基于图表示学习的消息回复关系判断方法

梁永明[1,2], 田 恬[1,2], 杨小雨[1,2], 张 熙[1,2], 邱莉榕[2,3]

[1]北京邮电大学网络空间安全学院 北京 中国 100876
[2]北京邮电大学可信分布式计算与服务教育部重点实验室 北京 中国 100876
[3]北京邮电大学计算机学院(国家示范性软件学院) 北京 中国 100876

摘要 微信、QQ 和钉钉等社交媒体都提供多对多聊天群组功能, 这些聊天群组包含海量信息, 对群组聊天内容进行有效分析, 获取有价值的关联信息, 是当前领域的研究热点。群组中用户间交互是群组实现的主要功能, 用户间消息回复是用户间交互实现的方式, 消息间的回复行为下隐藏着消息间和用户间的关系。群组消息间回复通常是隐式和非连续的, 大部分群组消息间没有指定明确的回复关系, 当前消息也不一定是上一条临近消息的回复, 回复关系要根据具体的聊天场景确定。当消息间没有显示指定回复关系时, 回复不易于分析和理解群组聊天内容, 阻碍了对群组聊天内容的整体性分析。本论文针对群组消息间的回复关系, 提出了基于图表示学习的消息回复关系判断方法, 该方法不同于以往方法仅使用部分群组要素, 是在综合学习消息的文本信息、发送消息的用户信息和上下文信息的基础上, 根据群组内容构建群组图和生成自适应消息图, 得到了多种群组要素信息和要素间关系组成的图结构, 利用图模型在图结构上进行群组消息的表示学习, 图模型输出群组消息的表示向量, 拼接消息对的表示向量并进一步预测群组消息间的回复关系。在消息间回复关系的学习过程中, 图模型通过任务学习更新图中消息节点, 同时更新图中用户节点向量表示, 经过用户向量分析实验验证了该模型输出的用户向量的有效性和合理性。在公开数据集和标注数据集上进行了对比实验和显著性检验分析, 结果显示模型在多个评估指标上大幅优于对比模型, 如在 F1 指标上, 比单纯依赖 BERT 的句子对分类模型提高了接近 20%。

关键词 图模型; 对话系统; 消息回复; 自然语言推理; 会话分析; 自适应构图; 群组分析
中图法分类号 TP391.4 DOI 号 10.19363/J.cnki.cn10-1380/tn.2021.09.15

The Method for Identifying the Conversation Responding Relationships using Graph Representation Learning

LIANG Yongming[1,2], TIAN Tian[1,2], YANG Xiaoyu[1,2], ZHANG Xi[1,2], QIU Lirong[2,3]

[1] School of Cyberspace Secunty, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Key Laboratory of Trustworthy Distributed Computing and Service(BUPT), Ministry of Education , Beijing 100876, China
[3] School of Computer(National demonstrative school of software), Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract Social media, such as WeChat, QQ and Ding Talk, all provide many-to-many chat groups. These chat groups contain a large amount of information. It is a research hotspot in the current field to effectively analyze the group chat content and obtain valuable related information. Interaction between users in a group is the main function of group implementation, and message reply between users is the way to realize interaction between users. The relationship between messages and users is hidden under the reply behavior between messages. The reply between group messages is usually implicit and discontinuous. Most group messages do not specify a clear reply relationship, and the current message is not necessarily the reply of the previous adjacent message. The reply relationship should be determined according to the specific chat scene. When there is no designated reply relationship between messages, the reply is not easy to analyze and understand the group chat content, which hinders the overall analysis of the group chat content. In this paper, aiming at the reply relationship between group messages, a method of judging message reply relationship based on graph representation learning is proposed. This method is different from the previous method, which only uses part of group elements. Based on the comprehensive study of text information, user information and context information of messages, it constructs a group graph and generates an adaptive message graph according to group content, and obtains a graph structure composed of various group element information and relationships among elements. The graph model is used to learn the representation of group messages on graph structure. The graph model outputs the representation vectors of group messages, splices the representation vectors of message pairs and further predicts the reply relationship between group messages. In the learning process of reply relationship between messages, the graph model updates the message nodes in the graph through task learning, and updates the vector representation of user nodes in the graph at the same time. The validity and rationality of

通讯作者: 张熙, 博士, 副教授, zhangx@bupt.edu.cn
本课题得到国家自然科学基金项目资助（61976026）资助。
收稿日期: 2021-04-30; 修改日期: 2021-08-09; 定稿日期: 2021-08-10

the user vectors output by the model are verified by user vector analysis experiments. A comparative experiment and significance test analysis are carried out on the public data set and the labeled data set. The results show that the model is significantly superior to the comparative model in many evaluation indexes, such as the F1 index, which is nearly 20% higher than the sentence classification model which only depends on BERT.

Key words graph model; dialogue system; conversation responding relationships; natural language inference; conversation analysis; adaptive graph construction; group analysis.

1 引言

微信、QQ、钉钉等社交媒体已经成为人们交流的主要渠道, 在这些媒体的聊天群组中, 群组用户频繁地发送消息互相回复进行交流。群组聊天是一个多方多轮的对话场景, 群组用户作为对话的参与者, 群组消息作为对话元素, 一段时间内发送的群组消息构成一个完整会话。不同于微博、Twitter、Facebook 等平台明确限定评论间的回复关系, 群组成员在通过发送群组消息进行交流时, 消息间回复关系通常是隐式和非连续的, 即没有指定的回复关系, 而且当前消息不一定是上一条临近消息的回复, 回复关系要根据具体的聊天场景确定。由于单条消息很难表达会话完整语义, 阻碍对会话内容的整体理解, 识别消息间的回复关系成为深入理解群组内容的关键。

针对群组消息间的回复关系识别任务具有很大的挑战。与之相关的是文本匹配任务, 比如自然语言推理(Natural Language Inference, NLI)[1]、问答任务(Question Answering, QA)[2]。在以往的文本匹配任务中, 自然语言推理的前提和问答任务的问题通常作为句子 A, 自然语言推理的假设和问答任务的答案作为句子 B, 将句子 A 和 B 合并组成长序列, 输入到模型中进行分类, NLI、QA 只考虑句子 A 和句子 B 中的信息来判断句子对的句子间的关系。而群组消息间关系判断更加困难, 不仅需要考虑当前消息的文本信息, 并且需要考虑发送消息的用户信息和当前消息的上下文信息, 如何有效融合上述信息, 更加精准的判断群组消息间的回复关系, 成为本论文要解决的重要问题。

群组消息间回复关系判断任务已经有了一些初步的研究工作, Li Y[3]提出增强的长短期记忆网络(Long Short-Term Memory, LSTM)[4]的自然语言推理算法(Enhanced LSTM for Natural Language Inference, ESIM)[5]通过消息匹配将对话转换为自然语言推理任务进行消息预测。Gu J C[6]提出应用 BERT[7]捕获消息间的交叉语义, 进一步学习消息间的关系。Henghui Zhu H[8]提出掩码 Transformer[9]结构学习消息间的回复关系。这些方法只使用消息的文本和上下文信息或其中部分信息, 未能充分建模消息回复场景下蕴含的丰富的图结构信息。

群组内容主要由用户、消息和消息中包含的单词三种实体组成, 如何利用群组中实体以及实体间的关系是本文要解决的主要问题。图结构可以保存节点和节点间的关系, 可以基于群组内容处理获得群组实体和实体间的关系, 转换为图结构节点和边保存信息, 群组用户、消息和单词作为图结构中的节点, 它们之间的关系作为图结构中的边。图模型可以很好地处理图结构中实体间的空间依赖关系, 如图卷积网络[10](Graph Convolutional Networks, GCN)和异质图注意力网络(Heterogeneous Graph Attention Networks, HGAT)[11]等, 在很多领域都得到了很好的应用效果。GCN 在同质图结构上进行图卷积操作同时编码局部图的结构特征和节点特征。HGAT 通过采用节点层和类型层组成的双层注意力机制, 学习异质图结构中不同相邻节点的重要性以及不同类型节点对当前节点的重要性进行节点表示学习。图模型在任务学习过程中融合节点特征和空间依赖关系, 优化并更新图结构中的任务节点与任务节点相关的其他节点的向量表示进行表示学习。针对深度挖掘群组内容判断消息回复关系的难题, 可以将群组内容转换为由群组实体和实体间关系组成的图结构, 输入图模型进行表示学习, 进一步预测消息间回复关系。

针对群组消息回复关系判断任务, 本论文提出了基于图表示学习的消息回复关系判断方法。首先, 基于群组内容构建全局群组图和生成自适应消息图, 得到群组用户、消息和单词和它们间的关系并保存到图结构。全局群组图包括群组用户、消息和单词三种实体以及它们之间的四种关系信息, 基于群组内容的全局信息统计并计算得到实体间关系, 实体间关系包括: 用户-用户、用户-消息、消息-单词、单词-单词。自适应消息图由消息和消息间关系组成。群组消息间具有一定的局部相关性, 消息序列中消息通常只与当前序列中的消息相关。不同消息序列中的消息间存在不同关系, 消息间关系难以直接计算得到。采用基于目标任务自适应消息图生成方法, 通过处理消息序列自适应地生成消息间关系组成的

消息图。然后, 从全局群组图中抽取与当前消息序列相关的节点和节点间关系得到局部图, 将局部图和自适应消息图合并得到合并子图, 作为图模型的输入。图模型在输入的合并子图上进行表示学习, 输出优化后的融合图结构信息的节点向量表示。从输出的节点向量表示中抽取消息对的向量表示, 输入到孪生网络(Siamese network)[12]预测消息对中消息间的回复关系, 完成群组消息间回复关系判断任务。

在公开数据集和标注数据集上实验验证了, 本论文提出的模型在多个评估指标上大幅优于对比模型, 包括基于 BERT 的句子对分类模型和短文本匹配通用模型 ESIM。经过用户向量分析实验, 表明该图模型输出的用户向量表示能够有效融合文本信息和回复结构信息。

本论文的主要贡献包括以下三点:

(1) 提出了一个基于图模型表示学习的群组消息回复关系判断方法, 是首个尝试使用图模型解决群组消息回复关系判断任务的方法。

(2) 提出了自适应生成消息图的优化, 处理不同的输入消息序列生成任务相关的局部消息图, 用于捕捉消息之间的隐式关联, 弥补人工构图的不足。

(3) 实验验证了所提方法的有效性, 而且验证了所获得的用户向量有效融合了文本信息和回复结构信息。

2 相关工作

2.1 会话消息分析

会话消息分析一直是会话研究的热点, 会话消息研究从一开始的短文本对话(Short text conversation, STC)研究, 经过了多轮两方对话研究, 到现在的多轮多方对话研究, 经过长期的发展。

STC 是给定一条输入消息, 预测合适的消息作为回复消息[13-14], 是单轮两方对话的形式, 预测的过程中只单纯考虑文本语义, 没有给定和考虑对话两方用户信息和历史会话信息。Li Y[3]提出的 Res-ESIM 通过文本匹配方法将对话转换为自然语言推理问题进行消息预测, 解决短文本消息对话关系预测问题。相比对完整对话进行建模, 对短文本对话建模更加简单直接, 应用场景也更广泛, 训练得到的短文本对话预测模型可以直接应用在聊天机器人[15]和自动回复场景[16]上。

两方多轮对话任务是给定两个用户和他们的历史会话消息, 生成或选择合适的消息作为当前消息的回复消息。回复消息生成任务大多数使用统计机器翻译方法[17]或序列到序列(sequence-to -sequence, seq2seq)[18-19]模型生成合适消息。回复消息选择任务通常是基于检索的方法使用排序模型[20-21]从候选消息中选择得分最高的响应消息作为回复消息。两方多轮会话任务场景只限于两个用户之间比较短的对话序列, 没有考虑对话双方的用户信息, 参考的上下文对话消息序列比较短, 消息生成或选择模型适用于机器回复的实际场景。

多方多轮对话是多个用户进行多次消息交互的会话, 是多个线程的会话相互交叉和纠缠的形式, 贴合实际聊天群组场景。多方多轮对话的对话序列中隐藏着会话结构, 群组消息之间的回复关系属于其中一部分结构。多方多轮对话任务通过建模会话结构, 将多个不相关的会话区分开是一个解纠缠[22-23]的过程。Kummerfeld J K[24]提出通过实现一个前馈神经网络构建一个由大量消息回复标记组成的会话解纠缠数据集, 促进聊天群组消息回复研究方法的发展, 本方法也将使用会话解纠缠数据集进行群组消息回复分析。最近, Gu J C[6]提出将 BERT 应用到会话解纠缠任务, 捕获消息间的交叉的语义, 挖掘深度语义判断消息间的关系。Zhu H[8]提出掩码 Transformer 结构学习消息间的回复关系。上述会话结构解纠缠的实现方法中通常只考虑局部消息序列关系区分多个会话, 很少考虑用户信息和全局用户关系信息。

不同于上述工作, 本论文采用基于图模型的群组元素及关系的表示学习方法挖掘群组内容信息, 学习群组消息间的回复关系。

2.2 图模型

图模型在处理图结构中实体之间的空间依赖性方面经历了很长时间的发展。图模型假设实体之间的空间依赖性是节点状态依赖于其邻居节点状态, 为了捕捉这种类型的空间依赖性, 通过信息传递、信息传播和图卷积等方面的研究, 研发得到各种类型的图模型。图卷积网络在捕获图结构局部空间依赖性方面取得了巨大的成功, 通过卷积操作将节点邻居的信息传递给当前节点来优化节点的向量表示。图卷积网络适用于处理同质图结构, 前期工作已有实验数据表明[11], 图卷积网络的卷积操作学习异质图结构的节点表示效果不太理想。最近, 处理异质图结构的异质图注意力网络出现, 异质图注意力网络的设计是为了学习文本语料库中主题、文本和实体信息以及它们之间关系信息进行文本分类。HAGT 学习异质图结构中信息, 通过采用节点层和类型层组成的双层注意力机制, 学习异质图结构中不同相邻节点以及不同类型节点对当前节点的重要性优化节点的向量表示。异质图注意力网络在处理异质图

结构方面相比于图卷积网络有了很大的提升, 但是异质图注意力网络处理的图结构是预定义的、静态的和全局的, 在一些节点关系未知和节点局部相关的图结构中, 异质图注意力网络缺少对应的处理方法和通用框架。

3 消息回复关系判断任务定义

聊天群组中用户间交互是群组实现的主要功能, 用户间消息回复是用户间交互实现的方式, 消息间的回复行为下隐藏着消息间和用户间的关系。本方法可以通过学习消息间的回复行为, 挖掘回复行为下消息间和用户间的关系, 预测消息间的回复关系, 同时优化图中节点的向量表示, 得到副产品优化后的消息和用户向量表示。下面对消息间回复关系判断任务进行详细介绍。

消息间回复关系判断任务是判断群组中的消息间是否存在回复关系, 如果存在回复关系则为正例, 不存在则为负例。给定群组中的一个对话序列, 序列为 $S_1, S_2, \cdots, S_N, S_{N+1}$, 消息 S_{N+1} 为当前消息, 本方法解决从序列中找出当前消息 S_{N+1} 的父消息 S_i 的问题。从序列中抽取当前消息回复的消息和当前消息未回复的消息与当前消息组成消息对, 作为正负样本, 例如 (S_i, S_{N+1}) 和 (S_j, S_{N+1}), 其中 S_i 是 S_{N+1} 回复的消息, S_j 是 S_{N+1} 未回复的消息中随机抽取的一条。

图 1 为构建正负样本的样例, 图中右侧消息数据是从 IRC ubuntu log[25]中随机截取的连续 N+1 条消息, $1, 2, 3, \cdots, N, N+1$ 是消息的序号, 消息序列中只展示了部分消息, 中间部分消息省略没有展示, 图中每条消息是"<用户名>:消息内容"的格式。N+1 条消息中的第 N+1 条为当前消息, 第 1 条消息到第 N 条消息为当前消息的上下文消息。图中线段形状的箭头曲线表示第 N+1 条消息是第 N–1 条消息的回复, 上下文消息中的其他消息是当前消息的未回复消息。当前消息和第 N–1 条消息组成正样本 $(N-1, N+1)$, 如图 1 中点状箭头所示。从未回复消息中随机抽取一条消息与当前消息组成负样本 $(N-3, N+1)$, 图中随机抽取的是第 N–3 条消息, 如图 1 中实线箭头所示。

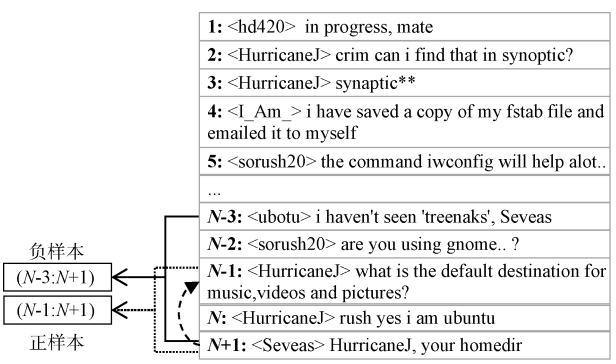

图 1 正负样本构建样例图

Figure 1 Positive and negative samples build sample graph

将构建得到的正负样本消息对输入到本模型中, 判断消息对是否存在回复关系并输出预测结果, 预测结果为 1, 证明组成消息对的两条消息间存在回复关系, 否则组成消息对的两条消息间不存在回复关系。

4 消息回复关系判断方法

如何学习群组消息的交互行为是本文提出方法的核心内容, 基于群组内容特性, 本算法提出以下方法优化群组消息交互行为学习方法。1)针对群组内容特性将群组内容转换为静态和动态两种类型图结构, 实现了聊天群组图结构的针对性构建方法。2)根据群组消息的局部性, 采用局部图抽取和自适应消息图生成方法构建, 得到局部会话的图结构。3)提出了图模型框架, 挖掘局部会话的异质图结构, 学习消息间的回复关系。接下来对本方法以及方法中实现以上优化方法的模块进行详细介绍。

图结构首先通过处理群组内容构建全局群组图和生成自适应消息图, 从全局群组图中抽取与需要判断回复关系的消息对相关的节点和节点间关系得到局部图, 将局部图和自适应消息图合并作为图模型的输入, 图模型在输入的合并子图上进行表示学习并输出节点的向量表示, 并从得到的节点向量表示中抽取消息对的向量表示, 输入到孪生网络模块预测消息对间的回复关系。本方法包括图的构建和生成、局部图获取和合并、异质图注意力网络(HGAT)、孪生网络(Siamese network)四个模块。

(1)图的构建和生成模块包括两个子模块: 全局群组图构建子模块、自适应消息图生成子模块。一方面, 全局群组图构建子模块基于群组内容的全局信息统计处理得到全局群组图。另一方面, 群组消息的消息间具有一定的局部相关性, 消息序列中消息通常只与当前序列中的消息相关, 具有一定的局部性, 不同消息序列中的消息间关系难以直接得到。自适应消息图生成子模块采用任务相关自适应的图学习方法, 基于目标任务通过处理输入的消息序列自

适应地生成消息间关系组成的消息图。当输入群组消息时, 自适应消息图生成子模块只考虑当前消息和当前消息前 N 条连续的历史消息组成 N+1 条消息, 前 N 条连续的历史消息中包含当前消息的回复消息和未回复消息, 将 N+1 条消息输入到自适应消息图生成模块, 模块输出大小为 $(N+1)\bullet(N+1)$ 自适应消息图。(2)局部图获取和合并模块由两个子模块组成, 分别为局部图获取子模块、局部图和自适应消息图合并子模块。消息序列具有一定的局部相关性, 当前消息通常只与当前消息序列中的消息、发送消息的用户和消息中的单词 3 种实体相关, 只考虑与当前消息相关的信息能大大减少噪声信息, 提高当前消息相关的任务效果。当群组消息输入时, 全局群组图中只有部分消息、用户和单词与当前消息相关, 当前消息和当前消息前 N 条连续的历史消息组成 N+1 条消息是当前消息的相关消息, 发送相关消息的用户组成的用户集是相关用户, 相关消息中出现的单词组成单词集是相关单词。从全局图中抽取相关消息、相关用户和相关单词组成的节点和节点间的边构成局部图, 局部图的节点集由相关消息、相关用户和相关单词组成, 局部图的边集由节点集中节点间存在的边组成。局部图获取子模块从全局群组图抽取与当前消息相关的消息、用户和单词组成当前消息的局部图。局部图和自适应消息图分别包含不同的节点关系信息, 局部图和自适应消息图合并子模块合并局部图和自适应消息图得到最终的合并子图, 作为异质图注意力网络模块的输入。(3)异质图注意力网络模块在输入的合并子图上进行表示学习, 利用节点层和类型层组成的双层注意力机制学习合并子图中的信息并输出更新后的节点向量表示。(4)孪生网络模块采用网络参数共享权值方式, 从前 N 条历史消息中抽取当前消息的回复消息或未回复消息, 与当前消息组成消息对作为输入, 消息对的向量表示从异质图注意力网络模块输出的节点向量表示中抽取得到。(5)孪生网络模块将两个不同的输入映射到相同的向量空间, 预测消息对的回复关系。最终将孪生网络的输出合并输入到全连接层中, 经过 soft max 处理, 得到消息对的预测结果。

图 2 为基于图表示学习的消息回复关系判断方法框架图, 图中简明地显示了该方法的整体结构和各模块之间的连接情况。

接下来按顺序对图的构建和生成、局部图获取和合并、异质图注意力网络(HGAT)、孪生网络(Siamese network)4 个模块分别进行详细介绍。

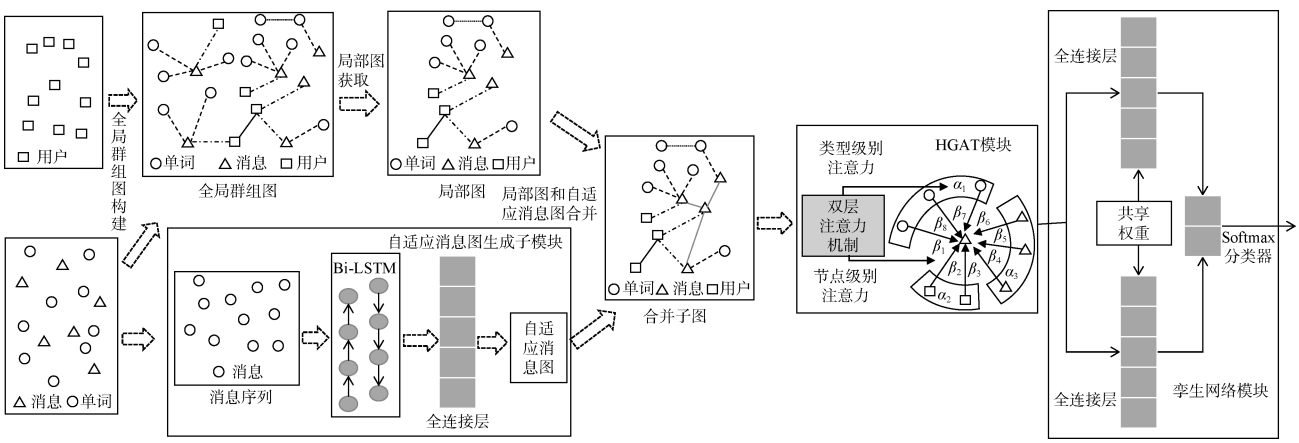

图 2 基于图表示学习的消息回复关系判断方法框架图

Figure 2 Framework graph of the method for identifying the conversation responding relationships using graph representation learning

4.1 图的构建和生成模块

图的构建和生成由两个子模块组成: 全局群组图构建子模块、自适应消息图生成子模块。全局群组图构建子模块基于群组内容构建全局群组图, 自适应消息图生成子模块基于目标任务处理消息序列自适应地生成消息间关系组成的消息图。

4.1.1 全局群组图构建子模块

该模块是基于群组内容的全局信息统计并计算得到全局群组图。图中有 3 种类型的节点: 用户节点、消息节点、单词节点。群组中的所有用户构成用户节点集, 群组用户发送的所有消息构成消息节点集, 所有消息中出现的全部单词构成单词节点集。图中有 4 种类型的边: 用户和用户之间的边、用户和消息之间的边、消息和单词之间的边、单词和单词之间的边。用户和用户之间采用基于时间窗口的全局用户点互信息方法(Pointwise Mutual Information,

PMI)[26]计算用户间的共现权重作为边权重; 根据用户发送消息, 在用户和用户发送的消息之间添加边; 消息和单词之间采用单词的全局 TF-IDF(Term Frequency-Inverse Document Frequency, TF-IDF)[27]计算单词的重要性权重作为边权重; 单词和单词之采用基于滑动窗口的全局单词 PMI 方法计算单词间的共现权重作为边权重。全局群组图如图 3 所示, 图中有 3 种类型的节点和 4 种类型的边, 分别使用不同类型的图形和线表示, 单词节点、消息节点和用户节点分别使用圆形、三角形和正方形表示, 用户与用户、用户与消息、消息与单词和单词与单词之间的边分别使用实线、点与线段、线段和点组成的直线表示对应节点对的节点间的边。

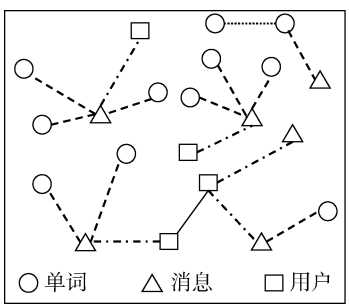

图 3 群组图

Figure 3 Group graph

通过计算节点和节点之间的边权重得到图 G 的邻接矩阵 $\boldsymbol{A}$, 邻接矩阵 $\boldsymbol{A}$ 中的 A_{ij} 元素计算公式为:

$$A_{ij}=\begin{cases} PMI(i,j) & i,j \text{ 都是单词节点;} \\ TF-IDF(i,j) & i,j \text{其中一个是消息节点，另一个是单词节点;} \\ 1 & i,j \text{ 是同一节点或} i,j \text{其中一个是用户节点，另一个是消息节点;} \\ PMI_{time\_window}(i,j) & i,j \text{都是用户节点;} \\ 0 & \text{其他;} \end{cases} \tag{1}$$

1) 成对单词节点间的权重计算

成对单词节点间的权重计算公式 $PMI(i,j)$ 为:

$$PMI(i,j)=\log\frac{p(i,j)}{p(i)p(j)} \tag{2}$$

$$p(i,j)=\frac{Q(i,j)}{Q} \tag{3}$$

$$p(i)=\frac{Q(i)}{Q} \tag{4}$$

Q 是群组消息中所有滑动窗口的总数量, $Q(i)$ 表示滑动窗口中包括单词 i 滑动窗口的数量, $Q(i,j)$ 表示滑动窗口中同时包含单词 i 和单词 j 的滑动窗口的数量。

2) 消息和单词节点之间的权重计算

消息和单词节点之间的单词重要性权重计算公式 $TF-IDF(i,j)$ 为:

$$TF-IDF(i,j)=TF_{ij}\times IDF_{ij} \tag{5}$$

$$TF_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{6}$$

$$IDF_{ij}=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|} \tag{7}$$

TF_{ij} 是单词 i 在消息 j 中出现的频率, n_{ij} 表示单词 i 在消息 j 中出现的次数, $\sum_k n_{kj}$ 表示消息 j 中所有单词出现的次数累加和; IDF_{ij} 表示单词 i 的逆文档频率, $|D|$ 表示聊天群组中的消息总数, $\left|\left\{j:t_i\in d_j\right\}\right|$ 表示包含单词 j 的消息数量。

3) 成对用户节点间的权重计算

成对用户节点间的共现权重计算公式 $PMI_{time\_window}(i,j)$ 为:

$$PMI_{time\_window}(i,j)=\log\frac{p_u(i,j)}{p_u(i)p(j)} \tag{8}$$

$$p_u(i,j)=\frac{Q_u(i,j)}{Q_u} \tag{9}$$

$$p_u(i)=\frac{Q_u(i)}{Q_u} \tag{10}$$

Q_u 是群组消息序列对应的用户序列中时间维度上滑动窗口的总数量, $Q_u(i)$ 表示滑动窗口中包括用户 i 滑动窗口的数量, $Q_u(i,j)$ 表示滑动窗口中同时包含用户 i 和用户 j 的滑动窗口数量。

最终得到全局群组图, 全局群组图表示为 $G=(V,E,X)$, 其中 $V(|V|=N)$ 和 E 分别是节点集和边集。$X=[x_1,x_2,\ldots,x_N]^T\in R^{N\times M}$ 是由消息和用户和单词节点的特征向量组成的矩阵, 其中 $\boldsymbol{x}_i$ 表示节点 i 的特征向量, M 是特征向量的维数。图的拓扑结构由邻接矩阵 $\boldsymbol{A}\in R^{N\times N}$ 表示, 假设每个节点都与其自身相连, 邻接矩阵 $\boldsymbol{A}$ 的对角元素设置为 1。引入 $\boldsymbol{D}\in R^{N\times N}$ 作为图的度矩阵, 其中 $D_{ii}=\sum A_{ij}$ 。最后计算得到归一化拉普拉斯矩阵 $\tilde{\boldsymbol{A}}=D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$。

4.1.2 自适应消息图生成子模块

自适应消息图生成子模块是基于目标任务通过处理输入的消息序列自适应地生成消息间关系组成

的消息图, 保存捕捉到的消息序列中消息间关系。自适应消息图由消息和消息之间的关系构成, 消息图中的节点为消息, 消息图中的边表示消息和消息之间的相关性强弱。现有的研究方法在构建图结构方面, 大部分是通过距离度量来计算节点对的节点间的相似性或相关性[1-2], 例如点积、余弦距离或欧几里得距离。这种计算图的方式, 一方面需求花费 $O(N^2)$ 级别的高时间和空间复杂度, 计算时间和内存成本随着图中节点数量的增加呈二次增长, 限制模型处理大图的能力[28]。另一方面, 点积、余弦距离或欧几里得距离这些距离度量方法并不适用于所有问题, 不同任务需要提取节点间不同的关系, 面向任务的节点对关系的生成方法更具有适用性和通用性。

自适应消息图生成子模块根据下游任务生成消息节点间关系组成的消息图结构。自适应消息图生成子模块的输入为群组的一个消息序列, 消息序列为 $S_1, S_2, \cdots, S_N, S_{N+1}$, 序列长度为 $N+1$, S_i 表示序列中的第 i 条消息, 为了学习序列中消息之间的前后位置关系以及消息间的语义关系, 采用双向长短期记忆网络(Bi-directional Long-Short Term Memory, Bi-LSTM)[29]从正序和反序两个方向学习得到所有位置的隐藏表示和最后的输出, 通过一个全连接层将 Bi-LSTM 的全部输出进行矩阵运算得到 $(N+1)^2$ 维度的输出矩阵, 作为消息序列的自适应消息图对应的邻接矩阵 $\boldsymbol{A}_{ada}$。

4.2 局部图获取和合并模块

局部图的获取和合并模块由两个子模块组成: 局部图获取子模块、局部图和自适应消息图合并子模块。局部图获取子模块从全局群组图抽取与当前消息序列相关的节点和节点间关系组成局部图, 局部图和自适应消息图合并子模块合并局部图信息和自适应消息图信息得到合并子图, 接下来对两个子模块分别进行详细介绍。

4.2.1 局部图获取子模块

聊天群组从被创建开始, 群组中每天都可能产生大量消息, 每条消息都有对应的时间戳和发送该消息的群组用户。分析群组消息时, 当前消息并不与所有群组消息都相关, 当前消息回复的消息只存在于当前消息的上文中(当前消息为会话的起始消息时, 上文中不存在当前消息回复的消息), 并且是与当前消息语义相关的消息, 群组消息关系具有空间相近性和语义相关性。通过考虑指定数量可能性大的上文范围, 缩小考虑的消息数量, 可以大大减少噪声信息, 提高当前消息回复关系判断的任务效果。从全局图中提取与当前消息序列相关的消息、用户和单词构成局部图。当前消息序列的局部图由以下节点组成:

- 消息节点: 当前消息及上文连续 N 条消息组成的大小为 $N+1$ 的窗口内的消息。
- 用户节点: 与消息节点的相关发送消息的用户。
- 单词节点: 出现在消息节点中的单词。

局部图相比于全局图, 有效地减少干扰消息, 降低计算时间和内存消耗。对大规模图结构的任务场景有更强的适用性。

4.2.2 局部图和自适应消息图合并子模块

局部图和自适应消息图分别包含不同的节点关系信息, 两种信息互相补充, 合并两种图结构信息作为异质图注意力网络模块输入的图结构。局部图获取子模块从全局群组图中抽取得到相关消息、用户和单词的节点和它们之间的关系组成局部图。采用滑动窗口的方式遍历群组消息, 当前消息及上文 N 条消息组成的大小为 $N+1$ 的滑动窗口得到消息序列 $S_1, S_2, \cdots, S_N, S_{N+1}$, 自适应消息图生成子模块基于该消息序列生成任务相关的自适应消息图。局部图和自适应消息图合并子模块合并自适应消息图和局部图得到最终的合并子图 $\tilde{A}$, 作为异质图注意力网络模块的输入。

4.3 异质图注意力网络(HGAT)模块

本部分使用 HGAT 模块来学习合并得到的合并子图信息, HGAT 采用节点层和类型层组成的双层注意力机制, 学习群组合并子图结构中不同相邻节点的重要性以及不同类型节点对当前节点的重要性进行节点表示学习。

合并子图由不同类型的节点组成, 异质图注意力网络采用异质图卷积方式处理异质合并子图。异质合并子图中不同类型节点拥有不同的向量表示, 利用对应的变换矩阵将不同类型的向量表示映射到公共向量表示空间。异质图卷积公式如下所示:

$$H^{l+1} = \sigma\left(\sum_{\tau \in \Gamma} \tilde{\boldsymbol{A}}_{\tau} \cdot \boldsymbol{H}_{\tau}^{l} \cdot \boldsymbol{W}_{\tau}^{l}\right) \tag{11}$$

上式中 $\tilde{\boldsymbol{A}}_{\tau} \in R^{N \times N_{\tau}}$ 是合并子图邻接矩阵 $\tilde{\boldsymbol{A}}$ 的子矩阵, 子矩阵中行表示节点, 列表示 τ 类型的相邻节点。τ 类型的转换矩阵 $\boldsymbol{W}_{\tau}^{l} \in R^{q^{l} \times q^{l+1}}$ 将使用 $\tilde{\boldsymbol{A}}_{\tau}$ 聚合 τ 类型相邻节点的向量表示 $\boldsymbol{H}_{\tau}^{l}$, 变换矩阵 $\boldsymbol{W}_{\tau}^{l}$ 考虑不同类型向量表示空间上的差异, 将它们映射到公共向量表示空间 $R^{q^{l+1}}$。将得到所有类型向量表示累加计算得到节点的第 $l+1$ 层向量表示 $\boldsymbol{H}^{l+1}$。$H_{\tau}^{0} = X_{\tau}$,

$\boldsymbol{X}_\tau$ 为 τ 类型节点初始向量表示。

给定一个特定类型的节点, 不同类型的相邻节点对当前节点产生不同程度的影响。相同类型的相邻节点可能会携带更多有用的信息, 同一类型的不同相邻节点也可能具有不同的重要程度。HGAT 使用一种的双层注意机制获取节点级和类型级的不同重要性。接下来分别对类型级和节点级的注意力权重计算方法介绍。

1) 类型级的注意力权重计算

类型级的注意力计算不同类型相邻节点对当前节点的重要性权重。给定一个特定类型的节点 i, 首先计算 τ 类型的向量表示 $\boldsymbol{h}_\tau = \sum_{i'} \tilde{A}_{ii'} h_{i'}$, 相邻节点 $h_{i'} \in N_i$ 为节点 i 的 τ 类型邻居节点, h_τ 是节点 i 的 τ 类型相邻节点向量表示 $\boldsymbol{h}_{i'}$ 的总和。基于当前节点向量 $\boldsymbol{h}_i$ 和类型 τ 的向量表示 $\boldsymbol{h}_\tau$ 计算类型级注意力得分, 计算公式如下所示:

$$a_\tau = \sigma(\mu_\tau^T \cdot [h_i \| h_\tau]) \tag{12}$$

上式中 μ_τ^T 是 τ 类型的注意力权重向量, $\|$ 表示向量拼接, σ 表示激活函数。最后通过使用 *soft*max 函数对所有类型的注意力得分进行归一化获得类型级的注意力权重:

$$\alpha_\tau = \frac{\exp(a_\tau)}{\sum_{\tau' \in \Gamma} \exp(a_{\tau'})} \tag{13}$$

2) 节点级的注意力权重计算

节点级的注意力权重计算获得不同相邻节点对当前节点的重要性, 降低噪声节点的干扰。给定类型 τ 的节点 i 和类型 τ' 的相邻节点 $i' \in N_i$, 基于节点 i 的向量表示 h_i 和邻居节点的向量表示 $h_{i'}$ 计算节点级注意力得分, 节点 $h_{i'}$ 类型级的注意力权重为 $\alpha_{\tau'}$:

$$b_{ii'} = \sigma(\boldsymbol{\delta}^T \cdot \alpha_{\tau'} [h_i \| h_{i'}]) \tag{14}$$

上式中 $\boldsymbol{\delta}^T$ 是节点级的注意力权重向量。最后使用 *soft*max 函数对节点级注意力得分进行归一化得到节点级的注意力权重:

$$\beta_{ii'} = \frac{\exp(b_{ii'})}{\sum_{j \in N_i} \exp(b_{ij})} \tag{15}$$

最后将包括类型级和节点级的双层注意力机制融合到异质图卷积公式中, 得到以下公式:

$$\boldsymbol{H}^{l+1} = \sigma(\sum_{\tau \in \boldsymbol{\Gamma}} \boldsymbol{B}_\tau \cdot H_\tau^l \cdot W_\tau^l) \tag{16}$$

上式中 $\boldsymbol{B}_\tau$ 表示注意力权重矩阵, $\boldsymbol{B}_\tau$ 的第 i 行第 i' 列的元素为 $b_{ii'}$。

4.4 孪生网络(Siamese network)模块

孪生网络采用网络参数共享权值方式将两个不同的输入映射到同一向量空间, 进行向量转换、非线性处理和拼接操作, 最终输入到目标函数中判断两条消息的相关性。

需要判断回复关系的消息对为 (S_i, S_{N+1}), 通过 HGAT 模块学习图结构信息和相邻节点信息得到向量表示输出 $(\boldsymbol{H}_i^l, \boldsymbol{H}_{N+1}^l)$, $\boldsymbol{H}_i^l$ 为 HAGT 模块最后一层输出的第 i 条消息(上文消息)的向量表示, $\boldsymbol{H}_{N+1}^l$ 为 HAGT 模块最后一层输出的第 $N+1$ 条消息(当前消息)的向量表示。$(\boldsymbol{H}_i^l, \boldsymbol{H}_{N+1}^l)$ 输入到孪生网络射到相同的向量空间得到向量表示对 $(\boldsymbol{u}, \boldsymbol{v})$, 对 $(\boldsymbol{u}, \boldsymbol{v})$ 进行多种拼接操作获取更多的交互信息, 将拼接结果输入到目标函数, 判断消息间是否存在回复关系。

为了获取数据对更多的交互信息, 目标函数的输入为进行多种拼接操作处理后的数据对, 将消息向量表示 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 与向量元素差 $|\boldsymbol{u} - \boldsymbol{v}|$ 拼接起来, 最终的目标函数公式如下所示:

$$\hat{y} = \text{softmax}\left((\boldsymbol{u}, \boldsymbol{v}, |\boldsymbol{u} - \boldsymbol{v}|) \cdot W_o\right) \tag{17}$$

上式 $W_o \in R^{3nxk}$ 为权重矩阵, 其中 n 是消息向量表示的嵌入维数, $3n$ 是拼接操作后向量的最终维度, k 是标记数据的类别数量。将目标函数的输出和真实标签通过交叉熵损失函数优化和训练模型, 交叉熵损失函数如下所示:

$$\text{Loss} = \frac{1}{N_{pair}} \sum_i -[y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{18}$$

上式中 N_{pair} 为消息对的数量, y_i 为第 i 对消息的真实标签, 真实标签为 1 表示第 i 对消息间存在回复关系, 为 0 表示第 i 对消息间不存在回复关系。$\hat{y}_i$ 表示本模型输出的第 i 对消息间的回复关系预测值。

5 实验

本节是在 IRC(Internet Relay Chat)数据集[24]和标注数据集上进行实验, 评估本模型在消息回复关系判断任务上的实验效果, 与对比模型和消融实验数据对比分析实验结果。本模型在 IRC 数据集上单词节点的向量表示使用 IRC 数据集 glove-ubuntu.txt 文件 GloVe-vectors 初始化, 消息节点的初始化向量表示通过累加该消息中所有单词节点向量表示求平均得到, 用户节点的初始化向量通过累加该用户发布的所有消息向量表示求平均得到。在标注数据集的微信、QQ 数据集上, 数据集中消息主要为中文形式,

单词节点向量表示使用 BERT 中文预训练模型 bert-base-chinese[7]初始化。为了保证与 IRC 数据集处理保持一致性, 消息节点和用户节点的向量表示采用与 IRC 数据集相同的处理方法初始化。接下来按顺序从实验数据、评价指标、对比实验和消融实验、实验结果、消息对向量距离分析、用户向量分析六部分进行详细介绍。

5.1 实验数据

5.1.1 IRC 数据集

本方法使用了 IRC 数据集评估实验效果, IRC 数据集是从 Linux 和 Ubuntu 的 Internet Relay Chat Log[25]收集的数据。IRC 数据集包含多个日期的聊天数据, 每个日期聊天数据由 1500 条左右消息数据组成。IRC 数据集的训练数据是由 153 个日期的数据组成, 日期之间是不连续的, 每个日期的数据包含四种类型的数据文件, 四种数据文件分别为.train-c.raw.txt、.train-c.ascii.txt、.train-c.tok.txt、.train-c.annotation.txt。.train-c.raw.txt 是从 IRC 聊天室中爬取特定日期连续消息数据组成的原始文件, .train-c.ascii.txt 是将.train-c.raw.txt 文件中的内容转换为 ascii 格式处理后的文件, .train-c.tok.txt 文件是使用自动标记符和占位符替换.train-c.ascii.txt 文件中的稀有单词处理后的文件, .train-c.annotation.txt 是标注数据文件。标注数据文件中每一行表示两条消息间的回复关系, 例如: “1002 1003 -”表示日志中消息 1003 是消息 1002 的回复。每个日期文件中的消息数据都是从 0 开始计数, 每条消息都是.train-c.raw.txt 文件中数据的一行。一条消息可以链接到它之前一条或多条消息, 每个链接关系分别在标注数据文件中给出。当消息链接到自身时, 表示它是一个新对话的开始, 系统消息也被计数和标注, 系统消息的标注也是链接到系统消息本身。标注数据文件从 1000 开始标注消息间的回复关系, 前 1000 条消息(包括第 1000 消息)没有标注回复关系, 只为标注数据提供上下文信息。IRC 数据集中提供 glove-ubuntu.txt 文件, 文件是在所有已爬取的 Ubuntu IRC 日志.tok.txt 文件上训练得到 Glove-vectors。

IRC 数据集中的数据是从 Internet Relay Chat 聊天室中爬取得到的, 聊天室中的用户可以不断加入和退出, 用户数量和参与用户是变化的。图模型输入的图结构是静态、全局的, 模型开始训练前需要将群组内容转换为图结构并保存下来, 图结构在模型训练过程中不变化。为了满足本方法图结构数据组成, 本部分合并多个日期文件数据组成包含训练集和验证集的数据集。

本方法旨在学习聊天群组中消息间的回复关系, 然而数据中系统消息和链接到自身的消息(一个新对话的开始)不存在回复关系, 系统消息提供的信息有限, 还会带来噪声。在数据处理的过程中, 将数据中的系统消息删除, 并将标注数据中链接到自身的非系统消息(一个新对话的开始)归为上下文消息, 为其他标注数据提供上下文信息, 不作为训练和验证标注数据使用。

实验数据来自于 IRC 数据集的训练集, 从 IRC 数据集的训练集中抽取两个子数据集: 子数据集 a 和子数据集 b, 实验数据的统计结果如表 1 所示。

表 1 IRC 数据集统计结果

Table 1 Statistics results of IRC dataset

数据集	包含日期文件数	消息数量	训练集包含日期文件数	训练集回复关系数量	训练集上下文消息数量	验证集包含日期文件数	验证集回复关系数量	验证集上下文消息数量
子数据集 a	10	12050	8	2479	7089	2	731	1751
子数据集 b	5	5936	4	1459	3308	1	333	836

数据集 a 数据集由训练集中前 10 个连续日期的聊天数据组成, 10 个日期中的前 8 个日期数据文件作为训练数据, 后 2 个数据文件作为验证数据。

数据集 b 数据集由 5 个连续日期的聊天数据组成, 5 个日期中的前 4 个日期数据文件作为训练数据, 最后 1 个日期数据文件作为验证数据。

2 个数据集中包含的聊天数据不交叉, 数据集间不存在重复数据。

5.1.2 标注数据集

标注数据集包括微信和 QQ 两个子标注数据集。微信和 QQ 两个子标注数据集是分别从社交工具微信和 QQ 中爬取的聊天群组历史数据, 通过人工标注群组消息间的回复关系构建得到。为了保证人工标注群组消息间回复关系的准确性, 每个标注子数据集的标注工作由两位标注人员标注, 在标注过程中, 产生不一致的标注结果时, 交予第三位标注者确定

最终的标注标签。标注人员标注过程中参考 IRC 数据集的标注规则进行标注, 标注规则如下所示:

- 如果当前消息没有回复任何消息, 是一个新对话的开始时, 将当前消息链接到消息本身, 回复关系标注为–1。
- 如果当前消息是感叹词、噪音消息, 将当前消息不链接到任何消息, 不作为实验数据, 回复关系标注为–2。
- 如果当前消息回复前面多条消息, 将当前消息链接到前面多条消息, 回复关系标注为前面多条消息的 ID。
- 如果当前用户发送多条消息回复前面的消息, 包括以下两种情况:
  - 当前用户发送的每条消息独立回复前面消息, 将每条消息链接到前面被回复的消息, 回复关系标注为被回复消息的 ID。
  - 当前用户发送每条消息是连续相关的, 后一条消息回复当前用户前一条消息, 将每条消息链接到当前用户的前一条消息, 回复关系标注为前一条消息的 ID。
  - 当前用户发送多条一样的消息时, 将重复消息链接到第一条原始消息, 回复关系标注为第一条消息的消息 ID。(比如: 一个问题被问了 3 次, 第 2 次和第 3 次连接到第 1 次)

表 2 标注数据集统计结果

Table 2 Label results of IRC dataset

标注数据集	群组个数	消息类型	消息数量	训练集回复关系数量	验证集回复关系数量
微信	1	文本	2729	1910	819
QQ	1	文本	2996	2097	899

按照上述标注规则对群组数据标注, 得到两个标注子数据集: 微信标注子数据集、QQ 标注子数据集, 标注数据集的统计结果如表 2 所示。

5.1.3 正负样本构建

从消息序列中抽取当前消息回复的消息和未回复的消息与当前消息组成消息对, 作为正负样本, 例如 (S_i, S_{N+1}) 和 (S_j, S_{N+1}), 其中 S_i 是 S_{N+1} 回复的消息, S_j 是 S_{N+1} 未回复的消息中随机抽取的一条。实验数据中存在回复关系两条消息组成消息对作为模型输入的正样本, 负样本通过以下方式获得: 滑动窗口大小为 N, S_{N+1} 为当前消息, 消息序列 $S_1, S_2, \cdots, S_{N+1}$ 中前 N 条消息为上文消息, S_i 为上下文消息中的一条, 随机从消息序列 $S_1, S_2, \cdots, S_{N+1}$(不包含 S_i)中抽取一条消息 S_j, 与当前消息组成消息对 (S_j, S_{N+1}) 作为负样本。为了保证滑动窗口中的每一条不存在回复关系的消息被学习到, 模型训练的过程最少训练 N 个 Epoch。

5.2 评价指标

基于图表示学习的消息回复关系判断方法是一种对消息间回复关系判断的二分类方法, 选择分类评价指标召回率(Recall)、精准率(Precision)、F1、准确率(ACC)四种方法作为本实验的评价指标, 下面对 Recall、Precision、F1、ACC 4 种方法详细介绍。

召回率(Recall): 召回率(Recall)为正确预测为 1 样本数量占真实为 1 样本数量的比率。

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{19}$$

上式中 TP 是预测为 1 预测正确的样本数量, FN 是预测为 0 预测错误的样本数量。

精确率(Precision): 精确率(Precision)为正确预测为 1 样本数量占预测为 1 样本数量的比率。

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{20}$$

上式中 FP 是预测为 1 预测错误的样本数量。

F1: F1 是由召回率(Recall)和精确率(Precision)计算得到, 同时考虑召回率和精确率, 相当于 Recall 和 Precision 的调和平均值。F1 的公式如下:

$$F1 = \frac{2 \times \mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{21}$$

准确率(ACC): ACC 是预测正确样本数量占预测样本的比率。ACC 计算公式如下:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{22}$$

上式中 TN 是预测为 0 正确的样本数量。

5.3 对比实验和消融实验

本部分在 IRC 数据集和标注数据集上评估本模型、对比模型和消融实验在消息间回复关系判断任务上的实验效果。本方法属于自然语言推理(NLI)任务实现的一种, 为了比较在 NLI 任务上的实验效果, 选择 NLI 方向的 ESIM 作为对比算法。本模型在 IRC 数据集上使用 glove-ubuntu.txt 文件中的 GloVe-vectors 初始化输入的向量表示, 在标注数据集上使

用 BERT 中文预训练模型 bert-base-chinese 初始化输入的向量表示, 基于初始化的向量表示 HGAT 模块对节点表示进一步更新。为了验证本方法的表示学习能力, 选择表示学习算法 TextCNN(Convolutional Neural Networks for Sentence Classification, TextCNN)[30]和 BERT 作为对比算法, TextCNN 采用和本方法的节点相同初始化方式, BERT 使用预训练模型初始化向量表示。对比算法包括 BERT、TextCNN、ESIM 3 种, BERT、TextCNN 属于向量表示学习算法, ESIM 是为自然语言推理设计的序列式推理算法。

为了验证提出的优化模块有效性, 进一步实现了本方法的多个消融实验。消融实验是去掉模型中部分优化模块后的实验, 最后通过与本模型和其他消融实验对比实验数据, 验证优化模块的有效性。本模型的消融实验包括去掉自适应消息矩阵、去掉孪生网络、同时去掉自适应消息矩阵和孪生网络三种消融实验。

下面对上述对比算法和消融实验进行详细介绍。

BERT: 使用预训练模型初始化的 BERT 作为对比模型, 在 NSP(Next Sentence Prediction)任务上使用数据集的训练集微调 BERT 预训练模型 N 轮, 统计模型在验证集上的评估数据作为实验结果。

TextCNN: TextCNN 是一种利用卷积神经网络对文本进行分类的算法, 可以获取文本的局部相关性特征。使用 TextCNN 分别获取消息对两条消息的特征向量, 拼接两条消息的特征向量转换后输入目标函数, 判断消息对两条消息间的回复关系。

ESIM: ESIM 是一种专为自然语言推理设计的综合 Bi-LSTM 和注意力机制实现局部推理和全局推理相结合的序列式推理算法。ESIM 采用和本模型相同的初始化方式。

消融实验介绍:

-自适应消息图: 去掉自适应消息图生成子模块后的实验。

-孪生网络: 去掉孪生网络模块后的实验。

-自适应消息图-孪生网络: 同时去掉自适应消息图生成子模块和孪生网络模块后的实验。

5.4 实验结果

接下来介绍本模型、对比模型和消融实验的参数设置: 学习率(Learning rate, lr)=1e-5; BERT、TextCNN、ESIM 模型中的其他参数设置为该模型的默认值。为了避免极端情况发生, 所有模型和实验运行 10 次, 窗口大小 N 设置为 40, 每次运行 40 个 Epoch, 计算 10 次平均结果作为实验的最终结果。

5.4.1 对比实验结果

本部分是在 IRC 数据集和标注数据集上评估本模型与多个对比模型在消息间回复关系判断任务上的实验效果, 分别在两个数据集的验证集上统计评估数据作为实验结果, 并对实验结果进行分析, 比较在消息间回复关系判断任务上的效果。

1) IRC 数据集上的对比实验

本部分是验证本模型和对比模型在 IRC 数据集上的实验效果, 统计所有模型在 4 个评估指标上的实验数据, 并将实验数据以表格的形式呈现。为了更直观地显示各模型的实验数据, 将 IRC 数据集的 2 个子数据集 4 个评估指标上的最优结果进行加粗处理, 方便查看。本模型和对比模型在 IRC 数据集 4 个评估指标上的统计结果如表 3 所示。

表 3 IRC 数据集评估结果

Table 3 Evaluation results of IRC dataset

数据集	评估指标	BERT	Text CNN	ESIM	本模型
数据集 a	Recall	71.88	49.03	70.30	**80.64**
	Precision	63.47	43.04	44.88	**71.90**
	ACC	69.51	54.59	51.91	**82.15**
	F1	67.41	45.84	54.78	**76.01**
数据集 b	Recall	63.41	40.92	54.54	**90.09**
	Precision	55.35	40.75	44.76	**75.70**
	ACC	64.48	56.59	56.06	**83.15**
	F1	59.10	40.83	49.16	**82.45**

为了验证模型的显著性, 在数据集 a 和数据集 b 上实现了 10-fold 交叉验证实验, 得到了所有模型在 ACC、F1 指标上的 10 次实验结果。在 ACC 和 F1 指标上, 基于得到的 10 次实验结果, 对我们模型和对比模型进行了显著性检验实验, P 值小于 0.05, 结果表明我们的模型是显著的。

2) 标注数据集上的对比实验

本部分是验证本模型和对比模型在标注数据集

上的实验结果, 统计所有模型在 4 个评估指标上的实验数据, 并将实验数据以表格的形式呈现。为了更直观地显示各模型的实验结果, 将标注数据集的微信和QQ标注子数据集4个评估指标上的最优结果进行加粗处理, 方便查看。本模型和对比模型在标注数据集 4 个评估指标上的统计结果如表 4 所示。

表 4 标注数据集评估结果

Table 4 Evaluation results of Label dataset

数据集	评估指标	BERT	TextCNN	ESIM	本模型
微信	Recall	47.34	21.48	49.60	**78.67**
	Precision	70.21	27.57	38.84	**71.00**
	ACC	63.04	59.54	51.63	**86.09**
	F1	56.55	24.14	43.56	**74.63**
QQ	Recall	50.87	34.59	52.71	**79.16**
	Precision	66.26	37.04	42.66	**68.59**
	ACC	74.00	60.60	57.71	**83.48**
	F1	57.55	35.77	47.15	**73.49**

为了验证模型的显著性, 在数据集 a 和数据集 b 上实现了 10-fold 交叉验证实验, 得到了所有模型在 ACC、F1 指标上的 10 次实验结果。在 ACC 和 F1 指标上, 基于得到的 10 次实验结果, 对我们模型和对比模型进行了显著性检验实验, P 值小于 0.05, 结果表明我们的模型是显著的。

3) 实验分析

表 3~表 4 分别展示了本模型和对比模型在 IRC 数据集和标注数据集上消息间回复关系判断任务的实验结果, 从表中可以看到本模型在两个数据集上都取得最优的实验结果, 证明了本方法在群组消息间回复关系判断任务上的有效性。BERT、TextCNN、ESIM 3 个对比模型中, 在两个数据集上的四个评估指标大部分表明 BERT 性能最好, BERT 模型参数使用预训练模型初始化, TextCNN、ESIM 模型参数随机初始化, 实验结果证明预训练模型在编码文本语义信息方面的优势。本模型、TextCNN 和 ESIM 的输入初始化方式相同, 两个数据集上分别采用 IRC 提供的 Glove-ubuntu 和 BERT 中文预训练模型 bert-base-chinese 初始化向量表示。其中 TextCNN 和 ESIM 相比, ESIM 模型的实验结果优于 TextCNN 模型, TextCNN 使用卷积神经网络学习单条消息的局部相关性特征; ESIM 综合 Bi-LSTM 和注意力机制实现局部推理和全局推理的结合, 深度挖掘消息对的交互关系信息。实验结果证明 ESIM 的局部推理和全局推理的综合推理在消息间关系判断方面的有效性。本模型相比于 TextCNN 和 ESIM 模型在两个数据集上都有较大提升, TextCNN 和 ESIM 模型只是简单使用了消息对中两条消息的文本和它们之间的关系信息, 本模型不仅使用了消息对的文本信息, 也融合了用户信息和上下文信息, 证明综合使用群组内容信息能有效提升判别消息回复关系能力。

5.4.2 消融实验结果

消融实验是本方法去掉部分优化模块后的实验, 在 IRC 数据集和标注数据集上实现消融实验并对比实验结果, 验证优化模块的有效性以及对回复判断任务的提升效果。

1) IRC 数据集上的消融实验

本部分是验证消融实验在 IRC 数据集上的实验结果, 统计所有模型在 4 个评估指标上的实验数据, 并将实验数据以表格的形式呈现。为了更直观地显示各模型的实验结果, 将IRC数据集的2个子数据集 4 个评估指标上的最优结果进行加粗处理, 方便查看。消融实验在 IRC 数据集 4 个评估指标上的统计结果如表 5 所示。

2) 标注数据集上的消融实验

本部分是验证消融实验在标注数据集上的实验结果, 统计所有模型在 4 个评估指标上的实验数据, 并将实验数据以表格的形式呈现。为了更直观地显示各模型的实验结果, 将标注数据集的微信和QQ标注子数据集 4 个评估指标上的最优结果进行加粗处理, 方便查看。消融实验在标注数据集 4 个评估指标上的统计结果如表 6 所示。

3) 实验分析

本部分将进一步分析各优化模块在回复关系判断任务上的提升效果, 表 5~表 6 分别展示消融实验在 2 个数据集上 4 个评估指标的实验结果。从表 5-6 中可以看到原模型在两个数据集上几乎都取得最优的实验结果, 证明了提出的两种优化有效地提升了本方法的消息回复关系判断能力。从表 5 观察得到,

表 5 IRC 数据集消融实验评估结果

Table 5 Evaluation results of IRC dataset ablation experiment

数据集	评估指标	-自适应消息图	-孪生网络	-自适应消息图 -孪生网络	原模型
数据集 a	Recall	**88.03**	80.32	87.87	80.64
	Precision	51.82	71.52	47.50	**71.90**
	ACC	47.92	82.15	51.78	**82.15**
	F1	65.23	75.66	61.66	**76.01**
数据集 b	Recall	79.67	89.94	74.26	**90.09**
	Precision	45.39	75.28	41.96	**75.70**
	ACC	50.84	81.90	51.10	**83.15**
	F1	57.83	81.95	53.62	**82.27**

表 6 标注数据集消融实验评估结果

Table 6 Evaluation results of Label dataset ablation experiment

数据集	评估指标	-自适应消息图	-孪生网络	-自适应消息图 -孪生网络	原模型
微信	Recall	**81.91**	77.01	72.18	78.67
	Precision	39.31	66.76	36.77	**71.00**
	ACC	42.37	76.47	45.28	**86.09**
	F1	53.12	71.51	48.72	**74.63**
QQ	Recall	72.49	78.12	70.89	**79.16**
	Precision	39.97	68.21	39.18	**68.59**
	ACC	48.15	**84.16**	47.87	83.48
	F1	51.52	72.82	50.46	**73.49**

3 个消融实验相比于原模型, 在数据集 a 的 F1 指标上分别下降 14.97%、14.63%、0.37%, 在数据集 b 的 F1 指标上分别下降 28.76%、24.76%、0.39%, 去掉自适应消息图模块两个消融实验 F1 指标下降的最多, 证明通过添加自适应消息图生成子模块可以有效提升本方法的消息回复关系判断能力。去掉孪生网络模块的消融实验和原模型相比, F1 值在两个子数据集上下降均不到 1%, 证明孪生网络模块可以提升本方法的消息回复关系判断能力, 只是提升效果有限, F1 评估指标上的提升效果不大。同理可分析表 6 实验结果, 消融实验在微信和 QQ 标注子数据集 F1 评估指标上的实验数据, 验证了提出的优化模块在消息回复关系判断任务上有不同提升效果, 证明了自适应消息图生成子模块可以有效提升本方法在消息回复关系判断任务上的能力。

表 5 中, 数据集 a 中, 去掉自适应消息图生成子模块的消融实验在 Recall 指标上取得最好的结果, 比原模型高出 7.39%, 可是原模型比该消融实验在 Precision 指标上提升 20.08%。原因是由于原模型添加自适应消息图生成子模块学习消息序列的局部消息间的关系, 忽略局部区间外的一些无关消息, 同时可能忽略一些存在关系的消息对, 导致实验结果在 Precision 指标上有巨大提升, 在 Recall 指标上数值有些许下降。去掉自适应消息图生成子模块的消融实验的实验结果在 Recall 指标上数值提升比例小于在 Precision 指标上下降的比例, 完全可以弥补在 Recall 指标上的数值下降, 证明添加自适应消息图子模块的有效性。同理可分析, 表 6 微信数据集中 Recall 评估指标取得最优结果, 其他指标远低于原模型的现象。

在表 6 中, QQ 数据集中去掉孪生网络模块的消融实验取得 ACC 评估指标上的最优结果, 在 ACC 评估指标上相比于原模型多出 0.68%。在 Recall、Precision、F1 3 个指标上, 原模型相比于去掉孪生网络模块的消融实验, 分别高出 1.04%、0.38%、0.6%。实验结果表明去掉孪生网络模块的消融实验在 ACC 评估指标上有些许提升, 添加孪生网络模块有助于提升原模型在消息回复关系判断任务的 Recall、Precision 和 F1 3 个指标上实验效果。

5.5 消息对向量距离分析

本部分对本方法的 HGAT 模块输出的消息向量进行分析, 比较真实标签中存在回复关系消息对和不存在回复关系消息对的两条消息间的向量距离, 分析 HGAT 模块输出的消息向量表示和任务相关性,

进一步评估本方法的表示学习效果。图 4 消息对向量距离分析图以箱线图的形式展示了存在回复关系消息对和未回复消息对的向量距离统计结果, 图中分别标明两种情况下第一四分位数(Q1)、中位数(M)、第三四分位数(Q3)、均值(Mean)、最大值(Max)等评估值, 直观地比较两种情况下的消息对向量距离。从图中可以观察两种情况下的均值(Mean), 回复消息对比未回复消息对的平均向量距离短, 说明在任务学习的过程中 HGAT 模块更新节点向量表示过程中, 消息节点向量表示可以学习消息对的回复关系信息, 使得存在回复关系消息对比未回复消息对的空间距离更接近。从 Q3 和均值的距离来看, 回复消息对向量距离的 Q3 和均值小于未回复消息对向量距离的 Q3 和均值, 说明回复的消息对向量距离的异常值数量较少和异常值异常程度小。经过统计得到回复消息对中异常值个数为 237, 占全部回复消息对的比例为 0.09172。未回复消息对异常值个数为 10560, 占未回复消息对的比例为 0.09367, 满足从图中得到的分析结果。

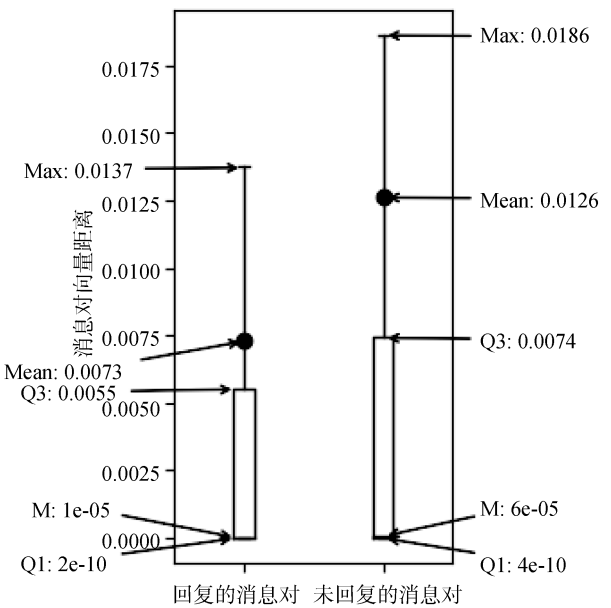

图 4 消息对向量距离分析图

Figure 4 Message pair vector distance analysis graph

5.6 用户向量分析

接下来对本方法的 HGAT 模块输出的用户向量表示进行分析, 分析统计真实标签得到用户间回复次数和用户间向量距离的关系, 评估 HGAT 模块输出的用户向量表示与任务相关性, 进一步验证本方法的表示学习能力。相比于 IRC 数据集, 标注数据集中的微信和 QQ 标注子数据集分别来自一个单独聊天群组, 数据集中的用户流入流出不频繁, 群组用户大部分都是固定的, 可以作为群组用户向量表示的分析对象。通过统计得到 QQ 标注子数据集参与群组聊天的用户有 76 位, 微信标注子数据集中参与群组聊天的用户有 193 位, 微信标注子数据集中参与用户数量更多。为了减小群组用户数量少导致用户分析不全面、不充分的影响, 接下来使用得到的微信标注子数据集用户向量表示进行用户向量表示分析。

为了分析 HGAT 模块输出的用户向量表示, 将用户间向量距离和用户间回复次数结合, 验证随着用户间回复次数增加用户向量间距离的呈现趋势。首先, 对用户间回复次数数量和用户间向量距离进行统计, 计算两个用户间相互回复次数总和作为两个用户间的总回复次数, 将用户间的总回复次数作为横坐标; 计算特定回复次数的所有用户间向量距离, 并计算向量距离累加和的平均值, 作为当前回复次数的用户间向量距离(纵坐标), 其中回复次数为 0 的向量距离是计算不存在回复的所有用户间向量距离累加和的平均值。用户间向量距离和回复次数的关系如图 5 所示。

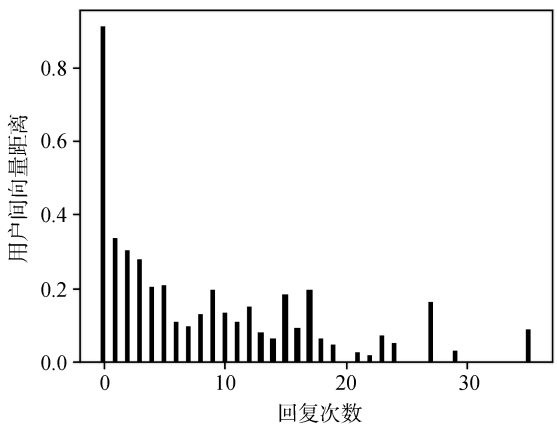

图 5 用户向量分析图

Figure 5 User vector analysis graph

从图 5 观察得到, 随着用户间回复次数的增加, 用户间向量距离整体呈现下降的趋势, 即用户间消息回复次数越多, 用户间距离越相近。由于本模型训练过程中用户数量和群组消息数量相对较少, 模型训练不充分, 少数回复次数较多的大于回复次数较少的用户间向量距离。图中呈现用户间向量距离整体趋势是下降的, 满足最初的方法预期。

6 总结和下一步展望

本论文提出了基于图表示学习的消息回复关系判断方法, 该方法处理群组内容转换为图结构, 通

过图模型综合学习消息的文本信息、发送消息的用户信息和上下文信息判断消息间的回复关系。首先针对群组特性基于群组内容构建了全局群组图和生成自适应消息图, 保存群组消息、用户、消息中单词三种实体信息和它们之间的关系。基于群组聊天的局部性, 从全局群组图中抽取与当前消息序列相关的消息、用户和单词构成局部图, 局部图与自适应消息图合并得到合并子图作为当前消息对输入的图结构。将合并子图输入到 HGAT 模块进行表示学习, 输出融合图结构信息的节点向量表示, 最后通过孪生网络转换消息对的向量空间并预测消息间的回复关系。在消息回复关系的学习过程中得到副产品——优化后的用户向量, 用户向量有效融合了文本信息和回复结构信息, 经过用户向量分析实验验证了用户向量的有效性和合理性。在公开数据集和标注数据集上通过对比实验验证本方法的有效性, 通过消融实验验证各优化模块提升本方法预测消息间回复关系的能力。

本方法是预测群组消息间的回复关系, 预测过程中只考虑消息所属会话部分消息间的局部语义, 没有考虑当前消息所属会话的整体语义, 忽略会话语义的完整性。后续研究会考虑会话语义的完整性, 进一步提升消息间回复关系的判断能力。

致　谢　在此向对本文工作提出指导的老师、同学表示感谢, 以及对提出建议的评审专家表示感谢。

参考文献

[1] Bowman S R, Angeli G, Potts C, et al. A Large Annotated Corpus for Learning Natural Language Inference[C]. *The 2015 Conference on Empirical Methods in Natural Language Processing*, 2015: 632–642.

[2] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100, 000+ Questions for Machine Comprehension of Text[C]. *The 2016 Conference on Empirical Methods in Natural Language Processing*, 2016: 2383–2392.

[3] Li Y D, Wang J, Lin H F, et al. Residual Connected Enhanced Sequential Inference Model for Natural Language Inference[C]. *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering*, 2019: 381-386.

[4] Graves A. Long Short-Term Memory[M]. Studies in Computational Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 37-45.

[5] Chen Q, Zhu X D, Ling Z H, et al. Enhanced LSTM for Natural Language Inference[EB/OL]. 2016: arXiv: 1609.06038[cs.CL]. https://arxiv.org/abs/1609.06038.

[6] Gu J C, Li T D, Liu Q, et al. Pre-Trained and Attention-Based Neural Networks for Building Noetic Task-Oriented Dialogue Systems[EB/OL]. 2020: arXiv: 2004.01940[cs.CL]. https://arxiv.org/abs/2004.01940.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. 2018: arXiv preprint arXiv:1810.04805.

[8] Zhu H H, Nan F, Wang Z G, et al. Who did they Respond To? Conversation Structure Modeling Using Masked Hierarchical Transformer[J]. *The AAAI Conference on Artificial Intelligence*, 2020, 34(5): 9741-9748.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You Need[EB/OL]. 2017: arXiv preprint arXiv:1706.03762.

[10] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[EB/OL]. 2016.

[11] Hu L M, Yang T C, Shi C, et al. Heterogeneous Graph Attention Networks for Semi-Supervised Short Text Classification[C]. *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019: 4823-4832.

[12] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks[EB/OL]. 2019.

[13] Wang H, Lu Z, Li H, et al. A dataset for Research on Short-Text Conversations[C]. *The 2013 Conference on Empirical Methods in Natural Language Processing*, 2013: 935-945.

[14] Shang L F, Lu Z D, Li H. Neural Responding Machine for Short-Text Conversation[EB/OL]. 2015.

[15] Feng S. *Dialogue Model Study and Development In Chat Bot System*[D]. Beijing: Beijing University of Posts and Telecommunications, 2014.
(冯升. 聊天机器人系统的对话理解研究与开发[D]. 北京: 北京邮电大学, 2014.)

[16] Ji Z C, Lu Z D, Li H. An Information Retrieval Approach to Short Text Conversation[EB/OL]. 2014.

[17] Ritter A, Cherry C, Dolan W B. Data-driven Response Generation in Social Media[C]. *The 2011 Conference on Empirical Methods in Natural Language Processing*, 2011: 583-593.

[18] Vinyals O, Le Q. A Neural Conversational Model[EB/OL]. 2015

[19] Mei H Y, Bansal M, Walter M R. Coherent Dialogue with Attention-Based Language Models[EB/OL]. 2016.

[20] Wang M, Lu Z, Li H, et al. Syntax-based Deep Matching of Short Textsl[EB/OL]. 2015: arXiv preprint arXiv:1503.02427.

[21] Fagin R, Kimelfeld B, Reiss F, et al. Document Spanners: A Formal Approach to Information Extraction[J]. Journal of the ACM, 2015, 62(2): 12.

[22] Shen D, Yang Q, Sun J T, et al. Thread Detection In Dynamic Text

Message Streams[C]. *The 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006: 35-42.

[23] Elsner M, Charniak E. You talking to me? a Corpus and Algorithm for Conversation Disentanglement[C]. *ACL-08*: *HLT*, 2008: 834-842.

[24] Kummerfeld J K, Gouravajhala S R, Peper J J, et al. A Large-Scale Corpus for Conversation Disentanglement [EB/OL]. 2018: arXiv preprint arXiv:1810.11118.

[25] IRC ubuntu log. http://irclogs.ubuntu.com/.

[26] Bouma G. Normalized (Pointwise) Mutual Information In Collocation Extraction[J]. *German Society for Computational Linguistics*, 2009: 31-40.

[27] Salton G, Buckley C. Term-Weighting Approaches In Automatic Text Retrieval[J]. *Information Processing & Management*, 1988, 24(5): 513-523.

[28] Wu Z H, Pan S R, Long G D, et al. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 753-763.

[29] Chen T, Xu R F, He Y L, et al. Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN[J]. *Expert Systems With Applications*, 2017, 72: 221-230.

[30] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. *The 2014 Conference on Empirical Methods in Natural Language Processing*, 2014: 1746–1751.

梁永明 于 2018 年在河北工业大学物联网工程专业获得学士学位。现在北京邮电大学网络空间安全专业攻读硕士学位。研究领域为数据挖掘、文本聚类、会话分析等。Email: liangyongming@bupt.edu.cn

田恬 于 2018 年在北京科技大学信息安全专业获得学士学位, 现在北京邮电大学计算机科学与技术专业攻读硕士学位。研究领域为深度学习, 数据挖掘, 虚假信息检测等。Email: tiantian_96727@bupt.edu.cn

杨小雨 于 2018 年在北京邮电大学通信工程专业获得学士学位, 现在北京邮电大学网络空间安全专业攻读硕士学位。研究领域为数据挖掘, 虚假信息检测等。Email: littlehaes@bupt.edu.cn

张熙 于 2012 年在清华大学计算机系获得博士学位。现为北京邮电大学网络空间安全学院副教授, 博士生导师, 可信分布式计算与服务教育部重点实验室副主任。研究领域为社交网络分析、数据挖掘、计算机体系结构。Email: zhangx@bupt.edu.cn

邱莉榕 于 2007 年在中科院计算所获得工学博士学位。现在北京邮电大学计算机学院(国家示范性软件学院)担任教授。研究领域为数据分析与处理、知识表示与知识发现等。Email: qiulirong@bupt.edu.cn