



基于深度学习的多方对话研究综述

张开颜, 张伟男*, 刘挺

哈尔滨工业大学社会计算与信息检索研究中心, 哈尔滨 150001

* 通信作者. E-mail: wnzhang@ir.hit.edu.cn

收稿日期: 2020-06-15; 修回日期: 2020-08-11; 接受日期: 2020-10-14; 网络出版日期: 2021-08-03

摘要 近年来, 随着深度学习技术的广泛应用, 人机对话研究取得了突破性进展. 但是, 目前的人机对话系统大多是在人机双方参与的假设下进行设计的, 而更具挑战性的人机多方对话的研究和应用尚不成熟. 本文将立足于自然语言处理领域, 对近几年基于深度学习的多方对话研究进展进行综述. 首先从人机对话角度出发, 整理多方对话系统的关键问题和已有解决方案; 然后, 梳理基于多方对话的其他自然语言处理任务; 之后, 总结已有多方对话研究的数据集, 并分析现有数据集的局限性和改进方案; 最后, 展望多方对话研究的未来发展趋势.

关键词 自然语言处理, 深度学习, 人机对话, 多方对话

1 引言

近年来, 随着互联网上社交数据的快速增长, 得益于深度学习技术的发展, 人机对话技术取得飞跃式进展, 同时, 因其在虚拟助手和社交机器人等领域的商业价值, 而广受学术界和工业界的关注. 目前人机对话的研究大多聚焦于人机双方参与的场景^[1], 然而, 现实生活中存在不少多方对话情况, 比如社交聊天群组、在线聊天室以及各种网络论坛等^[2]. 在多方对话场景中, 存在多个参与者, 人机双方对话系统将不足以应对这种情况^[3], 具有参与多人对话能力的多方对话系统则应运而生^[4].

在深度学习技术兴起之前, 已经存在一些基于手工规则、模板和机器学习的人机多方对话系统的研究, 但其大多体验效果不佳、领域迁移性差, 无法满足实际的娱乐和商业需求^[3]. Kennington 等^[5]提出了一个基于概率和规则的多方对话管理的游戏机器人, 其主要使用预设的对话模板来生成合适的回复, 可以与多个参与者进行问答游戏. Bayser 等^[4]设计了基于机器学习的人机多方对话系统, 由多个特定金融领域的机器人“专家”组成, 用户可以参与其中咨询金融理财问题, 其主要通过状态栈的方式进行多个机器人的对话管理, 并结合支持向量机等进行决策. Żarkowski^[6]介绍了社交机器人 EMYS 参与多方对话的实验系统, 其中机器人主导问答游戏, 两个参与者与机器人进行互动, 实验中主要使用手工制定的规则来进行对话轮转的控制. 除此以外, 还有一些多智能体系统的研究, 以及计算机视觉领

引用格式: 张开颜, 张伟男, 刘挺. 基于深度学习的多方对话研究综述. 中国科学: 信息科学, 2021, 51: 1217–1232, doi: 10.1360/SSI-2020-0176
Zhang K Y, Zhang W-N, Liu T. A survey of multi-party dialogue research based on deep learning (in Chinese). Sci Sin Inform, 2021, 51: 1217–1232, doi: 10.1360/SSI-2020-0176

域的研究, 这不在本文讨论范围内. 这些研究大多聚焦于多方对话管理和系统搭建, 而这些并不在自然语言处理领域范畴之内.

得益于深度学习技术的广泛应用, 自然语言处理领域已经展开不少多方对话研究. Ouchi 等^[7] 基于 Ubuntu Internet Relay Chat (IRC) Logs, 发布了一个多方对话数据集, 并进行了说话对象和回复选择的研究. Zhang 等^[8] 和 Zhu 等^[9] 在该数据集上进行了说话对象识别的研究, 进一步提升了实验效果. Hu 等^[10] 以及 Liu 等^[11] 在该数据集上, 进行了多方对话回复生成研究. 与此同时, 随着微信、Twitter 等社交媒体软件的广泛使用, 多方对话数据与日俱增, 许多自然语言处理任务也开始基于多方对话数据进行. Sun 等^[12] 在英语等级考试的听力阅读理解数据上, 进行多方对话的阅读理解研究. Chen 等^[13] 基于《老友记》(即 *Friends*) 对话数据进行多方对话情感识别的探索.

深度学习技术在多方对话研究上已经取得了一些进展, 但由于研究时间短、新的研究成果不断涌现, 已有的多方对话综述大多年份久远且着眼于多方对话系统构建和多方对话的社会角色分析. Traum^[14] 探讨了多方对话中参与角色和交互管理等方面的关键问题, 但是只进行了理论上的分析和建模, 而没有多方对话系统实现的说明. Uthus 等^[3] 介绍了多方对话背景、相关问题和解决方法, 以及研究趋势和未解决的问题, 受限于技术发展, 文中的模型和方法大多基于规则和机器学习. Bayser 等^[4] 总结了设计和构建多方对话系统的挑战和技术水平, 仍侧重于系统设计, 并且最后提出了多方对话混合结构, 但仅使用了规则和机器学习等方法进行实现. Seering 等^[2] 基于目前已有的聊天机器人文献, 对相关研究进行了全面的回顾, 最终给出了多方对话机器人在社交中可能扮演的角色和起到的作用.

本文将立足于自然语言处理领域, 对基于深度学习的多方对话研究进行回顾和梳理. 具体来说, 首先从人机对话角度出发, 介绍基于深度学习的多方对话系统, 并总结关键问题和已有研究方法. 然后, 介绍基于多方对话的自然语言处理任务, 包括机器阅读理解、情感分析等. 紧接着, 对已有的多方对话数据集进行总结, 比较分析各类型数据集的特点和局限性, 并给出一些改进方案. 最后, 展望多方对话发展方向, 为未来的研究工作奠定基础.

2 基于深度学习的多方对话系统

基于深度学习的多方对话系统以大规模多方对话数据作为训练语料, 利用深度学习算法学习多方对话模式. 在多方对话场景中, 整个对话历史可以抽象表示为一个三元组集合^[7]:

$$\mathcal{C} = [(a_{\text{sender}}^{(t)}, a_{\text{addressee}}^{(t)}, u^{(t)})]_{t=1}^T, \quad (1)$$

其中 $(a_{\text{sender}}^{(t)}, a_{\text{addressee}}^{(t)}, u^{(t)})$ 表示, 第 t 轮对话时, $a_{\text{sender}}^{(t)}$ 回复 $a_{\text{addressee}}^{(t)}$ 语句 $u^{(t)}$. 根据场景需求, 说话对象 $a_{\text{addressee}}^{(t)}$ 可以为空, 即说话对象不显式存在.

由此, 多方对话系统可以形式化描述为: 给定对话历史 $\mathcal{C}$ 作为输入, 计算机 (即 $a_{\text{sender}}^{(T+1)}$) 需要返回对应的回复语句 $u^{(T+1)}$ (也称消息或响应), 并按实际场景来判断是否选择说话对象 $a_{\text{addressee}}^{(T+1)}$. 本文通过调研, 将首先介绍满足该定义的多方对话模型, 这是实现多方对话系统的基本要求. 然后在此基础上, 从场景元素 $a_{\text{addressee}}$ 和 u 出发, 展开讨论说话对象识别和多方对话上下文的利用等关键问题.

2.1 基于深度学习的多方对话模型

人机双方对话中按照输入是否考虑历史对话信息可以分为单轮对话和多轮对话, 而多方对话一定是多人交互的多轮对话. 在人机多轮对话中, 相关研究大体可以分为两类基本模型: 序列到序列模

型^[15]和层次序列到序列模型等. 虽然多方对话的情况更加复杂, 但是人机多轮对话中的经典模型仍然可以借鉴.

事实上, 先前研究中的部分人机对话数据集本质上是多方对话数据. 著名的 OpenSubtitles 数据集^[16]是由 XML 格式的电影对话组成的, 主要为电影中多个角色之间的对话. 但由于该数据集缺少说话人标签, 因此相关文献中把连续的话语看作两个人的轮流对话^[17], 并使用序列到序列的模型进行多轮回复生成. 而经典的 Ubuntu 数据集^[18]则来源于多人在线聊天室, 通过显式的说话对象标签和一定的规则提取双人对话数据, 该数据集被广泛应用于多轮回复选择研究^[19].

在多方对话场景中, 当对某个参与者回复时, 可通过规则提取两个人的对话历史, 进而使用已有的人机双方对话模型建模. 这种方法的可行之处在于合理地提取对话历史, 这个问题将在后文多方对话上下文中进行详细说明. 除此以外, 多方对话具有多人多轮交互的特点, 而人机多轮对话的序列到序列模型并没有充分发挥这种图结构特点.

Hu 等^[10]将已有的基于序列的对话模型推广到基于图结构的神经网络模型 (graph-structured neural network, GSN) 进行多方对话回复生成. GSN 的核心结构是一个话语图编码器, 根据图中对话信息的流动进行话语编码. 具体来说, GSN 将对话历史中的话语作为图顶点, 根据对话中的回复关系构建边. 假设存在对话 $\{(p_1, u_1), (p_2, u_2), (p_3, u_3), (p_4, u_4)\}$ (其中 p_i 代表第 i 个说话人, u_i 代表其说的话), u_3 和 u_4 都是对 u_2 的回复, 因此分别建立 u_2 到 u_3 和 u_4 的有向边. 在话语图更新时, 作者还采用了双向信息流动和说话人信息流动机制 (通过在相同说话人的话语之间建立隐式边实现). 同时, 文中提出了一种基于门控循环单元 (gated recurrent unit, GRU) 的图更新算法, 比图卷积神经网络 (graph convolutional network, GCN) 能更适应对话中的图结构建模. 通过实验, 作者发现 GSN 在多方对话和双方对话回复生成中均能比基于序列的模型取得更好的效果. 如表 1 所示, GSN 在 Ubuntu Dialogue Corpus^[18]上进行回复生成任务的各指标上都取得了较好的结果, 但由于 GSN 需要根据回复关系进行边的构建, 因此其只适用于有显式的说话对象标签的数据集, 在缺少说话对象标签的情况下, GSN 将退化成序列结构.

2.2 说话对象

在现实世界中, 一个人的说话风格不仅与其自身有关, 而且也随说话对象的变化而变化^[20,21]. 在人机双方对话中, Li 等^[22]提出二元说话者模型, 将说话者和说话对象信息表现合并, 看作交流模型的向量表示, 将该表示加入模型后有助于提高生成效果. 与双方对话相比, 多方对话场景中说话对象更加重要, 显式指明说话对象, 可便于用户理解对话的进行, 能够减少交流的难度和歧义性. 对于多方对话系统而言, 合理地选择说话对象并面向说话对象进行个性化的回复是非常重要的.

Ouchi 等^[7]首次形式化定义了多方对话中说话对象和回复选择任务 (addressee and response selection, ARS), 即给定对话历史 C 和当前说话人 a_{res} , 系统需要从对话历史说话人集合中选择合适的说话对象, 并从候选回复集合中选择正确的回复. 作者还公开了基于 Ubuntu IRC Logs 的 ARS 任务基准数据集, 并给出了基准模型 DYNAMIC-RNN. 文献 [7] 首先按照说话人最后一次发言的倒序组织说话人向量, 以表征对话时序性对说话对象选择的影响; 然后更新说话人向量, DYNAMIC-RNN 通过统一的 GRU 编码器使用历史对话更新说话人向量表示, 进而通过最大池化说话人向量获得对话上下文表示, 从而分别选择说话对象和回复. 如表 1^[7,8,19,23,24]所示, 文献 [7] 使用说话对象 - 回复选择正确率 (ADR-RES)、说话对象选择正确率 (ADR) 以及回复选择正确率 (RES) 进行结果评价. 相比于不进行说话人向量动态更新的 Static-RNN 模型, DYNAMIC-RNN 有明显的效果提升. 除此以外, DYNAMIC-RNN 仅对当前说话人进行建模, 因此能广泛适用于各种多方对话场景. 考虑到 Ubuntu

表 1 说话对象和回复选择任务的各模型性能对比 ^{a)}

Table 1 Comparison of models' performance on the task of addressee and response selection

Model	ADR-RES	ADR	RES
Static-RNN [7]	48.67	60.97	77.75
Static-Hier-RNN [19, 23]	51.76	64.61	78.28
DYNAMIC-RNN [7]	53.85	66.94	78.16
SI-RNN [8]	67.30	80.47	80.91
WGAN* [24]	54.14	70.07	75.63

a) The first four models are tested on the Ubuntu ARS dataset. Here we show the comparison of the results, where the maximum turn of dialogue is set to 10 and the number of candidates is set to 2 (We skip the results under other settings, which are roughly the same). Specially, WGAN* uses an additional multi-lingual log corpus based on Ubuntu ARS dataset.

IRC 存在多种语言的交流频道, Sato 等 [24] 将 Ouchi 等的研究推广到多语言中. 作者首先使用语言自动识别工具从 Ubuntu IRC Logs 中获取多种语言的 ARS 数据集, 然后使用 Wasserstein-GAN (WGAN) 同时处理多种语言, 核心编码器仍然使用 DYNAMIC-RNN, 创新点在于通过 W-GAN 平衡高资源和低资源语言问题. 虽然文献 [24] 在先前 ARS 任务数据集中加入了更复杂的多语言数据集, 但其在个别指标上超越了 DYNAMIC-RNN 模型.

Zhang 等 [8] 认为 DYNAMIC-RNN 中仅使用每轮话语更新该轮说话人向量是不合理的, 根据现实情况, 对话中所有参与者都应有获取对话信息的能力. 同时, DYNAMIC-RNN 没法学习到说话对象的信息. 因此, 作者引入了说话人交互循环神经网络 (speaker interaction RNN, SI-RNN), SI-RNN 通过角色敏感机制更新说话人向量表示. 在每一轮对话中, 参与者基于该轮对话的角色 (说话人、说话对象或旁观者) 使用不同类型 GRU (即参数不共享) 更新其说话人向量. 除此以外, 文献 [8] 中还注意到说话对象和回复是相互依赖的, 因此将 ARS 任务视为联合预测问题. 具体来说, 训练时, SI-RNN 通过条件概率进行预测, 即给定正确说话对象标签来选择回复, 反之亦然; 推理时, 使用最大联合概率进行说话对象和回复的选择. 如表 1 所示, SI-RNN 模型在各指标上都取得了远优于 DYNAMIC-RNN 的结果, 但是由于其在每轮对话中对多种角色进行分别建模, 因此数据集中需要有明确的说话对象标签, 理论上来说, 数据集中说话对象标签的比例越高, 效果会越好. 在数据集缺乏说话对象标签时, SI-RNN 会退化为 DYNAMIC-RNN 模型, 此时, 将无法学习到说话对象信息, 其性能会大大降低.

Liu 等 [11] 提出了说话人感知上下文循环编码器 - 解码器框架 (interlocutor-aware contexts into recurrent encoder-decoder frameworks, ICRED), 并在 ARS 数据集上进行回复生成的研究. ICRED 说话人交互层 (speaker interaction layer) 借鉴了 SI-RNN 中的角色敏感更新机制; 说话对象记忆层 (addressee memory layer) 则是说话对象在对话中最后一次回复话语的向量表示 M_{tgt} ; 解码层 (decoder layer) 则结合当前说话人、说话对象经 Speaker Interaction Layer 编码后的向量表示以及说话对象记忆表示 (分别记为 A_{res} , A_{tgt} 和 M_{tgt}), 之后使用注意力机制进行回复生成. 相比于前文中的 GSN 模型仅对历史对话建模并进行回复生成, ICRED 模型则在已知当前轮说话对象的前提下, 对历史对话建模并针对说话对象进行回复生成, 因此其在应用中需要进行说话对象识别的额外操作, 普适性较差. Yang 等 [25] 则综合了 SI-RNN 和 ICRED 的思路在《老友记》数据集上进行回复生成. 具体来说, 文献 [25] 在提取了《老友记》主要角色的对话后, 首先针对每个角色使用角色特定 GRU 更新说话人向量表示, 并且预测出每个角色作为说话对象的概率, 紧接着将说话对象预测概率融入上下文向量表示中生成回复, 相比于没有融入说话对象的情况, 实验效果得到明显提升.

先前研究表明说话对象标签十分重要, 然而在真实多方对话中, 说话人通常不会明确地指定说话

对象. Le 等^[26]通过对 Ubuntu IRC 数据集进行统计分析,发现大约 66% 的话语缺少显式说话对象信息,因此提出 W2W 模型来补全多方对话中缺失的说话对象标签. W2W 模型以角色敏感机制为基础,提出伪说话对象概念,即对所有说话人向量使用最大池化操作,获得的向量作为缺少的说话对象表示.文献^[26]尝试使用说话人向量来更新话语,从而实现说话人向量和话语的信息更深层次的交互融合.此外,文献^[26]还考虑到下一个说话人有很大概率为当前说话对象的后验知识,因此在说话人向量编码中使用双方机制,即反向进行 SI-RNN 中的更新,拼接两个方向的向量作为最终说话人表示.

除以上显式选择说话对象以外,相关研究分析话语的回复关系结构的任务,可以看作隐式的说话对象识别.这类问题可以描述为:给定对话历史消息和当前回复消息,预测当前消息的回复对象,即对历史消息进行是否回复的标签分类,也称为 Reply-To 关系识别.之前的研究^[27]将 Reply-To 关系识别当作句对分类问题解决,并没有考虑语句在上下文中的关系. Guo 等^[28]提出了分别基于单词级别 (WL-LSTM-RT) 和语句级别 (SL-LSTM-RT) 的两种编码器,其中 WL-LSTM-RT 将所有对话历史按单词顺序拼接,然后通过长短时记忆单元 (long short-term memory, LSTM) 编码,最后使用隐层输出向量分类; SL-LSTM-RT 则使用层次 LSTM 结构,即先使用单词级别 LSTM 对消息语句编码,然后将语句编码向量再次使用语句级别 LSTM 编码获得隐层向量输出. Zhu 等^[9]提出了基于回复结构的层次掩盖 Transformer 模型,并采用两阶段预训练机制.具体来说,首先根据历史对话中的回复关系构建话语图,并通过一定规则从话语图获得语句的 Mask 矩阵;在两阶段训练时,第 1 阶段使用 BERT 预训练模型进行单词级别的话语编码,然后使用多层 Transformer 结合 Mask 矩阵进行语句级别编码,最后对每个历史消息语句进行分类.

2.3 多方对话上下文

合理地利用对话历史上下文的能力是建立对话系统的关键,它能够使对话保持活跃且具有吸引力.在人机双方对话中, Song 等^[29]拼接所有的对话历史进行对话生成; Serban 等^[30]使用基于注意力机制的层次结构进行编码,从而捕捉话语之间的关系信息; Zhang 等^[17]提出使用静态、动态注意力机制来学习不同时刻对话历史的影响.对比来说,多方对话中上下文的建模更具有挑战性:多方对话中存在多个参与者,多人的回复使得对话交互复杂,对话主题多样.并且往往充斥大量无用信息,此时对上下文的合理利用尤为重要.

如前文所述,可以对复杂的多方对话历史进行结构分离,进而更有针对性地进行回复. Zhang 等^[31]将多方对话的历史消息话语按照“@”回复关系,组织成一个树形结构,然后将树形结构分割成多个序列组合,进而使用说话对象所在分支的话语进行回复生成.文献^[31]实验结果表明使用较少的上下文信息来排除无关分支的话语,可以产生更好的结果.如表 2^[7, 10, 11, 18, 22, 23, 30~33]中 TreeSplit 结果所示,该模型在个别指标上远优于对完整对话历史进行建模的模型,但该方法需要数据集中有显式的说话对象标签.当缺乏说话对象标签时,将只存在单一对话分支,其效果与传统模型相同.

在多方对话场景中,多个会话可能同时发生,即同一个群组或聊天室中进行多个主题/线程的对话,从而导致对话日志中的会话线程混乱,这种情况在即时聊天频道 (例如 Ubuntu IRC 和 Slack 等) 中尤为常见.为了便于对话理解、自动摘要等工作的顺利进行,会话分离是多方对话中十分重要的工作,在相关研究中也称为会话解缠 (conversation disentanglement)^[34, 35]或线程检测 (thread detection)^[36].

这类问题可以描述为:给定对话历史上下文以及每个语句对应的线程 (或主题) 标签,求解当前回复消息的线程标签.之前的相关研究大多基于句对分类的方法,并使用语法规则等特征.随着深度学习技术的发展,基于神经网络的句对分类模型被应用到该问题中^[34],并出现了一些基于神经网络学习对话上下文信息的解决方案.

表 2 多方对话回复生成任务上模型性能对比^{a)}
Table 2 Comparison of models' performance on the multi-party dialogue response generation task

Dataset	Model	BLEU*/BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE*/ROUGE _L
ARS Dataset [7]	Seq2Seq [32]	8.86*	—	—	—	—	7.62*
	Persona Model [22]	9.12*	—	—	—	—	7.38*
	VHRED [30]	9.38*	—	—	—	—	7.65*
	ICRED [11]	10.63*	—	—	—	—	8.73*
Ubuntu Dialogue Corpus [18]	Context-Seq2Seq [33]	10.45	4.13	2.08	1.02	3.43	—
	TreeSplit [31]	11.73	6.06	4.28	3.29	4.86	—
	HRED [23]	11.23	4.60	2.54	1.42	4.38	10.23
	GSN [10]	13.50	5.63	3.24	1.99	4.85	11.36

a) Both the ARS Dataset and the Ubuntu Dialogue Corpus are from Ubuntu IRC Logs, the differences of which are mainly in data size, data pre-processing, and dataset splitting. All models on the ARS Dataset use current addressee information as a supervision signal. Both the TreeSplit and GSN models reply relations during a dialogue on the Ubuntu Dialogue Corpus.

Tan 等 [36] 提出了上下文感知线程检测 (context-aware thread detection, CATD) 模型, 该模型由 CATD-FLOW 和 CATD-MATCH 两部分组成. CATD-FLOW 使用 LSTM 对每个线程内的语句进行编码, 并将 LSTM 最后隐层输出向量和当前消息的向量表示拼接; CATD-MATCH 则从语义匹配的角度出发, 首先对会话线程和输入消息使用同一个 LSTM 编码, 从而获得更多语义层面的相似性, 紧接着使用注意力机制, 计算输入消息向量表示和会话线程上下文向量表示之间的相似性, 进而得到匹配向量表示. 文献 [36] 将 CATD-FLOW 和 CATD-MATCH 的最后向量表示以一定权重相加, 经过分类器得到输入消息属于该线程的概率.

Gu [35] 提出 Dialogue BERT (DialBERT) 模型, 并在 Kummerfeld 等 [27] 提出的数据基础上进行实验. 文献 [35] 将候选线程中最近的多个消息以及当前输入消息按照单词进行拼接, 消息之间用 SEQ 作为标识分隔, 然后输入到 DialBERT 进行编码; 通过 BERT 强大的特征学习能力, 可以得到每个历史消息和当前输入深度交互融合的向量表示; 紧接着使用双向 LSTM 对历史消息向量表示进行语句级别编码, 并将 LSTM 最后隐层向量使用分类器进行分类.

3 多方对话相关任务

随着互联网的发展, 在线聊天平台产生了大量的对话数据, 网络论坛也积累了海量的近似对话数据. 这些数据除了可以应用于对话系统的研究以外, 在用户友好性和社交媒体分析等领域中也存在许多研究和应用. 在深度学习兴起之前的研究中, 有许多基于规则或机器学习的方法, 来研究多方对话聊天界面的用户友好性, 如对消息历史进行主题分类并高亮 [37], 其实这与前文的会话分离本质相同. 得益于深度学习技术, 之前的研究得到了飞跃式的发展, 并且近年来也出现了许多更为高级的基于多方对话的研究, 比如对话阅读理解、对话情感识别等. 这些研究任务是基于多方对话进行的, 但尚未直接应用于多方对话系统的建模中, 因此本文将这类研究统称为多方对话相关任务.

3.1 多方对话阅读理解

机器阅读理解 (machine reading comprehension, MRC) 是让机器学会阅读和理解文章, 即对于给定的问题, 从相关文章中寻找答案. 近年来, 会话机器理解 (conversation machine comprehension,

CMC)^[38] 逐渐获得关注, 这类任务中给定开放领域的文本, 然后机器需要进行多轮对话, 回答与文本相关的问题. 本文将介绍的多方对话阅读理解与之不同, 是指从多方对话历史中寻找问题的答案. 由于多方对话历史大多是口语化表达, 由不同的说话人组成, 语句间的风格和主题可能有较大差异, 因此多方对话阅读理解的难度更大. 多方对话阅读理解的研究可以应用于信息搜索、自动问答等领域, 对于海量多方对话数据的分析和利用有一定的促进作用.

Sun 等^[12] 在英语听力阅读理解数据集 DREAM 中使用 Gated-Attention Reader 等多种经典多项选择阅读理解模型, 但模型大多仅学习到表层信息, 实验效果略强于基于规则的方法. 文献 [12] 分析效果差的原因除了多轮多方参与以外, 还在于 85% 问题需要多句推理, 34% 的问题需要常识知识, 总体来说更具有挑战性.

Li 等^[39] 使用基于 Transformer 结构的预训练模型来学习多方对话的层次上下文信息, 并在 Yang 等^[38] 提出的多方对话阅读理解数据集 FriendsQA 上取得了较好的实验效果. Li 等^[40] 则分析了基于多方对话的填空式阅读理解的挑战, 使用一种新的数据分割方法提升填空式阅读理解任务的难度, 进而重新评价先前的研究工作的效果, 结果表明多方对话阅读理解还有很大的进步空间.

3.2 多方对话情感分析

情感分析是自然语言处理领域的基础任务, 属于文本分析范畴. 而人机对话中的情感分析^[41] 则是对话系统较高层次的能力, 情感分析使得对话系统具有情绪感知能力, 进而能够增强系统的共情能力, 提高用户对系统的信任程度和参与度.

Chen 等^[13] 发布了多方对话情感检测数据集 EmotionLines, 并使用卷积神经网络 (convolutional neural network, CNN) 和双向 LSTM 分别编码句子级别和上下文级别的向量表示. 分类实验结果表明, 相对于传统的单文本情感识别, 多方对话的情境信息能够提高情感识别正确率.

Ghosal 等^[42] 提出一种基于图神经网络的对话图卷积网络模型 (DialogueGCN), DialogueGCN 利用说话人之间的相互依赖关系来建模用于情感识别的对话上下文. 具体来说, 将话语作为图节点, 按照对话时序关系在话语之间构建时序信息边, 并在相同说话人的前后两个话语之间建立说话人信息边, 然后使用 GCN 进行节点更新. 作者在多模态情感分析数据集 MLED^[43] 上进行实验, 实验结果表明该模型提高了多种类型情感的识别准确率.

3.3 多方对话说话人建模

人的个性化偏好、说话风格等人格化特征往往对其表达的话语有一定的影响. 相关研究已经表明, 对说话人进行建模有利于提高对话系统的回复效果^[8, 22]. 多方对话中, 对于不同说话人进行建模意义尤为重大.

Meng 等^[44] 在 2018 年提出使用说话人分类进行说话人建模的替代任务, 即给定候选说话人集合和一组话语, 该任务需要预测这组话语的正确说话人标签. 作者使用循环神经网络, 融合了时序特征和上下文特征进行预测. 其中, 时序特征遵循 Ouchi 等^[7] 提出的 DYNAMIC-RNN 模型的做法, 即对说话人按最后一次发言进行倒序排列, 并维护说话人向量; 上下文特征则使用层次循环神经网络对每个说话人最近的话语进行编码, 其向量表示作为该说话人的表示.

Jiang 等^[45] 提出了基于预训练的上下文表示和注意力神经网络来进行人格识别任务. 该任务可以描述为: 给定固定上下文长度的对话场景, 该场景中存在一个主要的说话人, 以及对应的人格标签. 其做法为拼接上下文作为输入, 然后使用预训练模型 RoBERTa 进行编码表示, 之后进行说话人的人格分类.

除此以外, Ma 等^[46]使用层次卷积神经网络在《老友记》对话数据集上进行说话人识别研究。

3.4 多方对话语篇解析

多方对话语篇解析 (discourse parsing) 旨在分析多方对话历史中话语的语篇结构和语义关系, 包括评论、致谢及问答等多种关系类型。

Afantenos 等^[47]定义了多方对话语篇分析任务并发布了 STAC 数据集, 但 Afantenos 等和之后 Perret 等^[48]主要使用基于手工特征和规则的统计学习方法 (最大熵和最大生成树) 解决。Shi 等^[49]提出基于深度序列模型的多方对话语篇解析模型, 模型采用深度神经网络自动学习特征, 并且在确定父节点的同时学习到包含路径信息的结构化表示。

Li 等^[50]发布了首个可以同时应用于机器阅读理解和语篇解析两个任务的多方对话数据集, 在该数据集上使用 GCN 学习语篇结构信息, 并加入到阅读理解模型的隐层中。实验效果表明语篇结构信息能够提升阅读理解的效果。Zhang 等^[51]将传统的问答关系 (question answering, QA) 扩展为更丰富的语篇关系类型, 并在多方对话数据上进行实验, 结果表明更丰富的问答关系类型可以提高问答抽取的效果。

4 多方对话数据集

4.1 多方对话数据集总结

对话数据集的可用性在很大程度上促进了用于人机对话的神经网络模型的发展。理论上来说, 多方交互场景中的交流文本等近似于对话的数据, 都可以作为多方对话数据集^[52]。本文通过调研, 对已有研究中使用的数据集进行归纳、梳理, 结果如表 3^[7, 9, 12, 13, 18, 24, 27, 38, 43~45, 47, 50, 51, 53~56]所示。根据来源类型, 可以将多方对话数据分为以下 4 类。

即时通讯和在线聊天频道。该类别主要是真实的人人对话。目前流行的即时通讯软件包括微信和 Telegram 等, 但受到隐私等政策保护, 一般无法获取到群聊对话记录。Guo 等^[28]通过导出私人的微信群聊对话进行实验, 但并未公布数据集。在线聊天频道, 如 Ubuntu IRC Logs, Slack 等, 是目前多方对话数据集的主要来源之一, 尤其是包含各种 Ubuntu 技术讨论的 Ubuntu IRC Logs, 聊天记录完全公开¹⁾、合法。因此, 出现了各种基于 Ubuntu IRC Logs 应用于各种多方对话任务的数据集。Uthus 等^[53]于 2013 年发布大规模多方对话数据集 Ubuntu Chat Corpus, 并讨论了包括会话分离、自动对话摘要和多语言技术等各种可能研究应用场景。Lowe 等^[18]在 2015 年发布了用于多轮对话研究的数据集 Ubuntu Dialogue Corpus, 虽然该数据集通过一系列的预处理从多方对话日志中抽取双方参与的对话数据, 并被广泛应用于对话的回复生成和选择研究中, 但其处理之前的源数据²⁾仍较多地应用于多方对话研究 (这里我们统计该数据集的源数据规模)。Ouchi 等^[7]在 2016 年发布了用于说话对象和回复选择的数据集 ARS Dataset, 包括 ARS 任务的数据处理脚本等。Kummerfeld 等^[27]在 2019 年提出了大规模、高质量具有多种语篇关系标注的数据集 IRC Disentanglement Dataset。Li 等^[50]于 2020 年基于 Ubuntu Chat Corpus 发布了带有丰富语篇关系标注和多个基于对话的问答数据集 Molweni, 该数据集可以应用于语篇结构分析和阅读理解等研究问题。

社交平台和网络论坛。随着社会媒体的发展, 社交平台和网络论坛等产生了大量多人交流数据。这种发帖 - 回复交流形式近似于对话, 因此被广泛应用于各种多方对话研究。常见的社交平台包括新

1) <https://irclogs.ubuntu.com/>.

2) <http://dataset.cs.mcgill.ca/ubuntu-corpus-1.0/>.

浪微博、Twitter 等. 新浪微博数据是具有代表性的短文本数据, 在文本情感识别和个性化回复生成上有较多应用, 但目前尚未有公开的多方对话数据集. 来自 Twitter 的数据大多应用于社交分析和回复生成^[22]等, Sordoni 等^[33]从 Twitter FireHose 获取对话数据, 但其构造为包含上下文、消息和回复的三元组, 破坏了原有的多方对话数据形式. 之后的研究也大多使用 Twitter 的三元组形式数据, 缺少多方对话相关的数据集. 目前流行的网络论坛类平台包括百度贴吧、豆瓣以及 Reddit 论坛等, 其中豆瓣和 Reddit 论坛数据在人机对话研究中使用较多. Wu 等^[57]提出的基于豆瓣小组的两人多轮对话数据在回复生成和检索中有广泛使用; 在多方对话方面, Guo 等^[28]和 Lv 等^[58]通过网络爬虫获取了豆瓣小组对话数据, 但均未公开相关数据集. 相对而言, Reddit 论坛数据³⁾更容易获取, 在多方对话研究中使用较多. Jiang 等^[34]对 Reddit 数据进行标注, 使其可以应用于会话分离任务. Zhang 等^[51]对 Reddit 数据标注话语行为和关系, 在此基础上, Zhu 等^[9]扩增了数据集规模, 但仅自动标注了回复关系.

影视剧和电视节目. 随着社会的发展, 人们对精神文化的需求提高, 出现了大量文学影视作品以及电视节目, 其中有不少近似于多方对话数据. Tiedemann 等^[16]发布 OpenSubtitles 数据集, 该数据集包含各种电影对白, 因缺少说话人标签而被看作双方参与的轮流对话, 较少应用于多方对话研究. 电视剧《老友记》是在多方对话研究中应用最多的数据集之一, 其场景对话数据集可从网络获取⁴⁾, 不少研究者对该数据集进行标注, 并应用于语篇解析、情感识别等研究. Yang 等^[38]对《老友记》对话进行标注, 包括多个问答对以及问题类型等信息, 发布了问答数据集 FriendsQA. Chen 等^[55]于 2016 年对《老友记》对话中所有人称、实体等进行角色指代标注, 并公开了数据集. Chen 等^[13]在 2019 年对《老友记》对话进行情感标注, 发布了 EmotionLines 数据集. Poria 等^[43]将 EmotionLines 拓展为了多模态多方对话情感识别数据集 MLED, 其中每个话语都带有情感标签, 并包含听觉、视觉和文本共 3 个模态的信息. Jiang 等^[45]创建了第 1 个基于多方对话的个性化数据集 FriendsPersona, 作者通过众包的方式给《老友记》中每个子场景的核心说话人标注了 5 个个性化特征, 进行人格识别研究. Chen 等^[56]在 2020 年提出了首个带有情感和关系标签的中文多方对话数据集 MPDD, 该数据集来源于电视剧本.

其他来源. 多方对话数据除以上来源外, 不少研究者也根据应用需求, 从其他渠道获取各种多方对话数据集. Sun 等^[12]发布了基于中国英语等级考试的听力阅读理解的数据集 DREAM, 是第 1 个基于多方对话的多项选择阅读理解数据集. Afantenos 等^[47]从在线游戏 The Settlers of Catan 中获取聊天数据, 并针对语篇结构分析任务进行标注, 发布了数据集 STAC. Zeng 等^[59]从美国最高法院庭审记录获得多方辩论的对话数据⁵⁾, 进行辩论过程中主题变化的研究.

4.2 多方对话数据集的分析

已有多方对话数据集的局限性对于深度学习模型来讲, 训练数据的不断增加能够带来性能的显著提升, 比如最近的 Blender^[60]和 PLATO^[61]等. 但如表 3 所示, 目前较多的多方对话数据集的规模较小^[45,54]、领域性较强^[47]. 并且大多数对话数据, 可能由于一些法律及道德原因无法公开^[28]. 而对于可以公开使用的大规模数据集, 比如 Ubuntu 和 Twitter 等数据, 大多从社交网络中获取, 不可避免地存在大量错误和噪音, 同时还存在一些局限性, 如表 4 所示. Ubuntu 数据集主要是关于 Ubuntu 相关技术问题的讨论, 其中涉及到较多的专业名词, 领域性较强, 目前在学术研究中应用较多^[7,18,24,27],

3) <https://files.pushshift.io/reddit/>.

4) <https://fangj.github.io/friends/>.

5) https://www.supremecourt.gov/oral_arguments/oral_arguments.aspx.

表 3 多方对话研究相关数据集
Table 3 Statistics of multi-party dialogue datasets

Type	Name	Total of dialogs/turns	Task	Source	Description
Online chatroom	Ubuntu Chat Corpus ^[53]	10.6k/34 M	–	Ubuntu IRC Logs	–
	Ubuntu Dialogue Corpus ^[18]	380k/1.75 M	Response generation	Ubuntu IRC Logs	–
	ARS Dataset ^[7]	7.3k/2.4 M	Addressee and response selection	Ubuntu IRC Logs	–
	multilingual ASR Corpus ^[24]	0.83 M/–	Addressee and response selection	Ubuntu IRC Logs	Multilingual
	IRC Disentanglement Dataset ^[27]	–/–	Conversation disentanglement	Ubuntu IRC Logs	–
	Molweni ^[50]	10k/88k	Discourse parsing and machine reading comprehension	Ubuntu Chat Corpus	–
	MPC Corpus ^[54]	14/520	Social phenomena analysis	Online conversation	–
	Reddit-Small ^[51]	9.3k/0.11 M	Discourse act classification	Reddit	–
	Reddit-Large ^[9]	76k/3.4 M	Conversation structure modeling	Reddit	–
	UseNet Corpus ^{a)}	47.8k/32.9 M	–	UseNet	–
Microblog and forum	Character Identification ^[55]	–/–	Character identification	Friends	–
	EmotionLines ^[13]	1k/14k	Emotion recognition	Friends	–
	MLED ^[43]	1.4k/13k	Emotion recognition	Friends	Multi-modal
	FriendsQA ^[38]	1.2k/–	Machine reading comprehension	Friends	–
	FriendsPersona ^[45]	0.7k/–	Personality recognition	Friends	–
	MPDD ^[56]	4k/25.5k	Emotion recognition	TV show scripts	Chinese
	Speaker Modeling Corpus ^[44]	–/–	Speaker identification	TV talk shows	–
	SATC ^[47]	2.1k/–	Discourse parsing	The Settlers of Catan	–
	DREAM ^[12]	6.4k/30.2k	Machine reading comprehension	English language exams	–

a) <http://www.psych.ualberta.ca/~westbury/lab/downloads/usenetcorpus.download.html>.

表 4 多方对话主要数据集对比^{a)}
Table 4 Comparison of main datasets

Name	Open-domain	Spontaneous	Spoken	Explicit addressee	Non-scripted	Chinese
Ubuntu IRC Logs	×	✓	×	✓	✓	×
Reddit	✓	×	×	✓	✓	×
Friends	✓	✓	✓	×	×	×
Twitter	✓	×	×	✓	×	×

a) (1) Open-domain dialogues are different from special domains or topic dialogues. For instance, a Ubuntu Dataset mainly refers to a technology discussion about Ubuntu. (2) The topics of spontaneous conversations are either casual or not pre-specified which closely mimic spontaneous and unplanned spoken interactions between humans. (3) Spoken dialogues tend to be more colloquial and generally well-formed as the user speaks in a train-of-thought manner where the speakers are face-to-face. (4) Explicit Addressee means that there is an explicit signal such as “@” that indicates a listener for speaking. (5) Non-scripted dialogues are different from scripted dialogues that are required to be dramatic, as latter are generally sourced from movies or TV shows.

但很难应用于开放域对话研究. Twitter 数据集来自于社交网络, 由于平台输入限制, 其中对话回复较短, 包含很多缩略语以及拼写错误等. 《老友记》数据集 (以及其他影视剧作品) 来自于剧本对话, 其对话和口语十分接近, 且大多缺少说话人标签. 除此以外, 来自社交网络和影视剧本的对话数据, 大多受对话之外的未知因素影响, 比如近期热门事件和剧情画面等, 这些都会加大模型的学习训练难度.

多方对话数据集的构建和完善. 如表 4 所示, 目前缺乏用于学术研究的中文多方对话数据集, 相关研究有从新浪微博^[62] 或者豆瓣小组^[57] 获取中文对话数据, 但公布的处理后数据并不适用于多方对话研究. 为了尽快构建中文多方对话数据集, 从新浪微博、豆瓣小组或者百度贴吧等中文论坛或者社交平台获取类似多方对话数据作为替代, 同时我们也做了一些尝试⁶⁾, 证明了获取的多方对话数据具有一定的可用性. 如前文所述, 从网络论坛或者社交平台获取的数据存在错误和噪音, 需要进行更加细致的数据清洗, 此时可以结合已有的高质量数据集 (或者人工标注一部分) 训练深度分类模型来辅助数据过滤. 先前的研究表明^[8,10], 说话对象标签是多方对话特有的, 并且在相关模型中应用能够显著提升效果, 但是在已有的对话数据集中, 说话对象标签甚至说话人标签大多是缺失的^[26]. 虽然目前有许多工作尝试了人工标注^[13,38,43,45,50,56], 但是限于人力和财力, 人工标注的数据量较少. 在这方面, 数据自动标注是可探索的方向, 比如影视及电视节目大多缺少说话人标签, 可以使用说话人识别等技术辅助标注^[63]. 而对于多方对话中缺少显式的说话对象标签问题, 可以使用 2.2 小节提到的说话对象识别算法^[26] 进行自动标注. 另外, 考虑到 Twitter 等社交平台的对话大多受近期热门事件和新闻的影响, 可以考虑将外部事件作为背景知识加入到对话历史中; 而对于影视剧, 还可以加入多模态信息完善数据集.

5 多方对话未来发展

随着深度学习技术在对话领域的广泛应用, 多方对话研究已经取得了一些进展, 但仍有许多研究方向值得深入探索. 根据已有研究, 本文总结了以下 3 个可能的研究方向.

(1) 多方对话结构信息的利用. 多方对话具有多人多轮交互的特点, 交互过程形成天然的图结构, 而传统基于序列的模型并没有充分利用这种结构信息. 最近研究者们提出了多种基于图结构的对话建模方法^[10], 但实际效果还没有得到广泛的验证.

6) <https://github.com/iseesaw/SMP-MCC2020>.

(2) 多模态多方对话的研究. 目前人机对话大多基于文本数据进行, 如何结合视听觉信息构建对话系统, 是值得探索的问题. 在人机双方对话中已经存在各种实体机器人的应用研究^[64], 一定程度上能够融合多模态信息, 从而达到深层次理解的目的. 现实场景中多模态多方对话的应用会更为广泛, 但目前的系统大多依赖于预设的规则进行对话控制.

(3) 多方对话的回复时机研究. 当前的对话系统以被动的回应为主, 而在真实场景下, 对话双方都应把握回复的时机, 比如是否等待对方说完再进行回复. 在多方对话中, 这个问题显得尤为重要, 不恰当的回复时机将显著降低多方对话质量^[65]. 但是, 传统的多方对话系统主要研究对话轮转问题 (turn-taking)^[66], 大多从整个交互系统层面而非对话系统自身来考虑.

6 结束语

目前多方对话已经取得了一定的关注, 但仍然缺少成熟的多方对话系统的应用, 很大程度上受限于当前研究的水平. 本文主要围绕基于深度学习技术的多方对话研究进行介绍, 分类梳理已有的研究成果, 整理了多方对话系统的说话对象和上下文利用两个关键问题及已有的解决方案, 总结了基于多方对话的相关任务、已有解决办法和已有的多方对话数据集, 希望对相关研究人员提供些许参考.

参考文献

- 1 Zhu Q, Cui L, Zhang W N, et al. Retrieval-enhanced adversarial training for neural response generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, 2019. 3763–3773
- 2 Seering J, Luria M, Kaufman G, et al. Beyond dyadic interactions: considering chatbots as community members. In: Proceedings of CHI Conference on Human Factors in Computing Systems, Glasgow, 2019. 1–13
- 3 Uthus D C, Aha D W. Multiparticipant chat analysis: a survey. *Artif Intell*, 2013, 199–200: 106–121
- 4 de Bayser M G, Cavalin P, Souza R, et al. A hybrid architecture for multi-party conversational systems. 2017. ArXiv:1705.01214
- 5 Kennington C, Funakoshi K, Yakahashi Y, et al. Probabilistic multiparty dialogue management for a game master robot. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Bielefeld, 2014. 200–201
- 6 Żarkowski M. Multi-party turn-taking in repeated human-robot interactions: an interdisciplinary evaluation. *Int J Soc Robot*, 2019, 11: 693–707
- 7 Ouchi H, Tsuboi Y. Addressee and response selection for multi-party conversation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, 2016. 2133–2143
- 8 Zhang R, Lee H, Polymenakos L, et al. Addressee and response selection in multi-party conversations with speaker interaction RNNs. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 5690–5697
- 9 Zhu H, Nan F, Wang Z, et al. Who did they respond to? Conversation structure modeling using masked hierarchical transformer. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI), New York, 2018. 4951–4958
- 10 Hu W, Chan Z, Liu B, et al. GSN: a graph-structured network for multi-party dialogues. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, 2019. 5010–5015
- 11 Liu C, Liu K, He S, et al. Incorporating interlocutor-aware context into response generation on multi-party chatbots. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, 2019. 718–727
- 12 Sun K, Yu D, Chen J, et al. DREAM: a challenge data set and models for dialogue-based reading comprehension. *Trans Assoc Comput Linguist*, 2019, 7: 217–231
- 13 Chen S, Hsu C C, Kuo C C, et al. Emotionlines: an emotion corpus of multi-party conversations. 2018. ArXiv:1802.08379

- 14 Traum D. Issues in multiparty dialogues. In: *Advances in Agent Communication*, 2004. 2922: 201–211
- 15 Zhu Q, Zhang W, Zhou L, et al. Learning to start for sequence to sequence architecture. 2016. ArXiv:1608.05554
- 16 Tiedemann J. Parallel data, tools and interfaces in OPUS. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, 2012. 2214–2218
- 17 Zhang W N, Cui Y, Wang Y, et al. Context-sensitive generation of open-domain conversational responses. In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, 2018. 2437–2447
- 18 Lowe R, Pow N, Serban I V, et al. The Ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, 2015. 285–294
- 19 Zhou X, Dong D, Wu H, et al. Multi-view response selection for human-computer conversation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, 2016. 372–381
- 20 Zhang W N, Zhu Q, Wang Y, et al. Neural personalized response generation as domain adaptation. *World Wide Web*, 2019, 22: 1427–1446
- 21 Song H, Zhang W N, Hu J, et al. Generating persona consistent dialogues by exploiting natural language inference. 2019. ArXiv:1911.05889
- 22 Li J, Galley M, Brockett C, et al. A persona-based neural conversation model. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, 2016. 994–1003
- 23 Serban I V, Sordoni A, Bengio Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models. In: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, 2016. 3776–3783
- 24 Sato M, Ouch H, Tsuboi Y. Addressee and response selection for multilingual conversation. In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, 2018. 3631–3644
- 25 Yang Q, He Z, Zhan Z, et al. End-to-end personalized humorous response generation in untrimmed multi-role dialogue system. *IEEE Access*, 2019, 7: 94059–94071
- 26 Le R, Hu W, Shang M, et al. Who is speaking to whom? Learning to identify utterance addressee in multi-party conversations. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 2019. 1909–1919
- 27 Kummerfeld J K, Gouravajhala S R, Peper J, et al. A large-scale corpus for conversation disentanglement. 2018. ArXiv:1810.11118
- 28 Guo G, Wang C, Chen J, et al. Who is answering to whom? Finding “reply-to” relations in group chats with long short-term memory networks. In: *Proceedings of the 7th International Conference on Emerging Databases*, 2018. 461: 161–171
- 29 Song H, Wang Y, Zhang W N, et al. Generate, delete and rewrite: a three-stage framework for improving persona consistency of dialogue generation. In: *Proceedings of Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation*, Online, 2020. 5821–5831
- 30 Serban I V, Sordoni A, Lowe R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, 2017. 3295–3301
- 31 Zhang H, Chan Z, Song Y, et al. When less is more: using less context information to generate better utterances in group conversations. In: *Proceedings of the 7th CCF International Conference on Natural Language Processing and Chinese Computing*, Hohhot, 2018. 76–84
- 32 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: *Proceedings of the 27th Advances in Neural Information Processing systems*, Montreal, 2014. 3104–3112
- 33 Sordoni A, Galley M, Auli M, et al. A neural network approach to context-sensitive generation of conversational responses. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, 2015. 196–205
- 34 Jiang J, Chen F, Chen Y, et al. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In: *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, 2018. 1812–1822
- 35 Gu J. DialBERT: a hierarchical pre-trained model for conversation disentanglement. 2020. ArXiv:2004.03760
- 36 Tan M, Wang D, Wang H. Context-aware conversation thread detection in multi-party chat. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

- Natural Language Processing, Hong Kong, 2019. 6457–6462
- 37 Uthus D C, Aha D W. Extending word highlighting in multiparticipant chat. In: Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference, Florida, 2013. 238–242
 - 38 Yang Z, Choi J D. FriendsQA: open-domain question answering on TV show transcripts. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, Stockholm, 2019. 188–197
 - 39 Li C, Choi J D. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. 2020. ArXiv:2004.03561v1
 - 40 Li C, Liu T, Choi J. Design and challenges of cloze-style reading comprehension tasks on multiparty dialogue. 2019. ArXiv:1911.00773
 - 41 Poria S, Majumder N, Mihalcea R, et al. Emotion recognition in conversation: research challenges, datasets, and recent advances. *IEEE Access*, 2019, 7: 100943
 - 42 Ghosal D, Majumder N, Poria S, et al. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019. 154–164
 - 43 Poria S, Hazarika D, Majumder N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations. 2019. ArXiv:1810.02508
 - 44 Meng Z, Mou L, Jin Z. Towards neural speaker modeling in multi-party conversation: the task, dataset, and models. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 8121–8122
 - 45 Jiang H, Zhang X, Choi J D. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. 2019. ArXiv:1911.09304
 - 46 Ma K, Xiao C, Choi J D. Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 49–55
 - 47 Afantenos S, Kow E, Asher N, et al. Discourse parsing for multi-party chat dialogues. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 928–937
 - 48 Perret J, Afantenos S, Asher N, et al. Integer linear programming for discourse parsing. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, 2016. 99–109
 - 49 Shi Z, Huang M. A deep sequential model for discourse parsing on multi-party dialogues. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, 2019. 7007–7014
 - 50 Li J, Liu M, Kan M Y, et al. Molweni: a challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. 2020. ArXiv:2004.05080
 - 51 Zhang A, Culbertson B, Paritosh P. Characterizing online discussion using coarse discourse sequences. In: Proceedings of the 11th International Conference on Web and Social Media, Quebec, 2017. 357–366
 - 52 Serban I V, Lowe R, Henderson P, et al. A survey of available corpora for building data-driven dialogue systems: the journal version. *Dialogue & Discourse*, 2018, 9: 1–49
 - 53 Uthus D C, Aha D W. The Ubuntu chat corpus for multiparticipant chat analysis. In: Proceedings of AAAI Spring Symposium on Analyzing Microtext, 2013. 99–102
 - 54 Shaikh S, Strzalkowski T, Broadwell A, et al. MPC: a multi-party chat corpus for modeling social phenomena in discourse. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, 2010. 2007–2013
 - 55 Chen Y H, Choi J D. Character identification on multiparty conversation: identifying mentions of characters in TV shows. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, 2016. 90–100
 - 56 Chen Y, Huang H, Chen H. MPDD: a multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, 2020. 610–614
 - 57 Wu Y, Wu W, Xing C, et al. Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics, Vancouver, 2017. 496–505
 - 58 Lv Z, Xu J, Zhao P, et al. Learning the structures of online asynchronous conversations. In: Proceedings of the 22nd

- International Conference on Database Systems for Advanced Applications, Suzhou, 2017. 19–34
- 59 Zeng J, Li J, He Y, et al. What changed your mind: the roles of dynamic topics and discourse in argumentation process. In: Proceedings of the Web Conference, Taipei, 2020. 1502–1513
 - 60 Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. 2020. ArXiv:2004.13637
 - 61 Bao S, He H, Wang F, et al. PLATO-2: towards building an open-domain chatbot via curriculum learning. 2020. ArXiv:2006.16779
 - 62 Shang L, Lu Z, Li H. Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 1577–1586
 - 63 Kuzu R S, Balci K, Salah A A. Authorship recognition in a multiparty chat scenario. In: Proceedings of the 4th International Workshop on Biometrics and Forensics, Limassol, 2016. 1–8
 - 64 Zhang W, Liu T, Qin B, et al. Benben: a Chinese intelligent conversational robot. In: Proceedings of ACL 2017 (System Demonstrations), Vancouver, 2017. 13–18
 - 65 Lan T, Mao X, Huang H, et al. When to talk: chatbot controls the timing of talking during multi-turn open-domain dialogue generation. 2019. ArXiv:1912.09879
 - 66 Levitan R, Benus S, Gravano A, et al. Entrainment and turn-taking in human-human dialogue. In: Proceedings of 2015 AAAI Spring Symposium Series, 2015. 44–51

A survey of multi-party dialogue research based on deep learning

Kaiyan ZHANG, Wei-Nan ZHANG* & Ting LIU

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China

* Corresponding author. E-mail: wnzhang@ir.hit.edu.cn

Abstract In recent years, with the extensive application of deep learning technology, breakthroughs have been made in the study of human-computer dialogue. However, most of the current human-machine dialogue systems are designed under the assumption that both parties are involved, and the research and application of more challenging multi-party human-machine dialogues are not yet mature. Based on the field of natural language processing, this paper will review the research progress of multi-party dialogue based on deep learning in recent years. First, from the perspective of human-machine dialogue, we sort out the key problems and existing solutions of the multi-party dialogue system; then, we introduce other natural language processing tasks based on multi-party dialogue; afterwards, we summarize the existing multi-party dialogue research dataset and make a comparative analysis of limitations on the existing dataset; Finally, we look forward to the future development trend of multi-party dialogue research.

Keywords natural language processing, deep learning, human-machine dialogue, multi-party dialogue



Kaiyan ZHANG is a graduate student at Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interests include human-computer dialogue systems, multi-party dialogue, pre-training for open-domain dialogue.



Wei-Nan ZHANG an associate professor at Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interests include human-computer dialogue systems, natural language processing, and information retrieval.



Ting LIU is a professor at Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His primary research interests are in natural language processing, information retrieval, and social computing.