{u_1,u_2,...,u_{n-1},u_n\}$, 本文所要处理的任务是, 在已有对话历史 $D=\{u_1,u_2,...,u_{n-1}\}$的基础上预测 u_n. 其中, 在第 t 步时, 模型根据对话历史 D 和前 t–1 步预测出的$\{y_1,y_2,...,y_{t-1}\}$来预测 u_n 的第 t 个单词 y_t. 本文使用了 Adam 优化器来训练模型, 训练目标是最小化交叉熵损失函数来学习模型参数θ, 损失函数定义如下:

$$L(\theta)=-\sum_{t=1}^{T_Y}\log P(y_t \mid y_1,y_2,...,y_{t-1},D;\theta) \tag{10}$$

测试过程中, 本模型使用了集束搜索进行预测, 即: 同时考虑多条生成序列, 每步解码选取概率最大的 n 个词进行考虑, 从而进一步保证对话回复生成的多样性.

3 实验设置与结果分析

3.1 数据集

DailyDialog[28]是抓取自英语学习网站的闲聊式多轮对话数据集, 该数据集中的对话是有英语学习者围绕生活中的常见主题来书写的双人对话. 相较社交网络的对话数据集, DailyDialog 数据集具有语法规范性较好、主题明确的特点; 电影台词的对话数据集的单条台词往往过短, 而 DailyDialog 更加符合日常生活中的闲聊式对话的特征[28]. 因此, 本实验选择了 DailyDialog 数据集进行实验. DailyDialog 数据集共有 13 118 组对话, 每个对话平均 7.9 轮, 每组对话平均含有 114.7 个单词, 每个话语平均含有 14.6 个词. 我们对原始数据集中的语料进行筛选, 选择了对话轮次大于等于 8 的数据, 并且要求训练数据第 8 轮话语中的单词数超过 4. 最后得到 200 条测试数据和 5 296 条训练数据(见表 1). 下面是对本实验实际使用的部分 DailyDialog 数据集进行的统计分析.

表 1 DailyDialog 的词汇数据统计

	数目	对话轮次	每个对话的平均词数	每个话语的平均词数
训练集	5 296	8	157.3	19.66
测试集	200	8	154.5	19.31

3.2 评测指标与参数设置

本实验采用了多种评估指标, 其中包括两种自动评测指标(BLEU,ROUGE)及人工评测指标. 以下是对这些评估指标的简要说明.

- BLEU[29]: BLEU 最早被应用于机器翻译领域, 后来也被广泛用于评估生成任务的文本质量. BLEU 评估模型生成的句子与实际句子之间的差异, 取值范围为 0.0–100.0, BLEU 值越高, 说明两个句子相似程度越高.
- ROUGE[30]: ROUGE 是评估自动文摘以及机器翻译的一组指标, 它将生成内容与参考内容进行比较计算, 得出相应的分值, 以衡量“相似度”. 本实验选取 ROUGE-1 的 *F-score* 作为记录的 ROUGE 值. 取值范围为 0.0–100.0, ROUGE 值越高, 说明两个句子相似程度越高.
- Human: 本实验借鉴了 Zhang 等人[31]的人工评分规则. 人工评分过程为 3 个测试人员各自独立进行, 每人对不同模型的生成结果进行评分, 最终记录 3 人评分结果的平均值. 规则设定的分数范围为 0–3: 0 分代表生成内容完全不可读, 1 分代表生成内容不正确或不相关, 2 分代表生成内容部分正确且

相关, 3 分代表生成内容完全正确且相关.

本实验的超参数的设定如下: 嵌入层的输出维度为 128, LSTM 的隐藏层维度为 128, 随机失活比例为 0.2, 优化器为 Adam 优化器[32], 学习率为 0.001, 批次大小为 16, 句子的最大长度为 20, 迭代次数为 25.

3.3 基准模型介绍

为了评估模型效果, 本文选取了 5 种基准模型作为对比, 分别是 Seq2Seq, Seq2Seq_multi, Dir-VHRED[5], ReCoSa[8], HRG[33]. 以下是对这些基准模型的简要介绍.

- Seq2Seq: 使用单层 LSTM 的一对一序列到序列模型, 将对话历史中的每一条话语拼接组合成单独一个输入, 解码预测输出时使用集束搜索算法, 以解码后输出内容作为对话回复.
- Seq2Seq_multi: 使用单层 LSTM 的多对一序列到序列模型, 将对话历史中的每一条话语都作为单独的输入, 解码预测时同样使用集束搜索算法, 以解码后的输出内容作为对话回复.
- Dir-VHRED(2019): Dir-VHRED 是基于变分自编码器和层次序列到序列结构的一种模型, 与传统方法不同的是, 它采用了狄利克雷分布作为潜变量的先验分布. 该模型的具体实现为: 第 1 步, 使用多层次编码, 先根据文本进行话语级别的编码, 再根据话语编码进行上下文编码; 第 2 步, 从上下文编码中随机采样得到一些狄利克雷随机变量, 随机变量再与上下文编码拼接形成最终编码向量; 第 3 步, 进行解码, 生成对话回复.
- ReCoSa(2019): ReCoSa 是基于 Transformer 结构的模型. 在具体流程中, 对话历史中的每一句话语都分别经过一个由多头注意力和前馈网络构成的编码器模型, 再将得到的多个表示进行注意力计算得到对话历史表达, 之后对话历史表达进入解码器模型进行解码, 最终得到生成的对话回复.
- HRG(2019): HRG 是基于条件变分自编码器的层次生成模型, 它使用表达重建模型来捕捉表达与辅助信息“意图”之间的层次关系, 又使用了表达注意力模型来有效地结合表达与内容, 从而更好捕捉语义信息.

3.4 实验结果与分析

(1) 与基准系统比较

表 2 列出了本模型与前述基准模型的自动评测指标结果.

表 2 DailyDialog 数据集上的自动评测指标结果

模型	BLEU	ROUGE
Seq2Seq	13.52	12.74
Seq2Seq_multi	16.71	16.61
Dir-VHRED	10.71	11.23
ReCoSa	17.29	18.37
HRG	18.30	18.09
Ours (Transformer Encoder)	16.58	16.73
Ours	**20.01**	**20.06**

表 3 列出了本模型与前述基准模型的人工评测指标结果, 该表格记录了各模型的得分比例, *Score* 为分数与比例的加权之和.

表 3 DailyDialog 数据集上的人工评测指标结果

模型	得分				
	0 (%)	1 (%)	2 (%)	3 (%)	Score
Seq2Seq	8	63	20	9	1.30
Seq2Seq_multi	5	62	23	10	1.38
Dir-VHRED	11	73	13	3	1.08
ReCoSa	2	57	26	15	1.54
HRG	2	64	24	10	1.42
Ours	**1**	**46**	**35**	**18**	**1.70**

可以看出, 在使用 BLEU, ROUGE 和人工指标的测评中, 本模型在 DailyDialog 数据集上取得了优于其他

所有模型的效果, 这说明本文提出的基于多种对话结构的对话生成模型是有效的. 在基准模型中, 多输入的 Seq2Seq_multi 要优于单输入的 Seq2Seq, 这是因为基于 LSTM 结构的 Seq2Seq 存在梯度消失的问题. 单输入情况下, 位置较前的信息的影响力有限, 模型仍然容易聚焦在当前的信息, 而忘记远端的内容, 难以表达长距离依赖关系. 而将输入拆解为多输入的方法, 显著降低了整体输入的序列长度, 能够有效地缓解长距离依赖的问题. 因此, 本模型在具体实现中也使用了多输入的 LSTM 编码器. Dir-VHRED 以 VHRED 模型为基础, 但效果在此处不甚理想. ReCoSa 有效地利用了 Transformer 模型, 效果要显著优于 Dir-VHRED 和 Seq2Seq. HRG 在 CVAE 的基础上进行改进, 并结合了文本以外的辅助信息意图进行建模, 效果与 ReCoSa 基本持平.

从表 2、表 3 的结果中可以看出, 本模型同样超越了 Dir-VHRED, ReCoSa 和 HRG 的表现. 这说明对于对话生成任务, 使用图神经网络结构来刻画对话关联结构的方法是有作用的.

(2) 不同影响因素比较

在实验中, 本文设计了多种图结构设计方法, 并对多种方法及其组合进行了效果评测. 实验结果可见表 4, 其中, basic 代表不具任何边结构的图, ts1 代表无权的文本相似关系图, ts2 代表有权的文本相似关系图.

表 4 不同图结构对模型效果的影响

模型	BLEU	ROUGE
Seq2Seq_multi	16.71	16.61
+ basic	17.12	17.68
+ text-based similarity-1 (ts1)	18.76	18.53
+ test-based similarity-2 (ts2)	19.28	19.75
+ turn switching (turn)	18.92	17.55
+ speaker-based (speaker)	18.84	18.52
+ ts1+turn	**20.01**	**20.06**
+ ts1+speaker	19.49	18.65
+ ts1+ts2	19.12	18.85
+ turn+speaker	19.56	18.68
+ ts2+turn	18.92	18.59
+ ts2+speaker	19.13	18.65

可以得到如下结论.

- 第一, 对比 Seq2Seq_multi 和 basic 的结果可知, 使用图神经网络处理对话生成任务是有效的.
- 第二, 采用单一图结构的方法之间的效果近似, 其中, 以有权的基于文本相似度的关联结构最好.
- 第三, 将图结构进行叠加后, 实验结果可以取得进一步的提升, 其中, 以无权的文本相似度关系图结合话轮转换关系图的复合图效果最好.

每种叠加图结构较单一图结构都有一定提高, 这说明本文对对话结构的多种建模方法是有意义的, 增加图结构的数目能够在一定程度上更好地刻画对话逻辑. 文本相似度关系图和说话人关系图都为模型提供了一种稳定的长期线索, 而话轮转换关系图提供的是突发的短期线索. 在几种复合图中, 无权的文本相似度关系图结合话轮转换关系图的复合图取得了最好的效果. 这是因为这种组合可以同时捕捉长期和短期线索, 从而兼顾了整体的对话历史背景和局部的短期内容更替. 此外, 无权的基于文本相似度的关联结构比较简单直接, 在复合图的情况下更容易被模型理解.

(3) 实例分析

本文选择基线模型中表现较好的 ReCoSa 作为对照模型, 对图 4 中的两个实例进行分析.

- 在示例 1 中, ReCoSa 未能有效地结合对话历史中的天气信息; 而本模型顺利地把握住了长期线索中的天气信息, 也抓住了短期线索中对天气变幻莫测的无常感.
- 在示例 2 中, ReCoSa 理解了打车回家的信息, 做出了送别的回复, 但未能抓住对话中的“share”所体现的合作意向; 本模型则成功理解了对话历史中的合作意向信息.

这说明本文使用的图结构能够更好地刻画对话内在的复杂逻辑结构, 从而在一定程度上提升对话生成的质量.

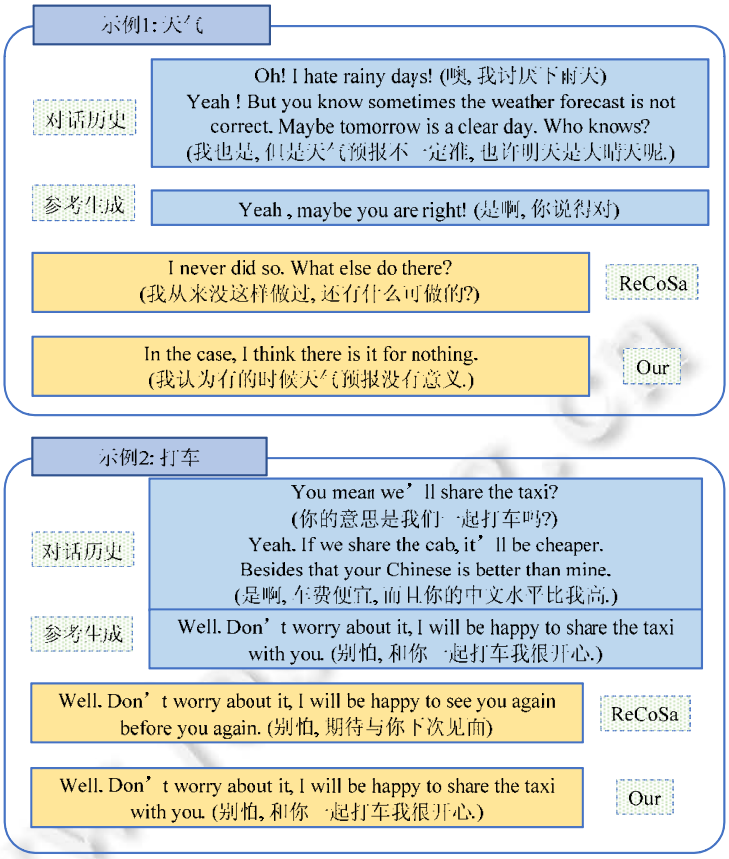


图 4　本模型与基线模型 ReCoSa 的生成内容对比示例图

4　结　语

本文提出了一种基于对话结构的对话生成模型, 该模型以序列到序列结构为框架, 通过图神经网络, 有效刻画了基于文本相似度的关联结构、基于话轮转换的关联结构和基于说话人的关联结构, 增强了模型对对话结构的理解能力, 从而帮助模型生成高质量的对话回复文本. 本文在 DailyDialog 数据集上分别测试了单一图结构和复合图结构的模型效果. 结果证明, 本方法在多个评估指标上超出了其他基线模型. 我们未来的工作方向大致有加入多模态信息、规范文本生成这两种, 对话作为一种现实世界的行为, 不只局限于文本一种媒介形式, 还有听觉信息、视觉信息等, 如明显的肢体动作、突然的打断插话, 因此, 加入多模态信息是有必要的. 另外言语规范、用词文明是维护社会公共秩序的重要组成部分, 机器生成的文本理应避免出现粗俗、不文雅的内容. 针对这一点, 我们可以借助强化学习的帮助来规范化文本生成, 从而防止出现文本生成内容出现恶意倾向.

References:

[1] Vinyals O, Le QV. A neural conversational model. arXiv:1506.05869v3, 2015.

[2] Serban I, Sordoni A, Bengio Y, Courville A, Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016. Article No.9883.

[3] Serban I, Sordoni A, Lowe R, Charlin L, Pineau J, Courville A, Bengio Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2017. Article No.10983.

[4] Chen H, Ren Z, Tang J, Zhao YE, Yin D. Hierarchical variational memory network for dialogue generation. In: Proc. of the World Wide Web Conf. 2018. 1653−1662.

[5] Zeng M, Wang Y, Luo Y. Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. 1267−1272.

[6] Gu X, Cho K, Ha JW, Kim S. Dialogwae: Multimodal response generation with conditional Wasserstein auto-encoder. In: Proc. of the 7th Int'l Conf. of Learning Represtations. 2019.

[7] Xing C, Wu Y, Wu W, Huang Y, Zhou M. Hierarchical recurrent attention network for response generation. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. Article No.11965.

[8] Zhang H, Lan Y, Pang L, Guo J, Cheng X. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 3721−3730.

[9] Wang XP. The power relationship between *Q* & *A* and women in TV interview. Journal of Dezhou University, 2010, 26(3): 44−47 (in Chinese with English abstract).

[10] Chen H, Liu X, Yin D, Tang J. A survey on dialogue systems: recent advances and new frontiers. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25−35. https://doi.org/10.1145/3166054.3166058

[11] Zhou X, Li L, Dong D, Liu Y, Chen Y, Zhao WX, Yu D, Wu H. Multi-turn response selection for chatbots with deep attention matching network. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol.1. 2018. 1118−1127.

[12] Zhang Z, Li J, Zhu P, Zhao H, Liu G. Modeling multi-turn conversation with deep utterance aggregation. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. 2018. 3740−3752.

[13] Yang L, Hu J, Qiu M, Qu C, Gao J, Croft WB, Liu X, Shen Y, Liu J. A hybrid retrieval-generation neural conversation model. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. 2019. 1341−1350. https://doi.org/10.1145/3357384.3357881

[14] Chen C, Zhu QQ, Yan R, Liu JF. Survey on deep learning based open domain dialogue system. Chinese Journal of Computer, 2019, 42(7): 1439−1466 (in Chinese with English abstract).

[15] Serban I, Klinger T, Tesauro G, Talamadupula K, Zhou B, Bengio Y, Courville A. Multiresolution recurrent neural networks: An application to dialogue response generation. Proc. of the AAAI Conf. on Artificial Intelligence, 2017, 31(1): 3288−3294.

[16] Bowman SR, Vilnis L, Vinyals O, Dai AM, Bengio S. Generating sentences from a continuous space. In: Proc. of the 20th SIGNLL Conf. on Computational Natural Language Learning. 2016. 10−21.

[17] Wang MY, Yu DY, Yan R, Hu WP, Zhao DY. Chinese multi-turn dialogue tasks based on HRED model. Journal of Chinese Information Processing, 2020, 34(8): 78−85 (in Chinese with English abstract).

[18] Zhao T, Ran Z, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol.1. 2017. 654−664.

[19] Shen L, Feng Y, Zhan H. Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 5497−5502.

[20] Hu D. An introductory survey on attention mechanisms in NLP problems. In: Proc. of the SAI Intelligent Systems Conf. 2019. 432−448.

[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. 5998−6008.

[22] Wu SS, Lin ZD. Research on dialogue generation mechanism of chat robot based on Seq2 seq and attention model. Automation and Instrumentation, 2020(7): 186−189 (in Chinese with English abstract).

[23] Bao S, He H, Wang F, Wu H, Wang H. Plato: Pre-trained dialogue generation model with discrete latent variable. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2019. 85−96.

[24] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol.1: Long and Short Papers). 2019. 4171−4186.

[25] Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, Gao J, Liu J, Dolan B. DIALOGPT: Large-scale generative pre-training for conversational response generation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. 270–278.

[26] Bruna J, Zaremba W, Szlam A, Lecun Y. Spectral networks and locally connected networks on graphs. In: Proc. of the 2nd Int'l Conf. of Learning Represtations. 2014.

[27] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[28] Li Y, Su H, Shen X, Li W, Cao Z, Niu S. Dailydialog: A manually labelled multi-turn dialogue dataset. In: Proc. of the 8th Int'l Joint Conf. on Natural Language Processing, Vol.1. 2017. 986–995.

[29] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 311–318.

[30] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Proc. of the Text Summarization Branches Out. 2004. 74–81.

[31] Zhang W, Song K, Kang Y, Wang Z, Sun C, Liu X, Li S, Zhang M, Si L. Multi-turn dialogue generation in e-commerce platform with the context of historical dialogue. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing: Findings. 2020. 1981–1990.

[32] Kingma D, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. of Learning Represtations. 2015.

[33] Zhang B, Zhang X. Hierarchy response learning for neural conversation generation. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. 1772–1781.

附中文参考文献:

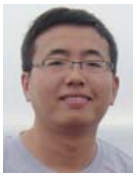
[9] 王晓培. 电视访谈中问答毗邻对与女性的权势关系. 德州学院学报, 2010, 26(3): 44–47.

[14] 陈晨, 朱晴晴, 严睿, 柳军飞. 基于深度学习的开放领域对话系统研究综述. 计算机学报, 2019, 42(7): 1439–1466.

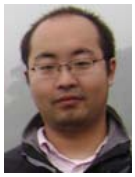
[17] 王孟宇, 俞鼎耀, 严睿, 胡文鹏, 赵东岩. 基于 HRED 模型的中文多轮对话任务方法研究. 中文信息学报, 2020, 34(8): 78–85.

[22] 吴石松, 林志达. 基于 seq2 seq 和 Attention 模型的聊天机器人对话生成机制研究.自动化与仪器仪表, 2020(7): 186–189.

姜晓彤(1997－), 女, 博士生, CCF 学生会员, 主要研究领域为自然语言处理.



李寿山(1980－), 男, 博士, 教授, CCF 专业会员, 主要研究领域为自然语言处理.



王中卿(1987－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.

周国栋(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.