智能回复系统研究综述

岳世峰[1,2], 林 政[1], 王伟平[1], 孟 丹[1]

[1](中国科学院信息工程研究所 信息内容安全技术国家工程实验室 北京 中国 100093)

[2](中国科学院大学 网络空间安全学院 北京 中国 100049)

摘要 网络舆情形成快、影响大, 如何对其进行智能导控一直是网络安全中的难点和重点。本文提出使用智能回复系统对网络舆情进行自动引导的观点, 然后对智能回复系统研究进行了综述。本文首先介绍了当前智能回复系统的主流研究方向, 如视觉问答、基于知识图谱的问答和推理问答等不同类型的智能回复系统;接着根据应用场景的不同分别介绍了垂直领域和开放领域的智能回复系统, 然后从技术手段上对实现智能回复系统的各种主流方法进行了详细的介绍和探讨。最后本文总结归纳了当前智能回复系统的自动评价方法以及当前智能回复系统存在的主要问题及未来可能的研究方向。

关键词 网络舆情; 深度学习; 问答系统; 生成模型; 检索模型

中图法分类号 TP391.1 DOI 号 10.19363/J.cnki.cn10-1380/tn.2020.01.03

Research on Intelligent Reply System: A Survey

YUE Shifeng[1,2], LIN Zheng[1], WANG Weiping[1], MENG Dan[1]

[1](National Engineering Laboratory for Information Security Technologies, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

[2](School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract Online public opinion develops rapidly and has a great impact on society. It has always been difficult and important to lead it in network security. In this paper, we propose the idea of using intelligent reply system to lead the online public opinion , and then we provide a general overview of the current intelligent reply system. First, we introduce the current mainstream research methods for the intelligent reply system, including the visual question answering, knowledge-based question answering and inference question answering; second, we demonstrate the task-oriented and untask-oriented intelligent reply syst em in terms of different application scenarios. After that we discussed the mainstream methods of building reply system. Finally we summarize the automatic evaluation methods of the current reply system and the main problems as well as future research directions in this field.

Key words online public opinion; deep learning; question-answer system; generation model; retrieval mode

1 引言

随着互联网的微信、微博等社交平台的日益流行, 社交网络已成为网络舆情生成、变化、发酵的主要平台。对于突发的重大、敏感事件以及其他公共议题事件, 网民通常会通过微博、微信等社交平台表达自身的观点和看法, 并迅速汇合成共同意见, 影响并推动舆情事件的走势与发展。与此同时, 一些不法分子通过社交网络的评论、留言等方式发布有害的网络内容, 从而误导民众, 对社会造成一定的负面影响, 这些留言和评论如果处理不当或者坐视不理, 负面影响则会被无限放大, 从而导致突发公共事件的衍生, 影响政府公信力和社会稳定。因此对可能影响社会稳定的网络舆情进行导控是十分必要的。

当前舆情导控的方法主要包括舆情采集、舆情分析、舆情预警、追踪导控四部分[1], 这些步骤它们通常需要大量的人工干预, 因此会耗费巨大的人力物力, 另一方面, 负面舆情通常具有数量大、范围广的特点。采用人工干预的方法也很难对所有负面舆情做出及时并且正确的回应。

对可能导致负面舆情的信息进行识别并做出正确的回应是智能回复系统的最新应用方向。与传统需要大量人工干预的舆情导控的方法不同, 智能回

通讯作者: 林政 Email: linzheng@iie.ac.cn

国家自然科学基金(No.61502478, No.61602467); 国家高技术研究发展计划(No.“863”计划)基金资助项目(No.2013AA013204)。

收稿日期: 2018-03-27; 修改日期: 2018-05-17; 定稿日期: 2019-12-19

复问答系统通过综合运用知识表示、信息检索、深度学习、强化学习、自然语言处理等技术, 能够在充分理解相关评论的内容后, 自动对评论进行及时且正确的回复, 从而对网络舆情进行正确的引导, 防止负面舆情扩散, 化解社会矛盾。

智能回复系统是指使用自然语言对用户输入的问句做出相应回复的智能系统, 如果将用户输入内容看作问题内容, 将回复看成答案时, 智能回复系统可以以对话系统的方式实现。智能回复系统不仅具有巨大的研究价值而且也有巨大的商业价值, 在家庭助手、语音助手、人工教学和智能对话等领域有着广泛的应用前景, 因此受到了工业界和学术界的研究关注。

互联网的发展和数据时代的到来为智能回复系统的发展带来了巨大的机遇, 一方面 Twitter, Facebook, 微博等社交平台聚集的大量活跃用户, 为智能回复系统的搭建提供了大量的公共对话语料; 另一方面在计算机视觉、语音识别、机器翻译等 AI 领域的研究都利用机器学习特别是深度学习的技术取得了突破性进展。近几年也出现了很多优秀的问答系统产品, 这些产品可以便捷地与用户通过自然语言进行交互, 如微软的小冰、iPhone 的 Siri、Google 的 Google 助手等, 正因为其不同于传统 APP 命令语言输入式的用户交互模式, 使其拥有了广泛的应用场景和用户群体, 可以预见未来智能回复系统将会有更加蓬勃的发展。

智能回复系统有很长的研究历史, 最早的问答系统是 20 世纪 60 年代的 ELIZA[2], 用于在临床中模仿心理医生, 1972 年 ELIZA 的变体 parry[3]在斯坦福诞生, 与 ELZA 不同, 它说话时会有自己的态度。1995 年 ALICEALICE[4]因为具有启发式能力, 能够和人类进行更有效的聊天, 使之成为 20 世纪最著名的聊天机器人。然而在 20 世纪早期的研究中, 智能回复系统主要是基于模板和规则的方法实现的, 对话的内容经常会局限于某个特定的场景[4-7], 例如 1988 年伯克利开发的智能机器人 UNIXConsultant, 它可以帮助人们学习如何使用 linux, 然而却不能和它在娱乐、体育等其他不同的话题上交流。

1990 年以后, 随着互联网的发展, 出现了大量可用来进行训练的公共语料, 这为搭建一个开放领域的智能回复系统提供了机遇。在此阶段, 人们主要利用信息检索或浅层语义理解技术在大量候选集中寻找对话的答案, 故检索式问答技术迅速发展, 如 1993 年 mit 开发的在线问答系统 START 系统和微软开发的 Enearta。2011 年, Ritter[8]首次提出可以把对话问题看成一个翻译过程, 即把一个问句翻译成一句回复。然而问答系统事实上比翻译问题更加困难, 因为一个问句可能有多个回复以及问句和答案很难像翻译系统一样进行对齐。如“一起去吃午饭吗?”的回答可以是“我吃过了”, 也可以是“还没有, 一起去吧”, 甚至可以是“抱歉, 一会约了朋友呢”。

近几年来, 一方面计算机的计算力得到了飞跃性提升, 另一方面, “大数据”时代的到来使我们有大量可用的对话语料, 并且神经网络的训练方法[9-10]也有了很大的进步, 这为使用神经网络进行智能回复系统的搭建和训练提供了机遇。2015 年, 基于神经网络的 NRM (Neural Responding Machine)模型[11]被应用到智能回复系统中, 该模型基于 Ritter 的思想将对话问题看成一个翻译问题, 然而与 Ritter 的 SMT[8](statistical machine translate)模型相比, NRM 模型的效果和性能却有了很大的改善和提升。NRM 模型也因此成为当前实现智能回复系统的主流方式。

从应用场景看, 当前智能回复系统可以分为限定领域的智能回复系统和开放领域的智能回复系统两种, 限定领域智能回复系统系统主要是帮助人们完成特定任务, 如酒店预定, 餐厅订餐等活动, 它通常由理解用户意图, 对话处理, 语言回复生成三部分组成, 典型产品有阿里小蜜和苹果的 siri 以及微软的 Cortana。该类型问答系统的问题输入是被限定的, 一旦问题超过该领域的范围, 问题将不被系统接受。与只能完成特定任务限定领域的任务型智能回复系统不同, 开放领域的智能回复系统能接受所有领域的问题并进行回答, 所以受到了越来越多的关注, 经典产品有小米的客服机器人以及微软的小冰。

从实现的技术方法看, 当前智能回复系统主要是基于检索模型实现的问答系统和生成模型实现的问答系统两种。在检索模型中, 系统会根据所给的问题从 Question-Answer 数据库中检索与该问题语义最相近问句所对应的答案, 作为该问句的答案。该方法的主要问题有两个: 一是数据库中 Question-Answer 对数量有限, 有可能检索不到用户提出的问题的答案; 二是问句之间的相似程度通常是通过词重叠率等表面特征来判断的[12], 对于语义相同但表述不同的问句, 可能无法检索到合适的答案。

在生成模型中, 智能回复系统会首先理解用户提出的问题, 然后逐字生成对应于该问题的答案。目前主流的方法是深度学习中 Seq2Seq 模型, 该模型首先在 encoder 端将问句编码为一个固定长度的向量, 再由 decoder 端将该向量解码为一个回复。该模型的主要问题是生成的答案很容易是通用的、单调的回

复, 例如: “我不知道”、“好的”等, 这样的回复包含的有用信息较少, 没有实质性的意义。此外, 生成模型还具有边界难对齐、易产生语法性错误以及无法考虑历史信息来保持对话一致性等问题。为了克服生成模型的不足和缺点, 生成模型的各种改进版本被提出, copynet[13]、ECM[14](Emotional Conversation model)等模型被证明可有效改善生成模型回复过于单调的问题, serban 提出的层次化生成模型则可以将对话历史加入到回复生成中来保持多轮对话中的回答一致性。可以预见, 基于神经网络的生成模型将成为未来研究的热点。

为了克服生成模型和检索模型各自的缺点, 融合生成模型和检索模型的混合模型被提出, 融合模型通常的做法是首先使用检索模型检索答案, 当检索不到问题答案时就用生成模型生成系统的答案, 或者将生成的答案和检索的答案一起作为候选集合, 进行重排序, 根据混合排序结果选择最优答案进行输出[15], 因为检索模型和生成模型具有各自的优势, 互为补充, 所以在实际应用中经常采用检索式和生成式的混合模型。

本文第 2 节介绍了智能回复系统主流的研究方向; 第 3 节根据应用场景都不同分别介绍了垂直领域的智能回复系统和开放领域的智能回复系统; 第 4 节从技术角度, 对目前主流智能回复系统的实现技术进行梳理; 第 5 节介绍了问答系统的评价标准并对其当前存在的问题进行了探讨; 第 6 节介绍了智能回复系统的可用的公共语料; 最后对全文进行了总结。

2 智能回复系统主流研究方向

2.1 视觉问答

视觉问答[16]是指用自然语言回答一个根据图片内容提出的问题, 如果要解决这个问题, 不仅需要理解图片的内容和问题的含义, 还需要理解文字和图片之间的关系。

当前实现视觉问答系统的方法主要包括基于贝叶斯框架的方法和基于深度学习的方法。

Kushal[17]利用贝叶斯框架实现了视觉问答系统, 该模型首先根据给定问题和相关图片预测答案类型, 然后计算在不同类别下每个特定答案的概率加和(也就是答案的边缘概率), 最后将概率最高的结果作为所问问题的答案。

基于深度学习的方法首先借助 LSTM(long short term memory)网络理解问题的含义, 然后使用卷积神经网络学习图像内容的表示[16]。最后, 通过将二者的向量表示采用直接相连的方法在自然语言问句和图片内容之间建立关联。Kim[17]和 Tao[18]在视觉问答任务上引入了注意力机制。注意力机制能够聚焦于图像、问题的重要部分, 从而给出更准确的答案。

2.2 基于知识图谱的问答

基于知识图谱的问答系统已经成为一种访问大型知识图谱的流行方式。通过访问知识图谱的结构化数据, 其可以使用自然语言来准确地回答事实性问题。

当前基于知识图谱的问答系统实现方法主要有基于规则、关键字搜索以及基于模板三种方法。

Mekhaldi [19-20]使用基于规则的方法将问题映射为结构化查询, 这种方法准确率高, 但是需要人工设计规则, 而且规则的泛化能力不强, 一个规则只能理解问题的一部分。

Unger [21]提出了基于关键字搜索的方法, 将问题映射为谓词再进行结构化查询, 但由于一个问题可以有不同的表示方法, 这一方法只能回答比较简单的问题。Unger[22]和 Yahya[23]在基于关键词搜索的方法上进行了扩展查询, 具体的做法是生成每个关键词的同义词之后, 使用同义词进行查询。

由于基于规则的方法泛化能力较弱, 且只能对问题做部分理解, 而基于关键词和同义词的方法不能全面理解用户提出的问题, 所以 Cai[24]提出了一种基于模板的问题表示方法, 这一方法不仅可以对问题的语义进行理解, 而且容易应用在大规模数据上, 使得模型具有较强的泛化能力。

2.3 推理问答

推理问答主要考验机器的智能理解能力, 它可以通过对已知知识的推理来得到未知的知识。推理问答的输入不仅有问题, 还有上下文, 它能够在阅读理解上下文之后, 对知识进行推理, 然后得到问题的正确答案。

当前推理问答系统的实现方法主要有基于规则的方法和基于记忆神经网络的方法。

早期的推理问答主要基于规则构建, 如 Etzioni[25]利用规则来表示知识库中实体之间存在的潜在关系。然而这种规则会随着关系数量的增加而出现爆炸性增长, 系统也不能明白这些规则的全部意义, 因此不适合在大规模数据上构建推理问答系统。

此后, Weston[26]将深度学习的技术应用在推理问答中, 提出的深度学习网络模型 Memory Network 具有长期记忆能力和推理能力。模型通过 IGOR 四个组件将长期记忆问题的能力和推理能力结合在一起, 从而能够对复杂的问题进行记忆并推理出准确的答案。Sukhbaatar [27]则对 Memory Network 进行了改进,

通过采取端到端的架构, 使模型能够直接使用问答语料对进行训练, 解决了 memory-network 需要大量监督学习的问题。

3 垂直领域智能回复系统VS.开放领域智能回复系统

根据应用场景的不同, 智能回复系统可以被分为垂直领域的智能回复系统和开放领域的智能回复系统。

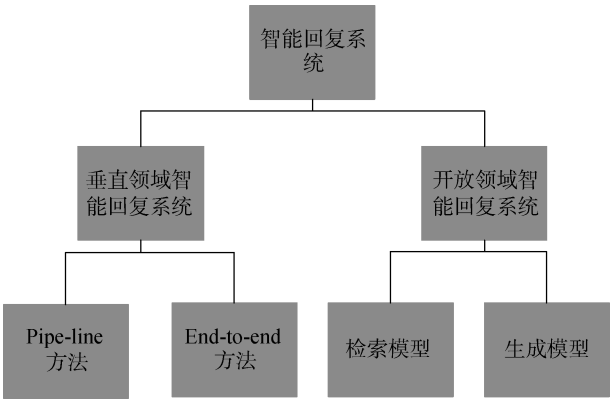

图 1 智能回复系统分类

Figure 1 The classification of dialogue system

3.1 垂直领域智能回复系统

垂直领域的智能回复系统, 也经常被称为任务型智能回复系统[28-30], 是早期人们主要使用的智能回复系统。这类智能回复系统主要是针对特定领域, 比如旅游语音助手、购物助手等。垂直领域智能回复系统的输入是被严格限制并且应该是可预测的, 比如进行酒店预定和公交车线路查询, 智能回复系统应该能够从问句中捕捉诸如抵达时间、出发地、目的地以及酒店或者公交车的基本信息, 然后做出相应回复。当前任务型智能回复系统的实现方式主要有 Pipeline Methods 和 end-to-end 两种。

3.1.1 Pipeline Methods

Pipeline 模型主要由 SLU (Spoken-Language Generation) 、NLG(Natural Language Understanding)、DM (Dialogue Management)三部分组成。

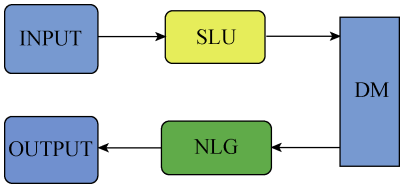

图 2 Pipeline Methods 架构图

Figure 2 Pipeline Methods framework

SLU 模块的作用是识别用户语言中的意图, 并将其转换为计算机可以理解的格式, 它通常包括用户意图探测[31-32]和 slot filling [33-35]两部分。

DM 模块的主要功能是协调管理问答系统的各个部分并根据 SLU 的传来的语义表示来制定回复用户的策略。当前比较成功的 DM 模型是基于局部可观测的马尔可夫决策过程[36-37]的。

NLG 模块负责将任务型智能回复系统 DM 产生的回复策略转换为自然语言对用户进行回复, 其性能好坏对用户体验有着极为重要的影响。NLG 部分的优秀与否主要由它的流畅性、可读性、充分性和多样性决定[38], 当前 NLG 的实现方法主要包括基于规则的方法[39]和基于语料的方法两种。

3.1.2 end-to-end 模型

在任务型智能回复系统中, Pipeline 的解决方法有不错的效果, 但是这样的方式需要将各个模块的训练分开, 需要大量的人工干预, 因此在模型泛化方面有很大的局限。相比于多模块的方式, end-to-end 模型只有单一的模块, 使得系统实现更为简洁, 便于训练和模型泛化。

Bordes[40]首次将 end-to-end 模型应用到任务型智能回复系统中, 通过使用 memory network 模型, 克服了 end-to-end 不适合任务型智能回复系统的缺点, 使得任务型智能回复系统可以进行单模块训练。

此后, Wen[41]通过对 end-to-end 模型进行了改进, 克服了 Bordes 模型[40]中不能获取外部知识的缺点。实验结果表明, 该模型在实际应用中可以帮人们出色地完成饭店预订和搜索等任务。

Zhao [42]等则首次将强化学习应用到端到端任务型智能回复系统中, 使得系统的表现更加卓越。

3.2 开放领域的智能回复系统

垂直领域智能回复系统主要关注特定任务的实现, 对话主题范围被限制在特定领域内, 超过对话主题范围的输入将不被接受。与垂直领域的智能回复系统不同, 开放型智能回复系统则可以在开放领域与用户进行闲聊, 对话主题也没有任何的范围限制。

目前开放领域问答系统的实现方法主要有基于检索的模型和基于生成的模型。

检索式模型利用选择算法从对话库中选择合适的回复, 因此回复具有语法正确、语义丰富的特点。然而因为语料库的对话语料是固定的, 因此问题必须限制在一定范围内。

生成模型则把开放领域智能回复系统的建立看

成一个翻译过程, 即在问题和答案之间形成翻译映射。该类方法并不是像检索模型一样直接返回答案, 而是根据用户所提的问题逐字地生成问题的答案。Ritter[8] 首次使用生成模型实现了开放式问答系统, 模型在 55%的情况下都优于传统的检索模型。随后, Vinyals [11]将深度学习应用到开放领域智能回复系统的生成模型中, 其提出的 NRM[11]方法极大提升了生成式问答系统的效果。基于深度学习的方法是目前实现生成式问答系统的主流方式。

4 检索模型 VS.生成模型

4.1 检索模型

基于检索模型的智能回复系统是一种针对大规模数据处理的方法。根据以自然语言方式提交的用户查询, 从大规模语料中检索到问题的答案并将其返回给用户。检索式问答系统需要正确理解以自然语言形式出现的用户的查询, 充分理解用户的查询意图, 并检索出与问句最为相关的的答案。因为检索式模型是利用选择算法从对话库中选择合适的回复, 因此回复具有流利、自然、信息丰富的特点。当前检索模型的实现方法主要可以分为可以改成基于浅层语义表示的方法和基于深度语义表示的方法。

4.1.1 基于浅层语义表示的方法

基于浅层语义表示方法实现的检索模型通常由过滤和重排序两部分构成, 他们通常使用人工提取的特征来计算问题与答案之间的相似度, 并以此来进行过滤与排序。

Jafarpour[43]首次利用检索模型来实现问答机器人, 它通过 bucket 过滤、MART 过滤排序和 AdaBoost 重排序三个步骤完成了开放领域智能回复系统的实现。

Wang[44]使用语义匹配、余弦值计算相似度等不同的匹配方式来过滤语料库中的答案得到答案的候选集, 然后通过对候选集中句子使用 ranksvm[45]进行重排序而得到所提问题的答案。

4.1.2 基于深度语义表示的方法

随着深度学习方法的兴起, 许多学者开始尝试将其应用在检索模型中, 通过构建诸多隐层的模型, 深度学习模型可以提取更深层的句子级特征, 从而提高检索模型的准确率。当前基于深度语义表示方法实现的检索模型主要可以分为单轮对话和多轮对话两类。

(1) 单轮对话

早期的智能回复系统主要是针对简单的单轮对话, 他们只根据当前信息检索答案而选择忽略复杂的对话历史。Lu[46]等首次使用 DNN 进行问句与答案的匹配, 该模型能在匹配过程中充分挖掘句子之间非线性和层次化的关系, 实验结果表明该模型超过了当时最出色的传统检索模型。

Hu[47]等首次将 CNN 应用在句子匹配过程中, 该模型能够在匹配过程中利用句子内部的结构和句子间的相互关系达到更好的匹配效果, 实验结果表明该模型在准确率和召回率上都超过了 DNN[47]模型。

Wu[48]提出人们在回答问题时, 不止应该考虑当前句子的信息, 还应该考虑当前对话的主题, 如问题“你吃饭了吗”, 人们除了回答“吃了”和“没有”之外还很有可能会讨论起三明治和汉堡等食物相关的话题。基于以上原因, Wu[48]将使用 LDA 模型[49]得到的 topic 加入到检索匹配过程中, 使得模型在匹配过程中不仅考虑了问题与答案的语义, 而且将问句的主题加入到匹配过中来引导模型挑选与问题更为匹配的答案。

(2) 多轮对话

单轮对话只是简单的考虑了当前的问题而完全忽略了对话的历史信息, 导致其很难在多回合对话中保持对话的一致性。与单轮轮对话不同, 多轮对话不仅需要考虑当前句子的语义信息, 更需要考虑对话历史和会话层的整体信息, 因此实现难度也比单轮对话更大。

	Context
utterance1	Human: How are you doing?
Utterance2	ChatBot: I am going to hold **a drum class** in shang hai
Utterance3	Human: Interesting! Do you hava coaches who can help me practice **drum**
Utterance4	ChatBot: of course
Utterance5	Human: Can I hava a free first lesson?
	Response Candidates
Response1	Sure.Have you ever played drum before? √
Response2	What lessons do you want? ×

图 3 多轮对话示例

Figure 3 A example of mul-turn dialogue

Ryan [50]针对检索模型的多轮对话问题, 提出了一个双向编码的模型, 该模型利用单词的词向量作为输入, 将 LSTM 的最后的状态向量作为对话历史与回复的语义表示进行问句匹配。实验结果表明, 使用基于 LSTM 的双向编码匹配模型的召回率比与传统检索模型提高了 33%。

类似于 Ryan[50]的工作, yan[51]首先使用使用双向递归神经网络分别得到对话历史和当前信息的表示, 然后使用卷积神经网络提取重要信息与答案进行匹配, 文中实验结果表明模型在 MAP 值和 nDCG 的评价标准上都超过了基于 LSTM 的多轮检索模型。

Wu[52]认为只是把对话历史和问句编码为一个固定长度的向量会丢失句子间的相关信息和重要的历史对话信息, 为了解决该问题, Wu[52]提出了序列化匹配模型 SMN(Sequential matching network), 该模型首先使用 CNN 从句子和词两个层次提取重要文本信息, 然后使用递归神经网络将对话历史和问句的重要文本信息相连接, 以此来保持句子间的相关信息。实验结果表示, 该模型的表现超过了当时最优秀的多轮对话检索模型。

4.2 生成模型

与检索模型不同, 生成模型则把开放领域智能回复系统的建立看成一个翻译过程。其不是像检索模型一样直接的返回答案, 而是根据用户所提的问题逐字的生成问题的答案。

Ritter[8]使用基于短语的 SMT 模型实现了第一个生成式智能回复系统, 在人工评价的标准下, 使用该生成系统的表现在 55%的情况下超过了传统的检索系统。

Vinyals [11]首次将深度学习应用到智能回复系统当中, 实验结果表明基于深度学习模型的生成式智能回复系统 NRM 的在各个评价指标下都明显优于 SMT 模型, 其实现方法 Seq2Seq 模型也因此成为当前实现智能回复系统的主流方式。

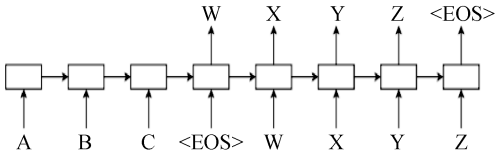

图 4 seq2seq 架构

Figure 4 The framework of seq2seq

如上图所示, 图中为 Sequence to Sequence 的基本架构, sequence-to-sequence 模型由 encoder 和 decoder 两部分组成, 该模型首先在 encoder 端将输入映射为一个固定长度的向量, 再在 decoder 端将向量解码为输出, 公式如下所示, 给定输入(x_1, x_2, …, x_T), 以下公式迭代可得($y_1, y_2, \cdots, y_T$)。

$$h_t = sigm\left(W^{hx}x_t + W^{hh}h_{t-1}\right) \tag{1}$$

$$y_t = W^{yh}h_t \tag{2}$$

一个良好的对话生成模型应该可以产生语法正确、信息量丰富并且上下文一致、逻辑清晰的回复。但是基于 sequence-to-sequence 实现的智能回复系统当生成较长回复时却容易产生不符合语法规则的句子, 也很容易生成“我不知道”, “好的”等通用却没有意义的回复。此外, 由于没有考虑对话的上下文, 从而模型不能在多轮对话中保持回答的一致性。Sequence-to-sequence 模型也没有考虑聊天对象的一致性[53], 因此对于同一用户所提出的语义相近的问题可能会产生不同的答案。与翻译模型相比, 还存在边界难对齐, 一个问题可能会有多种答案等问题。

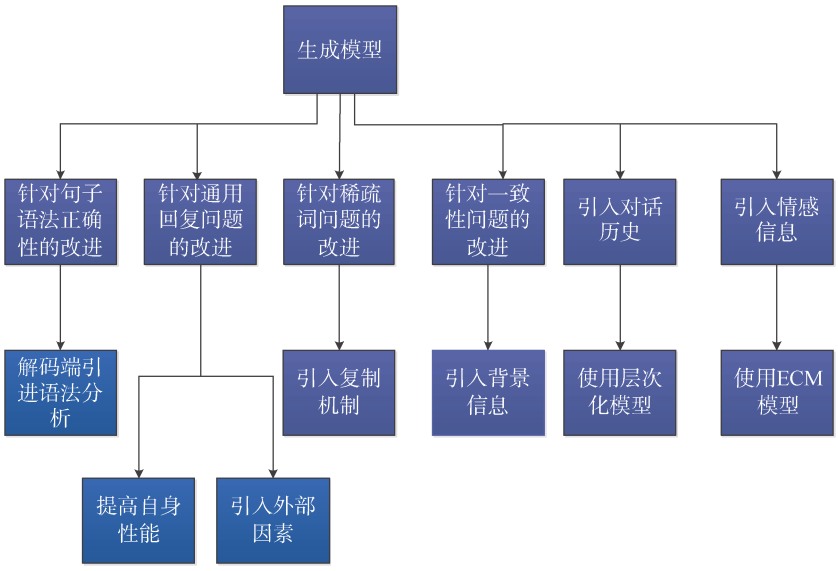

图 5 生成模型概览图

Figure 5 Generate model overview

为了解决以上问题, 许多研究者提出了很多 sequence-to-sequence 的改进方法, 包括引入复制机制、引入情感分析、引入背景信息解决回复一致性问题、引入句法分析、引入外部知识以及针对通用回复问题的改进。

4.2.1 针对通用回复问题的改进

Seq2Seq 模型总是倾向于产生“I don't know.”等单调通用而无意义的回复, 问题的出现主要与 Seq2Seq 模型使用最大似然函数作为目标函数和语料中通用回复出现的频率较大有关[54]。当前针对生成模型通用回复问题主要有两种观点, 第一种观点认为问句是输出答案产生的主要因素, 其他特征应该被提供给模型作为条件信息来引导模型生成具体的回复[55, 53]。第二种观点则认为应该注重于提高 Seq2Seq 模型自身的性能, 包括使用 beamsearch 及其变种[57]以及采用奖励模型来生成更长的回复。

为了解决 Seq2Seq 模型经常产生通用回复的问题, Li[54]使用最大互信息代替传统模型中的最大似然函数作为神经网络的目标函数, 通过对长回复进行奖励来引导模型生成更有意义的回复。文中的实验结果表明, MMI(Maximum mutual information)模型在 BLEU 和人工评价上都超过了传统的 sequence-to-seqence 模型。

Xing[56]为了解决 Seq2Seq 模型经常产生通用回复的问题, 通过对 attention 机制的改进, 将主题信息加入到 Seq2Seq 框架中, 然后利用主题模仿人类的先验知识来引导智能回复系统产生更有意义的回复, 该模型与标准的 Seq2Seq 以及 MMI 模型[54]做了比较, 其在 Perplexity 和人工评价方面的表现都明显优于基准模型。

Zhao[58]将 CVAE(Conditional Variational Auto-Encoder)模型引入到对话生成模型中, 使得模型能够在编码端捕捉到会话级别的多样性。模型使用一个隐藏变量来学习对话意图的分布继而产生不同的回复。为了克服优化中存在的困难, 模型还使用了 bag-of-word 损失来提升训练性能。实验结果表明, 相对于传统模型, 该模型所产生回复的多样性有比较显著的提高。

Zhou[59]使用隐藏向量对对话模型的的对话风格和对话主题等多种与回复有关的机制进行建模, 然后通过对不同回复机制的组合引导智能回复系统产生流利、语法正确、语义更为丰富的回复。实验表明该模型的回复多样性对比于传统 Seq2Seq 模型提高了 9.8%。

4.2.2 引入复制机制

因为词典大小是固定的, 所以传统的 sequence-to-sequence 模型不具备生成罕见词的能力。该问题被称为 OOV(out-of-vocabulary)问题, 其会使生成模型的对话效果大大下降[60]。

为了解决 OOV 问题对生成模型造成的不良影响, Gu[13] 在 Seq2Seq 模型中引入了复制机制, 该机制会抽取输入问题中的部分片段, 包括问句中的人名、实体名甚至很长的短语, 并把它们放到合适的位置上作为生成语句的一部分, 从而使得生成模型能够在一定程度上生成词表以外的稀疏词。模型与传统基于 rnn 的 Seq2Seq 在两个数据集上做了对比, 在 top1 和 top10(基于 beamsearch)的准确率都超过了传统的 Seq2Seq 模型。

4.2.3 引入个人信息

在传统对话生成模型中, 如果同一个人用不同的方法问相同的问题, 那么智能回复系统生成的答案就很可能会不同。该问题被称为 speaker consistency 问题。

message	Where do you live now?
response	I live in Los Angeles.
message	In which city do you live now?
response	I live in Madrid.
message	In which country do you live now?
response	England, you?
message	Where were you born?
response	I was born in Canada.
message	Where are you from?
response	England, you?
message	Where did you grow up?
response	I grew up in Texas.
message	How old are you?
response	16 and you?
message	What's your age?
response	18.
message	What is your major?
response	I'm majoring in psychology
message	What did you study in college?
response	English lit.

图 6 Speaker consistency 问题

Figure 6 Speaker consistency questions

为了解决对话生成模型中的 speaker consistency 问题, Li[53]将讲话者背景信息和讲话风格等个人特征引入到 sequence to sequence 模型中, 使得模型可以产生上下文一致的回复。实验结果表明, 对比于传统使用 LSTM 进行训练的 sequence-to-sequence 模型, 模型的 perplexity 和 BLEU 值都取得了突破性的提高。

4.2.4 引入对话历史

传统的 sequence-to-sequence 模型在生成回复

时通常只考虑到当前问题而选择忽略对话的上下文信息, 从而可能导致生成的回复与真实的情况并不相符。

图 7　多轮对话示例

Figure 7　A example of mul-turn conversation

为了将对话历史加入到回复生成的过程中, Sordoni [61]将层次化的端到端模型 HRED(hierarchical-recurrent-encoder-decoder)应用于开放领域的智能回复系统中。HRED 模型由编码端、解码端以及对话历史编码端三部分组成。通过对话历史编码端, HRED 模型可以将对话历史加入到回复生成的过程中, 从而使得生成的回复与对话上下文的逻辑和真实情况更加相符。

HRED[61]模型虽然在生成回复的过程中考虑了对话历史, 然而该模型却忽略了句子中不同单词的重要程度不同的特点, 从而可能导致模型在生成回复的过程中丢失重要信息, 为了解决该问题, serban[62]将 attention 机制[60]加入到层次化 sequence-to-sequence 模型中, 使模型能够在词语和句子两个层级对重要信息进行聚焦, 从而更好的生成相应的回复。

serban[25]提出了一个变分层次编码解码模型, 该模型将变分自编码器(VAE)[64] (Variational Auto- Encoder)加入到层次化对话生成模型中。它能够对生成过程中各个层次的随机变量建模, 从而生成语义更加丰富的句子。实验结果表明, 无论是使用自动评价标准还是人工评价标准, 结果都超过了传统的 sequence-to-sequence 模型和层次化模型 HRED。

4.2.5　引入情感分析

人们在交流过程中除了传递语义信息之外, 还经常使用带有情感色彩的词语来传递感情, 因此能够在对话中探测用户实时的情感并给予相应带有情感的回复能够提高用户对问答系统的满意度[65]。

Ghosh [66]提出了第一个能够传递情感信息的语言模型 Affect-LM(Affect-Language Model)。模型能够在不影响句子语法正确性的情况下, 根据输入的情感种类和情感强度生成不同的句子。大量的人工评价表明, 相比于传统语言模型, Affect-LM 能够极大的提高用户满意程度。

受启发于 Affect-LM , Huang[14]提出的 ECM 模型第一次将情感信息加入到智能回复系统中。ECM 是由情感类别词向量、内部情感记忆网络和外部记忆网络三部分组成的模型, 能够对情感因素建模并生成带有情感色彩的回复。人工评价表明该模型能够生成情感和内容都合适的回复, 并且人工评分也远高于传统的 NRM 模型。

4.2.6　引入外部知识

人们在聊天中回复的产生不止依赖于对话历史和当前信息, 也应该依赖于他们了解的背景知识[40]。Vougiouklis [67]首次提出一种基于数据和知识库的对话生成模型, 该模型能够通过访问知识库得到事实性信息并且能将该信息加入到答案生成过程中, 从而产生语义更加丰富的答案。与传统的 Seq2Seq 模型相比, 模型的 perplexity 值提高了 55%。

Yin[68]首次将访问外部知识的能力引入到 Seq2Seq 中。通过对传统 sequence to dequence 的解码端进行改进, 使其具有访问外部知识库的能力, 使得模型生成的回复语义更为丰富, 并能够回答简单的事实性问题。实验结果表明其在准确度上超过了传统的 NRM 系统。

4.2.7　引入语法分析

不同于其他序列化数据, 自然语言自身通常都具有一定的语法结构, 然而传统对话生成模型的解码端都是线性的生成结构, 因此在生成回复的过程中经常忽略句子间的语法结构[69], 从而影响了生成句子的流畅性和相关性。Filip [70]第一次在 Seq2Seq 的 decoder 端加入深度语法依赖树, 并应用于 NLG[71]。

此后, X2TREE 模型[69]将语言分析引入到生成模型的 decoder 端中, 使得模型能够在 decoder 端直接生成句法依赖树形式的回复。实验结果表明, 与传统的 Seq2Seq 模型相比, 该模型回复的可接受率提高了 11.5%。

4.3　检索和生成的混合模型

检索模型常常会产生长尾问题[15]以致于无法搜索到与用户问题最适合的答案, 而生成模型则很容易产生“好的”、“我不知道”等单调无意义的回复。

为了克服生成模型和检索模型各自的缺点, 融合生成模型和检索模型的混合模型被提出。融合模

型通常的做法是, 首先使用检索模型检索答案, 当检索不到问题答案时就用生成模型生成系统的答案, 或者将生成的答案和检索的答案一起作为候选集合, 进行重排序, 根据混合排序结果选择最优答案进行输出[15], 因为检索模型和生成模型具有各自的优势, 互为补充, 所以在实际应用中经常采用检索式和生成式的混合模型。

Alime[15]提出了一个由混合模型实现的开放领域的问答系统。该系统由检索模型、生成模型和重排序模型三部分组成。通过使用加入了注意力机制的 sequence-to-sequence 模型对检索模型的答案进行重排序、挑选、生成等步骤, 对模型的结果进行优化, 使得模型的性能超过了传统的检索模型和生成模型。

与 Alime 不同, Song[12]则提出了一种新的混合模型实现方法。模型首先使用检索模型检索到问题的答案, 然后使用检索模型的答案来引导生成模型生成更有意义的回复。

表 1　智能回复系统方法比较

Table 1　The comparison of intelligent reply system emotion analysis methods

		数据集	评价方法	结果	引用	优点	缺点
检索模型	BM25	Baidu douban	P@1	0.272	Ref[87]	检索模型产生的回复具有流利、自然、语义丰富的特点	因为数据库语料有限, 因此检索模型有可能检索到的答案与问题不相适应或者检索不到问题的答案
	SVM	weibo	P@1	0.574	Ref[58]		
	DNN	weibo	P@1	49.85%	Ref[46]		
	CNN	weibo	P@1	61.95%	Ref[47]		
	TANN	Twitter	MAP	0.557	Ref[48]		
	Lstm	Ubunru	R10@2	0.745	Ref[50]		
	SMN	Ubuntu	R10@2	0.847	Ref[52]		
	DL2R	Baidu douban	P@1	0.731	Ref[51]		
生成模型	Copynet	DS-I	Top1	61.2%	Ref[13]	生成模型根据问题住逐字的生成问题的答案, 确保答案与问题是完全对应的	生成模型很容易生成通用无意义的回复, 而且不能生成稀疏词, 生成过长的回复时还很可能会出现语法错误
	MMI	Open Subtitles	Bleu	1.74	Ref[54]		
	persona models	TV series dataset	Bleu	1.82	Ref[53]		
	kgCVAE	Sw	perplexity	16.02	Ref[58]		
	MARM	Tencent Weibo	Accept ratio	64.67	Ref[59]		
	ECM	STC	perplexity	65.0	Ref[89]		
	HRED	Twitter Dataset	perplexity	64.89	Ref[61]		
	VHRED	Ubuntu Dialogue	Embedding metrics	0.396	Ref[63]		
混合模型	AliMe	taobao chat log	Top1	0.60.36	Ref[15]	通过对不同模型的融合, 一定程度上避免了检索不到问题答案和容易生成通用单调回复的缺点	如何对不同类型的模型进行融合是问题的难点
	Song	sina weibo	bleu	1.06%	Ref[12]		

最后把生成和检索出来的答案一起作为新的候选答案集进行重排序, 从而得到原来问题的答案。模型在人工评分和 BLEU 值上都超过了检索模型和标准的端到端模型, 使用 biSeq2Seq 还使得智能回复系统的性能提高了 13%。

MILABOT[72]是一个深度强化对话机器人, 它能够和人类使用语音或者文本在娱乐、时尚、政治、体育和科技等领域进行流利地交流。该模型是语言生成模型和检索模型等的结合, 包括基于模板的模型、词袋模型、 sequence-to-sequence 模型和隐藏神经网络模型等多种模型。该模型能够通过在与真实用户的交流中使用利用强化学习不断提升自己, 大量的 A/B 测试显示, 该模型的表现超过了小冰、siri 等问答系统。

5　常用评价方法

在智能回复系统中由一个很重要的问题就是如何对其产生的回复进行评价。早期对智能回复系统生成回复好坏主要采用人工评价。人工评价具有较高的准确性, 然而在不断使用人工评价的过程中, 人们发现该方法比较昂贵和费时, 而且也不具有可扩展性, 评价人员的不同也不可避免地会引入评价的偏差, 因此人们开始将重点着眼于问答系统的自动评价标准。

近些年来问答系统的发展虽然很迅速, 但是如何制定问答系统的自动评价标准仍然是一个很具有挑战性的工作。当前智能回复系统自动评价方法主要有两类: 第一类是通过词语的重叠率来评价生成答案, 例如 BLEU 、METEOR 、ROUGE 等方法, 第二类是词向量统计方法, 这一想法主要来源于语言模型 word2vector[75]。

5.1 基于重叠词

因为智能回复系统的回复是无结构的, 因此自动评价智能回复系统生成回复的质量是一项很具有挑战性的工作。目前的基于词重叠评价方法是有偏差的, 并且与人类的判断关联度较小。当前的基于词重叠评价标准主要包括以下几个方法:

BLEU[73]的计算方法主要是比较候选答案和标准答案的 n-gram 集合, 该方法与词语的位置无关, 候选答案和标准答案的匹配程度越高, 该候选答案的 BLEU 值越高。

Perplexity[81]是衡量自然语言处理领域中语言模型好坏的指标, 该方法主要通过计算一个句子的困惑度来衡量答案的可靠性, 一个句子的生成可能性越大它的值越高。

NIST[74](National Institute of standards and Technology)方法是对 BLEU 方法的一种改进。它并不是简单的将匹配的 n-gram 片段数目累加起来, 而是求出每个 n-gram 的信息量(information), 累加后除以整个译文的 n-gram 片段数目。

METEOR[75]标准由 Lavir 发现召回率在评价指标中的意义后提出, 该标准基于单精度的加权调和平均数和单字召回率, 并具有同义词匹配等功能。研究结果表明, 与单纯基于精度的标准(例如 BLEU 和 NIST)相比, 其结果与人工判断的结果有更高的相关性。

5.2 基于词向量

与基于重叠词的方法不同, 基于词向量的评价方法先使用语言模型如 word2vec[76]计算每个词语的词向量, 然后再用该词向量得到句子级的向量, 机 sentence embedding, 最终候选答案与标准答案的相似度就可以用句子级向量的余弦距离等方式进行度量。当前常见基于词向量的评价方法主要有 Greedy matching, Embedding Average 和 Vector Extrema 三种。

Embedding Average 方法使用句子中所有词的词向量的平均值[76]作为句子的 sentence embedding, 然后使用余弦值来计算两个句子的相似度。

Vector Extrema 也是利用 sentence embedding 来计算句子相似度的方法, 然而不同于 Embedding Average 方法, Vector Extrema[77]方法利用向量抽取的方式在每一维上抽取所有词向量该维的极值来作为 sentence embedding 该维的值, 从而计算出句子级向量。

不同于以上两种方法, Greedy matching[78]并不直接计算句子的 sentence embedding。模型首先使用贪心的算法对句子和匹配语句的词进行两两匹配, 形成词匹配对, 然后使用 word-embedding 计算每个词匹配对的余弦距离, 选择其中最大的值作为两个句子的相似度。

5.3 其他方法

Liu[78]等发现传统的智能回复系统自动评价方式与人类的判断之间的相关度很小, 因此人们开始探索新的智能回复系统评价标准。

Liu[78]提出了一种基于机器学习的对话自动评价模型, 模型基于层次化 RNN 并使用半监督的方法来学习预测人类对回复的评分, 文中实验结果表明模型评分度和人类判断的相关度远远超过 BLEU。

Tao 等[79]提出以下三个观察结果:

(1) 生成的回复可能与标准答案的相同词语很少, 但仍然是一个很好的回复。

(2) 生成的回复可能与标准答案的语义不是很相近, 但相对于该问题仍然是一个很好的答案。

(3) 问题本身能对于区分答案的好坏提供很多有用的信息。

基于以上三个观察结果 Tao[79]又提出一种叫做 RUBER 的模型来评价生成回复的质量, 该模型首先计算生成回复和标准答案的相关度以及生成回复和问题的相关度, 然后使用启发式的思想合并两个得分来计算该回复的评分。实验结果表明该模型评价得分与人工评价的得分很相近。

6 语料资源归纳

对话语料是实现智能回复系统的基础, 其优秀与否直接影响着智能回复系统的性能。当前人们获取对话语料的主要来源是推特、ubuntu 和微博等开放的社交平台。

6.1 推特语料

推特是一个社交类型的网站, 用户可以在上面发布不超过 140 字的“推文”, 稍后其他用户可以对状态进行评价和交流。推特用户通过状态-回复之间的互动很适合用作智能回复系统的训练语料, 如 Ritter[8]为搭建生成式智能回复系统引入的约 1300000 对的推特对话语料如今被广泛使用。

6.2 Ubuntu 语料

Ubuntu 的日志中包括大量硬件和软件相关的问题和答案, 这些对话语料可以被认为是智能回复系统的训练语料。当前最著名的 Ubuntu 语料是 Ryan[84] 引入的, 该对话语料包含 100 万个多轮对话语料, 其中包含 700 万个句子和一亿个单词, 对话语料平均轮数是 8 轮, 最少为 3 轮。为搭建神经网络智能回复系统提供了大量无标签数据。

6.3 微博语料

微博是中国最流行的类推特社交服务型网站, 它是一个公共开放平台, 用户可以公开或者仅面向自定义的用户分组发布状态, 然后在该状态下能够收到其他用户的回复, 这些状态的回复通常具有灵活的形式和不同的主题。微博灵活公开的对话形式为对话系统的建立提供了大量的天然语料。Shang [80] 从微博中爬取了上亿的 post-question 对作为候选语料, 并对该语料进行去通用回复、过滤掉广告等处理, 得到了 4435959 个问答对作为问答系统的训练语料。Liu [41]使用爬虫技术从微博中抽取了大量微博对话语料, 并对数据进行了清洗和 label 标注等工作, 目前该数据已经公开。

表 2 公开语料对比

Table 2 The comparison of open data set

数据集	数据集大小	引用
Twitter	1300000	Ref[8]
Twitter couple	4232	Ref[85]
Sina weibo	4308211	Ref[80]
Sina weibo	618104	Ref[55]
Ubuntu corpus	930000	Ref[50]
Switch board	2400	Ref[86]
Movie trople	196308	Ref[61]

Wang[44]在微博上选中了 3200 位 NLP 和 ML 方向的专家用户, 然后用爬虫爬取他们的状态和对应回复作为 QA 问答训练语料, 该语料包括原始评论-回复对, 人工标注问答对以及每个评论对应的多个回复三个部分。

7 总结

本文在充分调研和深入分析的基础上对智能回复系统的研究进展进行了综述, 重点介绍了开放领域的智能回复系统、垂直领域的智能回复系统和当前实现智能回复系统的主流方法及其改进, 包括检索模型、生成模型和混合模型。

智能回复系统被认为是图灵测试的原始形式, 人们在自然语言处理领域对其进行的研究已经有很多年的时间, 然而智能回复系统尚有许多值得深入探索的问题。本文的最后, 我们基于大量的调研和近几年来的研究经验提出了一些该领域值得进一步挖掘的研究点, 希望对本领域的其他研究者有所启发:

(1) 生成回复过于单调等问题

当前实现智能回复系统的主流方式是使用 Seq2Seq 模型, 然而该模型却总是倾向于产生通用却没有实际意义的回复, 例如“我不知道”、“好的”等。虽然有很多学者提出了引入各种外部信息来缓解该问题, 但是它们通常都是针对特定因素进行优化, 因此很难在开放领域的智能回复系统中达到实用的程度。因此生成模型生成答案的单调性问题仍然亟待解决。

(2) 自动评价标准

对传统智能回复系统的自动评价主要包括基于词向量和基于重叠词两种评价标准, 然而这些的自动评价标准却与人类的评分只有很小的相关度, 因此可以预见如何制定可靠的智能回复系统的自动评价标准将会成为未来研究的热点。

(3) 生成稀疏词问题

当前生成模型因为使用概率模型预测单词, 因此有词表不能过大的问题, 当前主流方法是将低频词(稀疏词)设为 unk, 所以生成模型在回复生成过程中具有不能生成稀疏词的缺点, 从而导致生成回复的可理解性大大下降。因此如何将稀疏词引入到生成系统中也是一个值得研究的问题。

(4) 冷启动问题

当前基于深度学习模型的智能回复系统的训练通常都需要大量的语料, 然而在一些特定领域, 对话数据的收集是比较困难的。解决该问题的方法之一就是让对话模型具有使用少量的语料进行学习的能力。可以预见, 如何使智能回复系统具有快速自动学习的能力将成为未来研究的趋势。

参考文献

[1] Zhang X, Sun J H. Design of Public Opinion Monitoring System[J]. *Modern Electronics Technique*, 2015, 38(11): 98-102.
(张昕, 孙江辉. 舆情监测系统设计[J]. *现代电子技术*, 2015, 38(11): 98-102.)

[2] Weizenbaum J. ELIZA——-a Computer Program for the Study of Natural Language Communication between Man and Machine[J]. *Communications of the ACM*, 1966, 9(1): 36-45.

[3] Colby, K.,. Artificial Paranoia: A Computer Simulation of Paranoid Processes[J]. *Behavior Therapy*, 1976, 7(1): 146.

[4] Wallace R S. The Anatomy of A.L.I.C.E[M]. Parsing the Turing

Test. Dordrecht: Springer Netherlands, 2009: 181-210.

[5] Price, P. J. Evaluation of Spoken Language Systems: The ATIS domain[C]. *In Proceedings of the DARPA Workshop on Speech and Natural Language*, 2016.

[6] Hemphill C. T., Godfrey J. J., Doddington G. R. The ATIS Spoken Language Systems Pilot Corpus. *In Proceedings of the workshop on Speech and Natural Language*,2014.

[7] Dahl D. A., Bates M., Brown M., et al. Expanding the Scope of the ATIS Task: The ATIS-3 Corpus[M]. In Proceedings of a Workshop held at Plainsboro, New Jerey, USA, March 8-11, 1994.

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

[9] Tai K. S., Socher R., Manning, C. D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In ACL2015.

[10] Tieleman, T. , Hinton G. Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning. Technical report, 2012.

[11] Sutskever, Ilya, Oriol Vinyals. Sequence to Sequence Learning with Neural Networks[M]. Advances in neural information processing systems, 2014.

[12] Song Yiping. Two are Better than One: An Ensemble of Retrieval-and Generation-Based Dialog Systems. arXiv preprint arXiv:1610.07149 (2016).

[13] Gu J, Lu Z, Li H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence learning. arXiv preprint arXiv:1603.06393, 2016.

[14] Zhou H, Huang M L, Zhang T Y, et al. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory[EB/OL]. 2017: arXiv:1704.01074[cs.CL]. https://arxiv.org/abs/1704.01074.

[15] Qiu M H, Li F L, Wang S Y, et al. AliMe Chat: A Sequence to Sequence and Rerank Based Chatbot Engine[C]. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.

[16] Antol S, Agrawal A, Lu J S, et al. VQA: Visual Question Answering[C]. *2015 IEEE International Conference on Computer Vision (ICCV)*, December 7-13, 2015. Santiago, Chile. Piscataway, NJ: IEEE, 2015.

[17] Kushal Kafle ,Christopher Kanan. 2016. Answertype prediction for visual question answering[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 123-132.

[18] Akira Fukui, Dong Huk Park, Daylen Yang, et al. Ual Question Answering and Visual Grounding[C]. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016:235-242.

[19] Jin Hwa Kim, Kyoung Woon On, Jeonghee Kim, et al. Hadamard product for low-rank bilinear pooling. In International Conference on Learning Representations, 2017:236-245.

[20] S. Ou, C. Orasan, D. Mekhaldi, et al. Automatic Question Pattern Generation for Ontology-based Question Answering. In FLAIRS, 2008: 183–188.

[21] Unger C, Cimiano P. Pythia: Compositional Meaning Construction for Ontology-Based Question Answering on the Semantic Web[M]. Natural Language Processing and Information Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 153-160.

[22] C. Unger, L. Bühmann, J. Lehmann, et al. Template-Based Question Answering Over Rdf Data.

[23] M. Yahya, K. Berberich, S. Elbassuoni, et al. Natural Language Questions for the Web of Data[C]. EMNLP-CoNLL, 2012: 379-390.

[24] Cui W Y, Xiao Y H, Wang H X, et al. KBQA: Learning Question Answering over QA Corpora and Knowledge Bases[J]. *Proceedings of the VLDB Endowment*, 2017, 10(5): 565-576.

[25] Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, et al. Learning First-Order Horn Clauses from Web Text[C]. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2010:1088-1098.

[26] Jason Weston, Sumit Chopra, Antoine Bordes. Memory networks. arXiv:1410.3916v11.

[27] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, et al. End-to-end memory networks. arXiv:1503.08895v5.

[28] Walker M A, Passonneau R, E. Boland J L. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems[C]. 2001: 515-522.

[29] Jason Williams, Antoine Raux, Deepak Ramachandran, et al. The Dialog State Tracking Challenge[C]. *Proceedings of the SIGDIAL 2013 Conference*. 2013: 404-413.

[30] Ke Zhai, Jason D Williams. Discovering latent structure in task-oriented dialogues[C]. ACL: Meeting of the Association for Computational Linguistics. 2014: 120-126.

[31] Deng L, Tur G, He X D, et al. Use of Kernel Deep Convex Networks and End-to-end Learning for Spoken Language Understanding[C]. *2012 IEEE Spoken Language Technology Workshop (SLT)*, December 2-5, 2012. Miami, FL, USA. Piscataway, NJ: IEEE, 2012: 210-215.

[32] Dauphin Y, Tur G, Hakkani-Tur D, et al. Zero-Shot Learning and Clustering for Semantic Utterance Classification[EB/OL]. 2013: arXiv:1401.0509[cs.CL]. https://arxiv.org/abs/1401.0509.

[33] A. Deoras ,R. Sarikaya. Deep Belief Network Based Semantic Taggers for Spoken Language Understanding. In Interspeech, 2013: 2713-2717.

[34] Deng L, Tur G, He X D, et al. Use of Kernel Deep Convex Networks and End-to-end Learning for Spoken Language Understanding[C]. *2012 IEEE Spoken Language Technology Workshop (SLT)*, December 2-5, 2012. Miami, FL, USA. Piscataway, NJ: IEEE, 2012: 210-215.

[35] Tur G, Hakkani-Tur D, Heck L, et al. Sentence Simplification for Spoken Language Understanding[C]. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 22-27, 2011. Prague, Czech Republic. Piscataway, NJ: IEEE, 2011: 5628-5631.

[36] Young S, Gasic, M, Thomson B, et al. POMDP-based statistical spoken dialog systems[J]. *Proceedings of the IEEE*. 2013, 101(5): 1160-1179.

[37] Young S, Gašić M, Keizer S, et al. The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management[J]. *Computer Speech & Language*, 2010, 24(2): 150-174.

[38] Stent A, Marge M, Singhai M. Evaluating Evaluation Methods for Generation in the Presence of Variation[M]. Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 341-351.

[39] Danilo Mirkovic, Lawrence Cavedon. Dialogue Management Using Ccripts[C]. *EP Patent 1*, 2011: 891-897.

[40] Bordes A, Boureau Y L, Weston J. Learning End-to-End Goal-Oriented Dialog[EB/OL]. 2016: arXiv:1605.07683[cs.CL]. https://arxiv.org/abs/1605.07683.

[41] Wen T H, Vandyke D, Mrksic N, et al. A Network-based End-to-End Trainable Task-oriented Dialogue System[EB/OL]. 2016: arXiv:1604.04562[cs.CL]. https://arxiv.org/abs/1604.04562.

[42] Zhao T C, Eskenazi M. Towards End-to-End Learning for Dialog State Tracking and Management Using Deep Reinforcement Learning[C]. *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 1-10.

[43] Jafarpour Sina, Christopher JC Burges, Alan Ritter. Filter, Rank, and Transfer the Knowledge: Learning to chat[C]. *Advances in Ranking 10*, 2010: 2329-9290.

[44] Hao, Wang, Lu Zhengdong, Li Hang. A Dataset for Research on Short-Text Conversation[C]. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013: 935-945.

[45] Joachims T. Optimizing Search Engines Using Clickthrough Data[C]. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, July 23-26, 2002. Edmonton, Alberta, Canada. New York, USA: ACM Press, 2002: 133-142.

[46] Hu B T, Lu Z D, Li H, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[EB/OL]. 2015: arXiv:1503.03244[cs.CL]. https://arxiv.org/abs/1503.03244.

[47] Z. Lu ,H. Li. A Deep Architecture for Matching Short Texts[C]. *In International Conference on Neural Information Processing Systems*, 2013: 1367-1375.

[48] Wu Y, Wu W, Li Z J, et al. Topic Augmented Neural Network for Short Text Conversation[EB/OL]. 2016: arXiv:1605.00090[cs.CL]. https://arxiv.org/abs/1605.00090.

[49] Zhao W X, Jiang J, Weng J S, et al. Comparing Twitter and Traditional Media Using Topic Models[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 338-349.

[50] Ryan Lowe, et al., 2016, The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems[C]. *SIGDIAL 16*. 2015: 285-294.

[51] Yan R, Song Y P, Wu H. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System[C]. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, July 7-21, 2016. Pisa, Italy. New York, USA: ACM Press, 2016: 55-64.

[52] Wu Y, Wu W, Xing C, et al. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots[C]. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 496-505.

[53] Li J, Galley M, Brockett C, et al. A Persona-Based Neural Conversation Model. arXiv preprint arXiv:1603.06155. A Persona-Based Neural Conversation Model, 2016.

[54] Li J W, Galley M, Brockett C, et al. A Diversity-Promoting Objective Function for Neural Conversation Models[EB/OL]. 2015: arXiv:1510.03055[cs.CL]. https://arxiv.org/abs/1510.03055.

[55] Ji Zongcheng, Lu Zhengdong, Li Hang. An Information Retrieval Approach to Short Text Conversation. arXiv preprint arXiv:1408.6988, 2014.

[56] Xing C, Wu W, Wu Y, et al. Topic Aware Neural Response Generation[EB/OL]. 2016: arXiv:1606.08340[cs.CL]. https://arxiv.org/abs/1606.08340.

[57] Wiseman S, Rush A. Sequence-to-Sequence Learning as Beam-Search Optimization[EB/OL]. 2016: arXiv:1606.02960[cs.CL]. https://arxiv.org/abs/1606.02960.

[58] Zhao T C, Zhao R, Eskenazi M. Learning Discourse-level Diversity for Neural Dialog Models Using Conditional Variational Autoencoders[EB/OL]. 2017: arXiv:1703.10960[cs.CL]. https://arxiv.org/abs/1703.10960.

[59] Ganbin Zhou, Ping Luo, Rongyu Cao, et al. Mechanism-Aware Neural Machine for Dialogue Response Generation[C]. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2017:134 -142.

[60] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[EB/OL]. 2014: arXiv:1409.0473[cs.CL]. https://arxiv.org/abs/1409.0473.

[61] Serban I V, Sordoni A, Bengio Y, et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models[EB/OL]. 2015: arXiv:1507.04808[cs.CL]. https://arxiv.org/abs/1507.04808.

[62] Xing C, Wu W, Wu Y, et al. Hierarchical Recurrent Attention Network for Response Generation[EB/OL]. 2017: arXiv:1701.07149 [cs.CL]. https://arxiv.org/abs/1701.07149.

[63] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, et al. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. arXiv:1605.06069v3.

[64] Kingma D P, Welling M. Auto-Encoding Variational Bayes[EB/OL]. 2013: arXiv:1312.6114[stat.ML]. https://arxiv.org/abs/1312.6114.

[65] Partala T, Surakka V. The Effects of Affective Interventions in Human–computer Interaction[J]. *Interacting With Computers*, 2004, 16(2): 295-309.

[66] Ghosh S Y, Chollet M, Laksana E, et al. Affect-LM: A Neural Language Model for Customizable Affective Text Generation[EB/OL]. 2017: arXiv:1704.06851[cs.CL]. https://arxiv.org/abs/1704.06851.

[67] P. Vougiouklis, J. Hare, E. Simperl. A Neural Network Approach for Knowledge-Driven Response Generation[C]. *The 26th International Conference on Computational Linguistics*. 2016:3370-3380.

[68] J. Yin, X. Jiang, Z. Lu, et al. Neural Generative Question Answering[C]. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16*, 2016: 2972-2978.

[69] Zhou G B, Luo P, Cao R Y, et al. Tree-Structured Neural Machine for Linguistics-Aware Sentence Generation[EB/OL]. 2017: arXiv:1705.00321[cs.AI]. https://arxiv.org/abs/1705.00321.

[70] Dušek O, Jurčíček F. Sequence-to-Sequence Generation for Spoken Dialogue Via Deep Syntax Trees and Strings[EB/OL]. 2016: arXiv:1606.05491[cs.CL]. https://arxiv.org/abs/1606.05491.

[71] Bateman, John, Michael Zock. Natural Language Generation[M]. The Oxford handbook of computational linguistics. 2003.

[72] I. V. Serban, C. Sankar, M. Germain, et al. A Deep Reinforcement Learning Chatbot. arXiv preprint arXiv:1709.02349, 2017.

[73] Kishore Papineni, Salim Roukos, Todd Ward, et al. BLEU: a method for automatic evaluation of machine translation[C]. *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002:311-318.

[74] Doddington G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics[C]. *Proceedings of the second international conference on Human Language Technology Research* -, March 24-27, 2002. San Diego, California. Morristown, NJ, USA: Association for Computational Linguistics, 2002: 71-78.

[75] Denkowski M, Lavie A. Meteor Universal: Language Specific Translation Evaluation for any Target Language[C]. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 376-380.

[76] G. Forgues, J. Pineau, J.-M. Larcheveque, et al. Bootstrapping Dialog Systems with Word Embeddings[C]. *NIPS2014*, 2014:168-172.

[77] Lowe R, Noseworthy M, Serban I V, et al. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses[EB/OL]. 2017: arXiv:1708.07149[cs.CL]. https://arxiv.org/abs/1708.07149.

[78] Liu C W, Lowe R, Serban I, et al. How NOT to Evaluate your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation[C]. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 2122-2132.

[79] Chongyang Tao, Lili Mou, Dongyan Zhao, et al. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. arXiv preprint arXiv:1701.03079, 2017.

[80] Shang, Lifeng, Zhengdong Lu, Hang Li. Neural Responding Machine for Short-Text Conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[81] T. Mikolov, M. Karafi´at, L. Burget, et al. Recurrent Neural Network Based Language Model[J]. *In Interspeech*. 2010, 2(1): 1045-1048.

[82] Blei, David M., Andrew Y. Ng, et al. Latent Dirichlet Allocation[J]. *Journal of machine Learning research*. 2003, 3(1): 993-1022.

[83] C. D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval[M]. Cambridge University Press, 2008.

[84] Ryan Lowe, Nissan Pow, Iulian Serban, et al. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. arXiv:1506.08909v3, 2016.

[85] A. Sordoni, M. Galley, M. Auli, et al. A Neural Network Approach 293 to Context-Sensitive Generation of Conversational Responses. arXiv:1506.06714v1, 2015.

[86] Godfrey J J, Holliman E C, McDaniel J. SWITCHBOARD: Telephone Speech Corpus for Research and Development[C]. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 23-26, 1992. San Francisco, CA, USA. Piscataway, NJ: IEEE, 1992: 517-520.

[87] C. D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval[M]. Cambridge University Press, 2008.

[88] Zhao W X, Jiang J, Weng J S, et al. Comparing Twitter and Traditional Media Using Topic Models[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 338-349.

[89] Zongcheng Ji, Zhengdong Lu, Hang Li. An information Retrieval Approach to Short Text Conversation. arXiv preprint arXiv: 1408.6988, 2014.

[90] Zens R, Och F J, Ney H. Phrase-Based Statistical Machine Translation[M]. KI 2002: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 18-32.

[91] M. Ghazvininejad, C. Brockett, M.-W. Chang, et al. A Knowledge-Grounded Neural Conversation Model. arXiv preprint arXiv:1702.01932, 2017.

岳世峰 出生于 1993 年, 于 2019 年在中国科学院大学大学获得硕士学位。Email: yueshifeng@iie.ac.cn

王伟平 出生于 1975 年, 中国科学院信息工程研究所博士生导师、研究员, 主要研究方向为大数据与人工智能。Email: wangweiping@iie.ac.cn

林政 出生于 1984 年, 副研究员、硕士生导师, 她的主要研究方向为自然语言处理与情感分析。Email: linzheng@iie.ac.cn

孟丹 生于 1965 年，研究员、博士生导师，他的主要研究方向为分布式系统与系统安全。Email: mengdan@iie.ac.cn