# 检索式自动问答研究综述

赵 芸[1),2),3)] 刘德喜[1),3)] 万常选[1),3)] 刘喜平[1),3)] 廖国琼[1)]

[1)](江西财经大学信息管理学院 南昌 330013)

[2)](江西科技师范大学数学与计算机科学学院 南昌 330038)

[3)](江西财经大学数据与知识工程江西省高校重点实验室 南昌 330013)

**摘 要** 自动问答是人工智能和自然语言处理领域的一个研究热点,它最初是为了满足人们快速、准确地获取信息的需求,随着技术的发展,现有的自动问答模型大多无领域限制、可接收文本和语音输入.检索式自动问答是自动问答的重要技术路线,虽然近年来取得了丰硕的成果,但对这些成果进行总结分析的综述类文献或者比较早期、没有纳入新的成果,或者聚焦于某一个单独领域、没有从整体上进行总结分析.本文对问答模型的分类、技术方法、数据集和评价指标进行了比较全面的综述.首先,介绍自动问答的分类方法以及典型类型,总结了不同类型问答模型的特点以及常用的技术方法;然后,以检索式问答模型为主要对象,讨论常用的三类方法,分析了各类方法的特点以及难点,针对不同的难点,总结归纳了现有的改进技术;随后,介绍了检索式自动问答现有的评价方法和数据集;最后,总结现有方法存在的问题,并探讨了检索式自动问答将来的发展趋势和可能的挑战.

**关键词** 检索式自动问答;特征表示;深度学习;神经网络;评价方法;数据集

**中图法分类号** TP18 **DOI 号** 10.11897/SP.J.1016.2021.01214



# Retrieval-Based Automatic Question Answer: A Literature Survey

ZHAO Yun[1),2),3)] LIU De-Xi[1),3)] WAN Chang-Xuan[1),3)] LIU Xi-Ping[1),3)] LIAO Guo-Qiong[1)]

[1)](*School of Information Management, Jiangxi University of Finance and Economics, Nanchang* 330013)

[2)](*School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University, Nanchang* 330038)

[3)](*Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang* 330013)

**Abstract** Automatic question answer (Q&A) is a research hotspot in the field of artificial intelligence and natural language processing. It was originally designed to meet the needs of people to obtain information quickly and accurately. With the development of technology, most of the existing automatic Q&A models have no domain restrictions and can receive text and voice input. The retrieval-based method is an important technical route of the automatic Q&A, it has achieved fruitful results in recent years, but the related earlier reviews do not include new research results, or focus on a single field, which do not summarize and analyze the entire retrieval-based Q&A system. To provide researchers a quick overview of automatic Q&A, this paper makes a full survey on the classification, technical methods, data sets and evaluation indicators of automatic Q&A. First, we introduce the classification methods and several typical types of automatic Q&A, and then summarize the characteristics of different types of automatic Q&A and common technical methods. Secondly, taking the retrieval-based Q&A model as the main research object,

收稿日期:2019-12-06;在线发布日期:2020-12-25.本课题得到国家自然科学基金项目(61762042,61972184,62076112)资助.**赵 芸**,博士研究生,主要研究方向为社会媒体处理、自然语言处理.E-mail:zhaoyun_hust@163.com.**刘德喜**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为社会媒体处理、信息检索、自然语言处理.E-mail:dexi.liu@163.com.**万常选**,博士,教授,主要研究领域为 Web 数据管理、情感计算.**刘喜平**,博士,教授,主要研究领域为 Web 数据管理、文本挖掘.**廖国琼**,博士,教授,主要研究领域为社会网络挖掘、物联网数据管理.

we analyze three main methods, including rule-based methods, the methods based on statistical or machine learning and the methods based on deep learning. Then we focus on analyzing the characteristics and existing problems of these methods. For the rule-based method, the core idea is to understand the semantic information by using manual rules. These methods have the advantages of high accuracy and strong interpretability. But the biggest issue in the rule-based methods is that manual rules have limited coverage and are usually very costly. For the methods based on statistical or machine learning, the most important point of them is the representation of text. These methods could be self-optimized, so they have good scalability. However, they still require much manual effort in the design of the text representation, and the quality of the text representation directly affects the performance. In recent years, the methods based on deep learning are widely used in automatic Q&A, which greatly promotes the development of automatic Q&A. They can learn the complex text representation, relying on the nonlinear transformation. We also introduce various approaches for fusing external feature into the methods based on deep learning, including concatenation, attention mechanism, multi-channel mechanism, and multi-layer mechanism, etc., thus have greatly improved the performance of automatic Q&A. In addition, we successively introduce the evaluation methods and data sets of retrieval-based Q&A. Finally, we point out some issues commonly existing in retrieval-based Q&A, and discuss the possible future developing trends and challenges of retrieval-based Q&A.

**Keywords**　retrieval-based Q&A; feature representation; deep learning; neural networks; evaluation method; data set



## 1　引　言

自动问答让计算机能用准确、简洁的自然语言回答用户用自然语言提出的问题.其研究兴起的主要原因是人们对快速、准确地获取信息的需求.自动问答是目前人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向.

早期的问答模型主要基于知识库,例如问答式专家系统、受限语言的数据库查询系统等.Green 等人[1]设计了一个计算机程序 Baseball,用于回答关于棒球的问题,这可以说是最早的自动问答系统;LUNAR[2]用于回答关于月球岩石和土壤成分化学数据相关的问题;SHRDLU[3]只能回答和响应关于积木移动的问题等.这些系统在当时都取得了不错的性能,但是它们大多是受限的.一是语言受限,即只能采取少数几种特定的语言模式,一旦使用比较随意的语言,系统的性能将会大大下降;二是知识领域受限,这些系统一般只能回答某个特定领域的问题.

由于早期的研究积累,自动问答已经初具雏形,之后对于自动问答的研究主要集中于开放领域.为了促进非限定域自动问答的发展,信息检索评测组织(Text REtrieval Conferenee, TREC)自 1999 年开始,设立了非限定域问答的评测任务,是 TREC 中历时最长的评测任务之一.

尽管自动问答近年来取得了丰硕的成果,但对这些成果进行总结分析的综述类文献大都是较早的.而随着自动问答的发展,其研究和应用不断扩展,尤其是信息查找类问答、阅读理解和闲聊式对话均是近年研究的热点.研究者对这些工作进行总结分析时,大都聚焦于某一个单独领域,例如基于知识库的事实类问答[4]、CQA[5-7](Community Question Answering)、Chatbot[8-10](闲聊机器人)等,对整个自动问答体系的分类总结,以及对检索式自动问答模型的详细分析和对比工作并不多见.

虽然自动问答模型类型众多,但采用的方法大多基于规则、基于统计或机器学习以及基于神经网络,因此我们对自动问答的分类、技术方法、数据集和评价指标进行详细梳理,结合最新文献,对检索式自动问答模型的具体技术方法、评价指标以及数据集进行较全面地归纳和总结.

本文第 1 节介绍背景;第 2 节主要介绍自动问

答的分类以及典型类型，阐述各类自动问答的特点以及技术要求等；第 3 节详细介绍检索式自动问答模型的三类方法，分析各类方法的特点与难点，针对不同类别的方法，总结归纳现有的改进技术；第 4 节介绍检索式自动问答常用的客观评价指标以及数据集；第 5 节阐述检索式自动问答现存的挑战和对未来发展趋势的思考.

## 2 自动问答的分类及应用

自动问答依据不同的维度或角度可以有不同的分类结果，如图 1 所示.

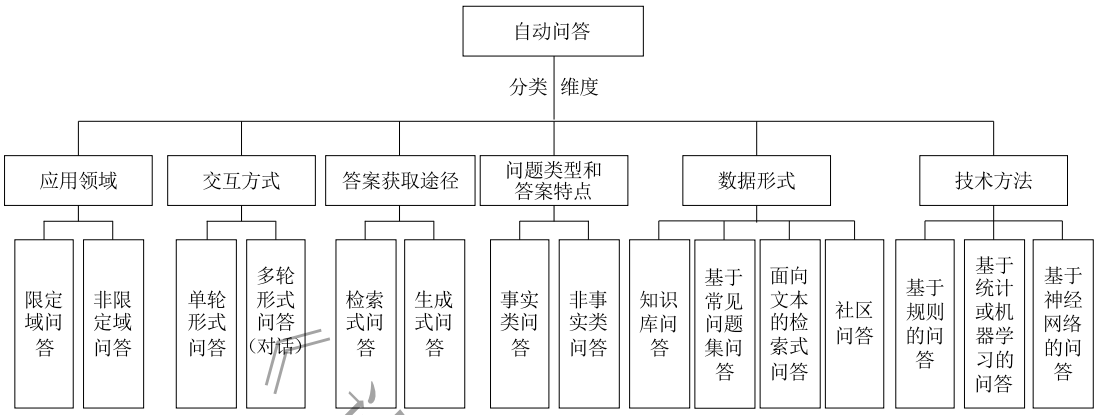


图 1 不同维度下自动问答的分类

按照应用领域或问题与答案涉及的领域是否受限，自动问答可以分为限定域自动问答和非限定域自动问答. 限定域自动问答只限于应用在某个领域、范围或任务上，或者问题与答案只涉及某个特定领域，例如只限定于医学或企业的某项业务等；非限定域自动问答则无领域限制，或者不针对某一特定的范围或任务.

按照交互方式，自动问答可以分为单轮形式问答和多轮形式问答（对话）. 单轮形式的自动问答只包含一个问题和相应答案（回复）；而多轮形式的自动问答则包含多个问题和多个答案（回复），问题之间、答案之间以及问题和答案之间大都存在某种关联，或遵从一定的逻辑，形成多轮交互.

按照答案的获取途径，自动问答可以分为检索式自动问答和生成式自动问答. 检索式自动问答是从大量的语料数据中，根据用户输入的问题，寻找最为合适的内容作为回复. 从语料中选出的最终答案可以是一段文字、一句话或者一个词. 而生成式自动问答是根据当前查询的问题，由模型自动生成相关答案进行回复.

按照问题类型和答案特点，自动问答可以分为事实类自动问答和非事实类自动问答. 事实类问答一般针对某一个事实提问，其答案具有唯一确定性，例如 When、Where、Who 等类型的问题；而非事实类问答的答案不具备唯一性，如 How、Why 等类型的问题，以及以闲聊为目的的“问题”.

按照数据形式，自动问答可以分为知识库问答、基于常见问题集问答、面向文本的检索式问答以及社区问答. 知识库问答使用的知识库通常是由知识构成的结构化数据；基于常见问题集问答的数据集由问答对构成，具有数据质量高、答案准确等特点，但数据规模不大；面向文本的检索式问答的数据通常是通过搜索引擎获得的各种文本内容；而社区问答的数据集则是通过搜集各种问答社区（例如雅虎问答、百度知道等）的帖子构建而成的.

按照采用的技术方法，自动问答可以分为基于规则方法的问答、基于统计或机器学习方法的问答和基于神经网络深度学习方法的问答. 基于规则的方法是根据人们预先制订好的规则检索或生成答案；基于统计或机器学习的方法首先根据知识构建特征，然后基于训练数据训练模型，最后利用模型检索候选答案，或对候选答案进行打分、排序，或直接生成候选答案；基于神经网络的深度学习方法是一种表示学习方法，它较少由人工参与特征设计，而是由机器自动发现检索、打分、排序或生成所需要的特征表示.

尽管图 1 从不同维度对自动问答进行了分类，但事实是，各种分类结果之间存在相互交融的现象.

例如限定域、非限定域问答中既可以包含事实类、非事实类问题，也可以根据需要采用单轮或多轮的问答形式；面向文本的检索式问答以及社区问答既可以包含事实类的问题，也可以包含非事实类问题；对话既可以回答限定域的问题，也可以回答非限定域的问题，既可以回答事实类问题，也可以回答非事实类问题.

自动问答的分类维度丰富多样，而技术方法的选择与自动问答的类别有很强的关联性. 下面以“应用领域”这一维度以及当前学界和工业界的热点问题“对话”为例，对自动问答的任务定义、模型要求、特点、技术要点等问题进行阐述，也进一步说明自动问答的多样性和复杂性.

### 2.1　限定域自动问答

早期的研究为了使得普通用户能从庞大的数据库中找寻自己所需的内容，同时避免专业、复杂的计算机操作，推出了限定域的自动问答. 例如，1973 年，为了地质学家能够方便地访问、比较和评价采集的月球岩石和土壤成分的化学分析数据，Woods[2] 设计了一个原型系统 LUNAR，使得人与计算机的交流可以直接使用英语，而无需采用专业的计算机查询语言. Hendrix 等人[11] 提出了一种使用自然语言访问大型数据库的方法. 因为采用自然语言进行提问，是早期自动问答的基本要求，自然语言理解(Natural Language Understanding)成为自动问答一个研究热点. ELIZA[12] 也是一个限定域自动问答系统，开发者希望它能帮助许多患有心理疾病的病人，让它富有同情心，能像知心朋友一样给人安慰. 许多患有心理疾病的病人与它聊过后，对其的信任度超过了真正的人类心理医生. SABORI[13] 则是一个用于辅助心理健康治疗的非事实类自动问答系统，它首先询问用户一个和行为建议有关的问题，然后根据用户的回答，提供新的行为建议.

随着计算机的普及以及网络的快速发展，限定域的自动问答常被用于完成特定任务[14-16]. 例如公交助手、机票或火车票预订、寻找餐馆、自动播放音乐、辅助心理治疗等. 这时候的自动问答更像一个助手，不仅能回答问题，还能触发动作，并且利用自动语音识别(Automatic Speech Recognition，ASR)技术，使得系统可以接收语音提问.

早期限定域自动问答由于语言和领域的限制，因此基于规则的方法被广泛采用. 随着深度学习技术的日益成熟，近年来，其被广泛应用于面向任务型的限定域问答，取得了良好的成果.

### 2.2　非限定域自动问答

依据问题类型和答案特点，非限定域自动问答又可以进一步分为事实类自动问答和非事实类自动问答，两者拥有不同的特点和技术要求. 非限定域事实类自动问答需要回答的问题一般具有确定答案，而非限定域非事实类自动问答的回答则可能仅仅是一种观点、解释或者评论，其注重的是回答是否为提问者提供了必要的信息，如图 2 所示.

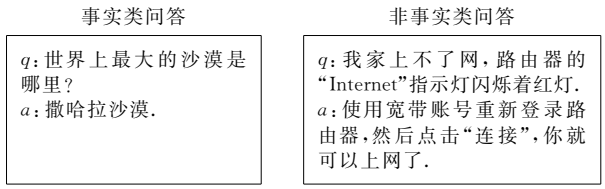


图 2　事实类和非事实类问答示例

(1) 事实类自动问答

非限定域事实类自动问答由于数据形式的不同，可以进一步细分为基于知识库的问答和基于常见问题集的问答.

基于知识库的问答和基于常见问题集的问答都属于一般性的知识问答模型，前者是通过学习各类知识(Freebase、Wikipedia、雅虎问答、百度百科等)，从而帮助模型挖掘出“问题”和相应的“答案”；后者则是从已有的“问答对”中找出与用户问题相匹配的“问题”，然后将“答案”返回给用户. 它们的问题一般以“wh”开头的单词，即 what、who、when、where 等，而回答一般是实体、地理位置、时间等. 事实类问答通常需要对问题进行分类，明确问题类型，不同类型的问题后续采用的回答策略是不同的；其次明确问题的主题，即问题的核心，例如“世界上最大的沙漠是哪里？”，它就和地理、沙漠有关；最后完成回答的抽取.

阅读理解是自然语言处理领域的热点问题，它也可以看作是特殊的事实类自动问答. 阅读理解是给定问题与相关的文章，要求机器能理解文章内容，并从给定的文章或段落中找寻问题的答案. 尽管阅读理解与事实类问答存在一定的差异，但两者还是具有很大的共性：首先，阅读理解的问题大多属于事实类问题；其次，阅读理解的答案具有唯一确定性，且来源于自然语言文本，而事实类的自动问答的答案来源也可以是自然语言文本；最后，阅读理解的评价指标与事实类自动问答基本一致.

非限定域事实类自动问答的回答按照精度可以分为词、词组、短语和句子.常用的方法有基于规则的方法、基于机器学习的语义解析方法和基于深度学习的方法.其答案的评估准则主要有精确率、召回率等.

(2)非事实类自动问答

非事实类自动问答与事实类自动问答最大的不同是,其答案不具有确定性.非事实类自动问答的问题可以是观点类、方法型,其更在乎回答的相关度、合理度,与事实类自动问答相比,其回答通常较长.

非事实类自动问答根据数据形式的不同,可以分为基于 Web 的非事实类问答和非事实类社区问答等,其中基于 Web 的非事实类问答的问答数据主要依靠搜索引擎获得的 Web 文档,而非事实类社区问答的数据主要来自百度知道、知乎、雅虎问答等问答社区.虽然社区问答中存在事实类的提问,但其问题的很大比例是有关"观点"的提问和信息的查找.

非事实类的问答模型主要采用基于统计或机器学习的方法和基于深度学习的方法,也有少量采用基于规则的方法.

### 2.3 对话

对话是一种多轮互动的问答形式,与其他自动问答类型相比,对话弱化了问答的概念,注重对话功能,因此研究者通常将对话中的问答对定义为 message-response,并且 message 和 response 的长度都偏短.

Chatbot 是对话的一种应用,它使得人与机器的交流变得更方便,被广泛应用于多个领域.它可以是面向任务的[10],像是面向多用途的代理人,例如 Siri①、Amazon Alexa②、Microsoft Cortana③ 和 Google Assistance④ 等,也可以是非面向任务的,即注重闲聊属性的闲聊机器人[10],例如带有情绪识别的微软 XiaoIce[17].

由于闲聊式对话模型主要的任务就是聊天,因此除了保证回复与信息的相关性和合理性之外,还需要注意对话的可持续性.已有学者或企业在现有的闲聊式对话模型中引入情绪处理,使得该类问答模型变得更有"温度".微软推出了的小冰就是一款具有同理心的社交聊天机器人,能够识别用户的情感需求,然后通过给予鼓励或其他情感信息,提供一种引人入胜的人际交流,从而在交流中抓住人的注意力.Zhang 等人[18]、Colombo 等人[19]均在问答模型中引入情绪分类器,然后根据识别的情绪类别生成相应回答.这些策略使得机器生成的回复更拟人化,从而增强聊天的可持续性.

各典型自动问答类型的介绍及特点如表 1 所示.由于自动问答是一个很宽大的研究范围,根据不同的标准会有不同的分类方法,并且在具体应用时会发生一定的重叠.

**表 1 自动问答典型类型的介绍及特点**

| 名称 | | 定义 | 模型要求 | 特点 | 技术要点 | 使用的方法 |
|---|---|---|---|---|---|---|
| 限定域自动问答 | | 只能处理某个领域或者某个特定范围的问题,主要是面向特定任务. | 能很好地完成特定任务 | 问题来自某个特定范围 | 自然语言理解、规则设定 | 基于规则的方法;基于统计或机器学习的方法;基于神经网络的深度学习方法 |
| 非限定域自动问答 | 事实类问答 | 面向事实类的问题,该类问题一般具有唯一的确定答案. | 能准确抽取或生成回答 | 答案简短,一般是实体、地理位置、时间等 | 问题分类、问题主题识别、答案抽取或生成 | |
| | 非事实类问答 | 面向非事实类的问题,不具有确定答案. | 回答必须与问题具有高相关性 | 答案与问题相比,通常较长 | 相关性的判断 | |
| 对话 | | 既可以针对限定域,也可以面向非限定域,还可以是不以信息获取或特定任务为目的的闲聊 | 除具有限定域自动问答和非限定域自动问答的要求外,在以闲聊为目的时,回答不仅与问题具有高相关性,还要使得话题具有可持续性. | 除具有限定域自动问答和非限定域自动问答的特点外,在以闲聊为目的时,信息与回复均较短 | 除综合应用限定域自动问答和非限定域自动问答的技术外,在以闲聊为目的时,要提升回复的拟人度、可持续性等 | |

尽管依据不同的标准,问答模型有不同的分类,但是它们之间又有某种的联系,例如检索式问答模型与生成式问答模型均使用了三类技术方法,但是由于两者自身的特点,三类技术方法在两种问答模型中的使用存在差异,受篇幅限制,也为了更好地整理相关研究,本文以检索式问答模型为主,详细介绍检索式问答模型主要采用的方法以及各类方法的特点、难点、相关评测、评价指标以及数据集.图 3 中展示的是不同的方法在检索式自动问答模型中的应用.

① https://www.apple.com/es/siri/
② https://developer.amazon.com/alexa
③ https://www.microsoft.com/en-au/windows/cortana
④ https://assistant.google.com/

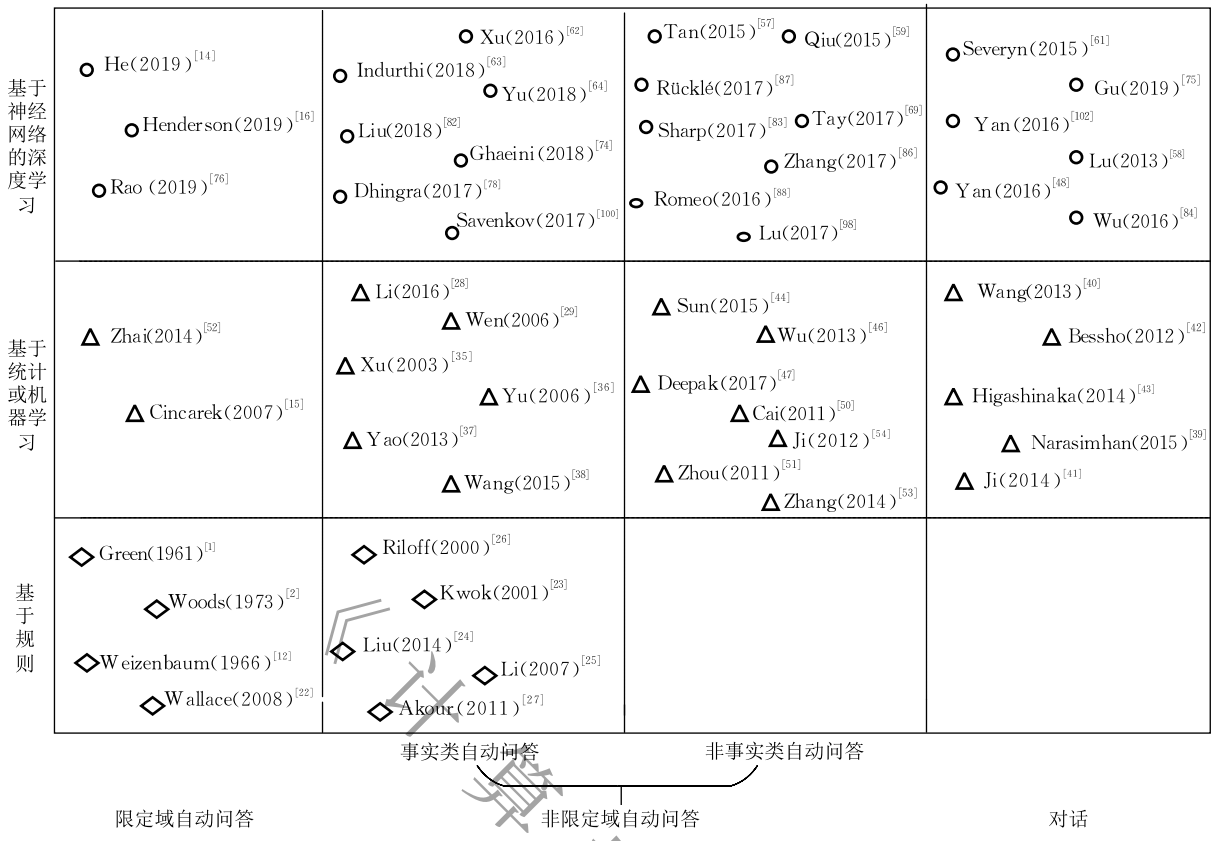


图 3　检索式自动问答模型的分类示意图

## 3　检索式自动问答模型

虽然自动问答的任务不同，但构建为检索式自动问答的方法是类似的，即根据用户给定的问题 $q$，从已有的数据集、知识库、网页或者文档中检索与 $q$ 相关的回答，并返回最为合理的回答 $r^*$. 检索模式一般有三种，第一种是检索与 $q$ 最为相似的历史问题 $p$，将历史问题 $p$ 的回答作为 $r^*$，如式(1)所示；第二种是直接计算数据集、知识库、网页或者文档中的候选回答与 $q$ 的语义相似性，检索出得分最高的回答，如式(2)所示；第三种则是综合前两种检索模式，检索时既考虑问题相似性，又考虑回答的相似，返回综合得分高的回答，如式(3)所示.

$$r^*=answer(\arg\max_p score(q,p)) \tag{1}$$

$$r^*=\arg\max_r score(q,r) \tag{2}$$

$$r^*=\arg\max_{(p,r)} score(q,(p,r)) \tag{3}$$

非事实类社区自动问答和对话，其数据集通常来源于问答社区(如雅虎问答、百度知道等)或社交媒体(如 Twitter、新浪微博等)，因此，自动问答的检索对象通常是由主题帖(视为历史问题 $p$)和其回复帖或评论(视为候选回答 $r$)构成的候选问答对；而基于 Web 的非事实类自动问答，检索对象主要是 Web 上的页面或文档，以及更细粒度的段落或句子.

基于常见问题集的自动问答，其检索对象通常为已有数据集中的问答对；而基于知识库的事实类自动问答，为了更好地处理语义结构复杂的问题，其检索对象除了已有数据集或网络中的候选文档外，还涉及知识库(Freebase、DBpedia 等)等结构化或半结构化数据.

很多限定域自动问答的检索对象仅来源于领域知识库，若没有检索到候选问句或无法匹配到预先设定的模式，就无法对给定问题进行回答. 为了解决这一缺点，有研究者将限定域自动问答的检索对象进行扩充，不仅检索本地的领域知识库，还对在线问答社区等更广泛的信息源进行检索.

阅读理解可以看成是将检索范围限定在指定文档 $d$ 中的特殊检索形式. 阅读理解主要有填空型阅读理解、子句抽取型阅读理解、判断型阅读理解等. 对于填空型阅读理解，其检索对象通常是文档 $d$ 中的名词实体或动词；对于子句抽取型阅读理解，其检索对象是文档 $d$ 中的子句；对于判断型阅读理解，可视为以待判断的句子为查询 $q$，在文档 $d$ 中检索是否存在语义相同的内容.

检索式自动问答模型主要采用的方法可分为三类，各类方法及其特点和难点如表 2 所示.

表 2 检索式自动问答模型各类方法的特点与难点

| 模型 | 特点 | 难点 |
|---|---|---|
| 基于规则 | 模型准确率高,可解释性强,但覆盖率有限,且需要人工制订规则,不易扩展. | 规则的制订以及规则的覆盖率 |
| 基于统计或机器学习 | 模型对各类数据的适用性较强,且能自我优化,模型扩展性好. 但需要人工设计特征表示器. | 特征表示器的设计 |
| 基于神经网络的深度学习 | 能自主学习复杂函数来构建预测或分类模型. 但是模型参数庞大,可解释性差,训练速度偏慢. | 神经网络的选取与搭建、外部特征的有效融合 |

基于规则的方法主要用于限定域自动问答,这些规则大多是由人根据语法语义制订的,系统再根据这些规则为当前查询的问题寻找正确的回答. 基于规则的方法最大的缺点是制订规则不仅工作量大,而且需要深入了解问题和候选回答的语法语义甚至世界知识,同时模型的扩展性不好,一旦该问答模型需要变换语种或任务,则需要重新制订规则.

随着互联网的快速发展,网络文本数据快速增长,统计方法对检索式问答模型的重要性也在增加. 基于统计或机器学习的方法主要是基于数据预测回答. 这类方法能够很好地解决数据的异构性,并且不受结构化查询语言的约束. 基于统计或机器学习的方法可以从训练数据中构建知识库,然后使用该知识库来回答新问题. 此外,随着机器学习模型的进步,使得使用该类方法的问答模型能够随着时间的推移进行自我优化,大大增强了系统的可扩展性,具有良好的经济效益.

基于神经网络的深度学习方法被引入问答模型,为问答模型提供了更大的提升空间. 传统的基于机器学习的方法需要设计特征提取器,用于文本的特征表示,文本特征表示的质量对问答模型最终的效果具有极大的影响,而特征提取器的设计需要领域知识的介入. 基于神经网络的深度学习方法是一种表示学习方法,它允许向机器输入原始数据,并自动发现预测或分类所需的表示方法. 它可以有多个层次,每层由许多简单但非线性的模块组成,每个模块将一个层次的表示(从原始输入开始)转换为一个更高、更抽象的层次表示. 由于是非线性变换,因此神经网络可以学习非常复杂的函数表示.

### 3.1 基于规则的方法

基于规则的方法[2,12,20-22]在早期的自动问答模型中被广泛采用. 在早期,研究者常常通过制定规则来理解语义信息,例如 Woods[2]在 LUNAR 系统中使用一种基于规则的通用语言解释器,用来“理解”用户输入的自然语言信息. 在此后的研究中,研究者还使用规则来选取或生成答案. 在检索式自动问答模型中,基于规则的方法主要应用于事实类自动问答模型.

在事实类自动问答模型中,一般需要先对问题进行分类,然后根据不同类型问题的特点,制定不同的答案抽取规则. 例如 Kwok 等人[23]、Liu 等人[24]制定规则对问题进行分类,其中 Liu 等人[24]还使用自然语言工具对问题进行语法分析,然后提取主语、宾语作为问题的关键词;Li 等人[25]提出一种特殊的基于语义模式的规则对问题进行分类;Riloff 等人[26]提出了一个基于规则的阅读理解模型 Quarc,基于语法规则构建决策树,根据问题类型分别定义答案抽取路径的规则集,使得问题“Tajmahal 在哪里”和问题“谁在一美元钞票上”的答案提取过程所采用的路径是不同的;Akour 等人[27]同样使用一系列的规则构建阅读理解问答模型,该模型主要针对阿拉伯语. 为了提高答案匹配的准确度,词法分析、词性标注、词干提取和命名实体识别等都被引入这类模型中.

由于限定域的问答模型一般只能处理某个领域或者某个特定范围的问题,方便规则的制定,因此基于规则的方法在这类问答模型中表现较好. 但是人工制定规则不仅工作量大,而且模型不具有普适性,扩展性不佳. 尤其对于非限定域的自动问答,因为其缺乏领域限制,人工制定规则难度加大,即使制定出规则,也会因为规则的覆盖率有限,导致模型效果不佳.

### 3.2 基于统计或机器学习的方法

由于互联网的大爆发,获得大量训练数据成为可能,基于统计或机器学习的方法变得可行.

#### 3.2.1 基于统计或机器学习方法的事实类问答模型

事实类问答模型通常需要利用机器学习的方法解决问题分类和答案抽取两个模块.

(1) 问题分类

问题的类别决定了答案的类别,它可以提高检索正确率以及指导答案的抽取. 目前被广泛接受的英文事实类问题的分类是 Li 等人[28]提出的,他们将问题分为 6 大类 50 小类;文勖等人[29]根据中文的特点,将中文事实类问答的问题分为 7 大类 60 小类.

根据问题的常用类别，研究者通常使用支持向量机(SVM)、最大熵或随机森林等机器学习算法训练问题分类器[30-33]，使用的特征有词汇特征(词袋、词频、$n$-gram 等)、词性特征、句法分析特征、语义分类特征(WordNet、HowNet 等)以及命名实体的类名等.紧接着，根据问题的分类，借助依存句法分析等语法分析工具[24,34]获取问题的句法关系，结合语义分类特征(WordNet、HowNet 等)提取关键词，为信息检索提供搜索词.

(2)答案抽取

答案抽取包含候选答案抽取和候选答案排序两个步骤.首先，根据问题分类，利用自然语言处理工具抽取相应的词/词组，例如 Xu 等人[35]提出了基于命名实体识别的答案抽取算法.然后，通过排序算法对候选答案进行排序，例如余正涛等人[36]使用潜在语义分析，采用向量空间模型计算相似度，根据相似度的得分进行排序.Yao 等人[37]将答案抽取视为序列标注问题，利用条件随机场(CRF)在候选句子中找到答案所在，其所使用的特征除了词性特征、依存句法特征和命名实体特征外，还借助从 Tree Edit Distance(TED)模型中提取的特征，用于对齐回答句树与问题树，从而提高答案抽取的准确率.Wang 等人[38]发现依赖语法、框架语义、词嵌入和共指关系的使用，对于理解“谁对谁做了什么”具有重要作用.Narasimhan 等人[39]则将目光转向句子信息，通过构建模型识别相关句子，构建它们之间的关系，从而影响答案的预测.

### 3.2.2　基于统计或机器学习方法的非事实类问答模型或闲聊对话

与事实类问答模型不同，非事实类问答模型或闲聊对话更侧重相关性的判断.对于检索式问答模型，可以利用机器学习模型对候选的回答进行甄别排序[40-46]，找出最为相关的回答.而该类技术的关键是要设计特征提取器，用于原始数据的特征表示.

由于该类自动问答模型更像是信息检索模型的一种高级形式，因此有学者尝试将经典的搜索算法引入问答模型，例如 BM25 模型.这类搜索算法大多是通过比较问题与回答中出现的相同词汇或同义词汇来判断两者的相关性.由于语义不相同的词汇也可能反映问题与回答之间的相关性，因此，语言鸿沟(Lexical Gap)[47]问题是这类方法的主要挑战，例如“磁盘已满(disk full)”的问题与含有其解决办法“格式化(format)”的回答是具有高相关性的.

为了解决语言鸿沟的问题，从数据中抽取更丰富的特征来改进模型的方法被证明是有效的，例如基于翻译模型提取映射特征[48-51]、基于主题模型提取主题特征[48,50,52-54]以及基于潜在语义分析提取语义特征[40-41,44,47-48]等.

(1)基于翻译模型提取映射特征

翻译是用不同的语言表达相同的内容，因此，可以利用翻译模型学习问题和回答中词语或短语之间的相关性.尽管问答对在语义上不符合平行语料库的标准，但是使用翻译模型来处理它们可以提高检索精度.对于前面例子中“disk”与“format”之间的语言鸿沟问题，$p$(“format”|“disk”)的高概率可以提高包含单词“format”的候选回答的分数，以响应涉及磁盘问题的用户查询.也有研究者将其用于更粗粒度的文本中，例如短语级别[48,51].

但是将机器翻译模型移植到问答领域，会遇到一个新的挑战，“词对齐”问题.因为问题和回答的文本在语义上并不等价，因此存在大量无法对齐的词语.为了解决“词对齐”问题，Ritter 等人[55]提出一种基于统计机器翻译的自动问答方法.该方法通过设置限制条件和特征来解决词对齐中的词汇重叠问题，并且不再依赖经典的对齐方式来提取短语对，而是在“问题-回答”的并行文本中考虑所有可能的短语对，并采用基于关联的过滤器进行过滤.

(2)基于主题模型提取主题特征

利用主题将表述同一主题的词语聚合在一起，用以解决语言鸿沟问题.例如通过主题分析，获悉词语“disk”与“format”隶属于同一主题，从而调整其相关性得分，提高模型效果.

也有学者基于相似问题拥有相同或相似答案的假设来检索回答，因此利用主题模型分析问题与问题之间的相似性，例如 Cai 等人[50]运用主题模型 LDA[56]获得当前问题 $q$ 与历史问题 $Q$ 的主题分布，然后计算两者之间的相似度，$P_{TM}(q|Q)$ 的高概率会提高与 $q$ 具有相近主题的历史问题 $Q$ 的得分.

(3)基于潜在语义分析提取潜在语义特征

从词语的潜在语义特征获得词语间的潜在联系.Ji 等人[41]使用 Wu 等人[46]提出的映射方法，将问题 $q$ 与候选回答 $r$ 的原始向量表示分别映射到低维特征空间 $L_q$ 和 $L_r$ 上，问题与回答之间的相关性打分用向量内积表示，如式(4).

$$LatentMatch(q,r)=q^{\mathrm{T}}L_qL_r^{\mathrm{T}}r \tag{4}$$

实验表明潜在语义分析能获得有意义的语义信息.例如，问题 $q$ 中出现词语 Italy 能更好地与

“Sicily”、“Mediterranean sea”和“travel”等词匹配，从而提高问题与回答的匹配精度.

(4) 其他特征

Savenkov 等人[45]借助知识库 Freebase，通过提取知识库中的信息来提高事实类问答模型的准确性. Yan 等人[48]使用卷积神经网络(CNN)提取有助于判定问题 $q$ 与候选回答 $r$ 之间存在因果关系或对话关系的特征. Wang 等人[40]计算问题 $q$ 与候选回答 $r$ 中相同字符串的最长长度. 这些特征在问答模型中同样具有普遍性，有助于提高问答模型的精度，它们可以单独使用，也可以混合使用.

基于统计或机器学习的方法对训练数据集的容量、标注质量和特征选择有极大的依赖性，这类方法多用于检索式问答模型.

### 3.3 基于神经网络的深度学习方法

近年来，深度学习模型在语义分析、机器翻译、文本摘要等多种自然语言处理任务上取得了显著的成绩[57]. 深度学习相比其它机器学习的优势在于，它能通过神经网络自动学习数据的特征表示.

对于检索式问答模型，可视为对候选回答进行分类预测，即将查询问题 $q$ 与候选回答 $r$ 送入神经网络模型中，对所有候选回答进行分类或者排序.

基于神经网络的检索式问答模型[58]采用的是监督学习的方法，神经网络的任务就是根据每一个问题 $q$ 对所有的候选回答进行分类或者排序. 如果任务被定义为分类问题，则需要学习一个函数，它为每一个问答对输出一个二进制数，“0”表示负类(即不是该问题的回答)，“1”表示正类(即是该问题的正确回答). 若是对所有的候选回答进行排序，则需要学习排序函数，它为每一个问答对输出能表示两者相关性的分数，分数越高，相关性越大. 这些方法大都基于典型的神经网络结构，例如卷积神经网络(CNN)[59-67]和循环式神经网络(RNN、LSTM、GRU等)[57,68-82]等.

该类模型的训练方式主要有两种，pairwise 方式和 pointwise 方式. pairwise 方式每次送入一个问题($q$)、一个正向回答(positive reply)以及一个负向回答(negative reply)，其损失函数是使得正向回答与问题的相关性得分要尽可能地大于负向回答与问题的相关性得分. pointwise 方式则是采取分类的思想，每次训练时送入一个问题($q$)、一个回答以及一个标签，标签表明该回答对于当前问题是正确的还是错误的，其损失函数多采用交叉熵. 通常，pairwise 方式比 pointwise 方式的训练时间更长.

在基于深度学习构建问答模型时，无论使用何种神经网络，都需要与其他特性(如词汇重复特征和 IDF 权重等)相结合才能发挥更好的作用. 将特征与神经网络融合的方法有多种，常用的有，直接拼接(Concatenate)[61,69,83]、注意力机制(Attention)[84-96]、多通道(Multiple Channel)[62,97-103]以及堆叠[81,104]等.

(1) 直接拼接

将人工提取的外部特征与神经网络提取的特征进行向量的拼接，对神经网络提取的特征进行补充，然后再进行分类或排序. 例如 Severyn 等人[61]提取 *query* 和 *document* 之间去除停用词之后的相同单词的个数以及相同单词的 IDF 权重之和作为外部特征 $x_{feat}$，将其与 CNN 提取的特征进行直接拼接，组成一维新的特征，然后送入全连接层进行分类预测.

(2) 注意力机制

注意力机制与直接拼接法不同，它不是直接补充神经网络提取的特征，而是希望模型在提取特征的过程中，重点关注某些重要信息，例如问题 $q$ 或候选回答 $r$ 中的关键词或关键短语，提高这些重要信息在原特征中的权重，从而提高问题和回答的匹配精度.

Chen 等人[89]利用问题 $q$ 与候选回答 $r$ 的词向量矩阵计算注意力矩阵 $\boldsymbol{A}\in R^{|q|\times|r|}$，其元素 $A_{i,j}$ 的计算如式(5)所示，用于衡量问题 $q$ 中第 $i$ 个词与回答 $r$ 中第 $j$ 个词之间的语义距离.

$$A_{i,j}=\frac{1}{1+\|\boldsymbol{F}_q[i,:]-\boldsymbol{F}_r[j,:]\|} \tag{5}$$

其中 $\boldsymbol{F}_q$(或 $\boldsymbol{F}_r$)表示问题 $q$(或候选回答 $r$)的词向量矩阵，$\|\cdot\|$表示欧几里得距离.

Du 等人[105]利用 Bi-LSTM 在给定的文本中找出问题的回答. 为了更精准地定位回答，Du 等人利用双线性函数来计算注意力权重，以此指导模型关注重要信息.

Kundu 等人[90]则利用问题与候选回答的词相似度矩阵，通过行或列的归一化处理获得问题(或候选回答)中每个词对候选回答(或问题)的相关性，以此作为问题(或候选回答)的注意力权重.

Gu 等人[106]在应对阅读理解问题时，利用时态卷积网络(Temporal Convolutional Network，TCN)捕获问题间的主题转移特征，然后再利用注意力机制将捕获的特征融入主模型，从而提高模型的效果.

虽然注意力权重的计算方法各不相同，但目的都是突出有助于提高系统性能的重要信息.

(3) 多通道机制

多通道机制则是增加神经网络的宽度，对多源数据进行处理.

对于事实类自动问答模型，常常借助知识库来提升答案的准确性[62,100]. Savenkov 等人[100]除了将当前问题和候选语句送入神经网络外，还将与候选语句中名词实体有关的结构化的知识信息作为输入送入模型，以辅助答案的抽取和判断.

延续对话的主题，可以使得对话模型获得良好的逻辑性和连贯性，为此，研究者经常在输入端增加上下文内容. 图 4 所示的例子显示，如果仅仅针对当前问题 $q_0$ 检索最佳回答，则 $A_2$ 比 $A_1$ 具有更高的语义相关性，但是结合上下文信息($c_1$ 和 $c_2$)，发现综合整轮对话的逻辑性和连贯性，$A_1$ 比 $A_2$ 更适合作为 $q_0$ 的回答.

$c_1$：今天天气好差，天气转凉的跨度好大！
$c_2$：是不是很恐怖，今天已经结冰了.
$q_0$：那你最近要怎么度过？
$A_1$：那就冬眠吧.
$A_2$：我一会去上班.

图 4　上下文对问答影响的示例[84]

Yan 等人[102]利用多通道融合上下文特征，其模型拥有 4 个通道，分别整合了当前查询 $q_0$(Original query)、历史查询 $p$(Antecedent post)、利用上下文信息构建的扩展查询 $q_i$(Reformulated query)、以及候选回答 $r$(Candidate reply). 其中上下文信息取当前对话进程中除当前查询以外的历史问答对，历史查询 $p$ 则是当前待考察的候选回答 $r$ 在数据集问答对中的真实历史查询. 分别计算候选回答 $r$ 与扩展查询 $q_i$ 的相关性 $f(q_i,r)$、扩展查询 $q_i$ 与当前查询 $q_0$ 之间的相关性 $h(q_i,q_0)$、扩展查询 $q_i$ 与回答 $r$ 对应的历史查询 $p$ 之间的相关性 $g(q_i,p)$，形成最终的预测分数，如式(6)所示. 该模型不仅利用了历史查询，同时还整合了上下文信息，使得问答模型具有更高的检索精度和更强的对话逻辑性.

$$F(q_0,r)=\sum_{i=0}^{|Q|}(h(q_i,q_0)\sum_{p}(f(q_i,r)\cdot g(q_i,p))) \tag{6}$$

而 Wu 等人[84]不仅使用多通道，还引入了注意力机制. 其利用多通道学习上下文的特征，然后使用注意力机制找出上下文特征中有助于检索的关键词或短语. 上下文信息包含当前对话的聊天背景，结合聊天背景和当前问题检索出的回答，不仅具有良好的语义相关性，还具有较好的逻辑性和连贯性.

(4) 堆叠机制

通过使用相同模块或相似模块进行堆叠，增加神经网络层次，从而实现多轮推理，用以应对事实类问答或阅读理解的复杂推理问题.

Sukhbaatar 等人[107]首先提出了一个单层的端到端记忆网络，该网络拥有输入模块、输出模块和答案预测模块. 由于单层网络的推理能力有限，作者将单层的记忆网络进行堆叠，输入模块和输出模块不变，但是问题的语义表示与文档进行多次交互更新，在每一次交互过程中，不断理解问题，进行多轮推理.

Dhingra 等人[81]同样使用了堆叠机制，但是与 Sukhbaatar 等人不同的是，在堆叠过程中，问题的语义信息保持不变，在多层网络中使用门控注意力机制不断地过滤文档信息，进而增强模型对文档的理解能力.

Sordoni 等人[104]在堆叠过程中交替对问题和文档进行提炼，从而提升模型的推理能力.

(5) 其他改进方法

在非限定域的问答模型中，会出现一些容易引起歧义的问题以及稀疏性问题，即使添加了众多外部特征也很难得到有效的处理. 因此有学者利用多通道机制将用户的社交关系引入模型[98-99]，以提高模型的效果.

Google 推出的 BERT(Bidirectional Encoder Representations from Transformers)，是一个基于多层次的双向 Transformer 编码器的特征表示模型，与近年优秀的特征表示模型不同，BERT 模型在 NLP 任务中仅仅只需要添加输出层即可，不需要根据特定任务对 BERT 进行实质性的体系结构修改，大大增进了模型的可移植性. Devlin 等人[108]将 BERT 用于包含自动问答的 11 种自然语言处理任务，均取得了非常优异的结果.

阅读理解因为答案来自给定文本，在阅读理解任务常用的数据集中存在无法回答的问题，例如 SQuAD1.1、SQuAD2.0 和 NewsQA，其中 SQuAD2.0 包含 43 000 个无法回答的问题[109]. 虽然可以利用这些数据训练模型用以判断问题是否有答案，但是其更大的效用是被用来进行对抗攻击训练. 在阅读理解任务中，存在输入信息有变化，但模型预测结果仍不改变的情形，具体如图 5 所示，研究者称之为模型敏感性. 而对抗攻击训练可以有效提高模型的敏感性.

给定文本：附近的西班牙殖民地圣奥古斯丁袭击了**卡洛琳堡(Caroline)**，几乎杀死了所有保卫它的法国士兵. 西班牙人改名为**圣马特奥堡(San Mateo)**……
真实问题：西班牙人进攻后，**卡洛琳堡**更名为什么？
预测结果：San Mateo
相似问题：西班牙人进攻后，***罗伯特·奥本海默(Robert Oppenheimer)***更名为什么？
预测结果：San Mateo

图 5 阅读理解模型敏感性示例[109]

研究者往往会通过生成高质量的对抗示例，用以进行更为有效的对抗攻击训练. 例如 Ribeiro 等人[110]使用了一些简单的干扰，例如将真实问题中“Who is”替换为“Who's”；Jia 等人[111]和 Wang 等人[112]对真实问题在语义上实施小改动，从而生成用于对抗训练的问题；Ebrahimi 等人[113]在字母级别进行变动，生成对抗数据；Welbl 等人[109]则是通过替换真实问题中的命名实体来生成对抗示例.

# 4 自动问答评测、评价指标以及相关数据集

## 4.1 自动问答评测

目前国际上影响力较大的自动问答评测主要有 TREC、CLEF(Cross Language Evaluation Forum)和 NTCIR(NACSIS Test Collections for IR).

(1) TREC① 是文本检索领域最具权威的评测会议，它由美国国防部高等研究计划署与美国国家标准和技术局联合主办. 该评测会议从 1992 年开始，每年举办 1 次. 从 1999 年(TREC 8)开始引入问答系统的专项评测，2002 年(TREC 11)是该评测会议最后一次组织跨语言的问答系统评测.

(2) NTCIR② 是由日本国家科学咨询系统中心主办，建立的目的是为了促进基于日语的信息检索与自然语言处理研究，该会议从 2002 年开始，每 2 年举办一次. 该评测会议在一开始就有 Question Answering Challenge(QAC)任务，并于 NTCIR-5 开始加入繁体中文语料集.

(3) CLEF③ 是由欧盟资助，主要针对欧洲语言进行的信息检索评测会议. 该会议由 2003 年开始，每年 1 次.

中国科学院自动化研究所 2005 年初步建立了一个汉语的问答模型评测平台 EPCQA[114]. 其目的是为了完善以汉语为主的问答数据集，从而推动汉语问答模型的发展.

## 4.2 评价指标

检索式自动问答模型或问题分类的评价指标较为客观，评价方式也较为成熟，例如 *Precision*@*N*、*Recall*@*N*、*F*1@*N*、*MAP*、*MRR*、*NDCG*、*CWS* 和 *K*-Measure.

假设 $Q=\{q_1,q_2,\cdots,q_{|Q|}\}$ 是所有待查询的问题的集合，$|Q|$ 表示问题的数量，$\hat{r}$ 是模型根据问题 $q$ 返回的答案集合，$r$ 为相应的理想答案集合，$|\hat{r}|$ 和 $|r|$ 分别为模型返回结果集合以及理想答案集合的大小.

(1) *Precision*@*N*、*Recall*@*N* 和 *F*1@*N*

*Precision*@*N* 反映前 *N* 条返回结果的精确率，*Recall*@*N* 反映前 *N* 条返回结果的召回率，*F*1@*N* 是两者的调和平均值，如式(7). 其中 $\hat{r}_i$ 为 $\hat{r}$ 中排名第 $i$ 的返回结果，$I$ 为指示函数.

(2) *MAP*

*MAP*[115](Mean Average Precision)衡量模型返回的正确答案的数量和排序位置. 模型的正确答案越多且越靠前，*MAP* 的值越高，如式(8)所示. 其中 $\hat{r}_i$ 为 $\hat{r}$ 中排名第 $i$ 的返回结果，$I$ 为指示函数，$rank(r_j,\hat{r})$ 表示 $r_j$ 在 $\hat{r}$ 中的排名位置.

(3) *MRR*

*MRR*[116](Mean Reciprocal Rank)评价第一条正确答案在模型返回结果中的位置，排名越靠前越好. 它以相关文档在结果中的最好排序取倒数作为它的准确度，如式(9)所示. 其中 $r_i$ 为 $r$ 中排名第 $i$ 的理想答案，$rank(r_i,\hat{r})$ 表示 $r_i$ 在 $\hat{r}$ 中的排名位置.

(4) *NDCG*

*NDCG*[117](Normalized Discounted Cumulative Gain)，不仅考察返回结果中正确答案的数量、位置，还考察这些返回结果的信息量或者分值. 信息量或分值越高的答案返回的越多且越靠前，*NDCG* 的值越大，如式(10)所示. 其中 $rel_i$ 是第 $i$ 个返回结果的分值或信息量(通常由人工设置)，$IDCG(q)$ 是返回结果为理想结果(即全部理想结果都被模型返回且按照分值高低从大到小的排序)时的 $DCG(q)$ 得分.

(5) *CWS*

*CWS*[114](Confidence Weighted Score)，是 2002 年 TREC 问答评测任务的评价指标之一. 参与评测的系统根据对问题回答的置信度对问题进行排序，评测时根据此排序中问题的回答是否正确来打分，如

① https://trec.nist.gov
② http://research.nii.ac.jp/ntcir/index-en.html
③ http://www.clef-campaign.org

式(11)所示.其中 $Count(i)$ 是前 $i$ 个提问中被正确回答的提问数.

(6) $K$-Measure

$K$-Measure[118]是 2004 年 CLEF 的 Pilot Task 任务的评测指标,它鼓励返回不同的正确答案,但是对于错误答案需要惩罚,如式(12)所示.其中 $Score(a)$ 是问答模型返回的候选答案 $a$ 的打分,$Judgement(a)$ 是人工对候选答案 $a$ 的评分,若评分为"1"表示该候选答案为正确答案,为"0"表明该候选答案已经出现过,为"−1"则表示该候选答案不是正确答案.

(7) Exact Match($EM$)

$EM$[119]是精确匹配,即与正确答案完全匹配的百分比.常用于事实类问答或阅读理解任务.

$$\begin{aligned} Precision@N&=\frac{1}{|Q|}\sum_{q\in Q}\left(\frac{1}{N}\sum_{i=1}^{N}I(\hat{r}_i\in r)\right)\\ Recall@N&=\frac{1}{|Q|}\sum_{q\in Q}\left(\frac{1}{|r|}\sum_{i=1}^{N}I(\hat{r}_i\in r)\right)\\ F1@N&=\frac{P\times R\times 2}{P+R}\end{aligned}\tag{7}$$

$$MAP=\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|r|}\sum_{j=1}^{|r|}\left(\frac{1}{rank(r_j,\hat{r})}\sum_{i=1}^{rank(r_j,\hat{r})}I(\hat{r}_i\in r)\right)\tag{8}$$

$$MRR=\frac{1}{|Q|}\sum_{q\in Q}\max_{r_i\in r}\left(\frac{1}{rank(r_i,\hat{r})}\right)\tag{9}$$

$$DCG(q)=\sum_{i=1}^{|\hat{r}|}\frac{rel_i}{\lg(i+1)}\quad NDCG=\frac{1}{|Q|}\sum_{q\in Q}\frac{DCG(q)}{IDCG(q)}\tag{10}$$

$$CWS=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{Count(i)}{i}\tag{11}$$

$$K\text{-Measure}=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{\sum_{a\in\hat{r}_i}Score(a)\times Judgement(a)}{\max\{|\hat{r}_i|,|r_i|\}}\tag{12}$$

### 4.3　自动问答数据集

(1) 标准数据集

三大主流评测会议均发布了多个关于问答任务的标准数据集.由于 TREC 的数据集主要是英文的,因此其应用最为广泛.TREC 的问答评测任务大多数是面向事实类问答,例如"What are pennies made of?"这样的事实类问题.许多研究者[44,100]使用的是 TREC8 到 TREC12 的数据集,共包含 1000 多个事实[44]类问题.$Precision@N$、$Recall@N$ 和 $F1@N$ 是其常用的评价指标.

与 TREC 相同,NTCIR 和 CLEF 发布的问答评测任务也是面向事实类的.NTCIR 以日语语料集为主,除此之外 NTCIR 还发布过繁体中文和英文的语料集.其主要使用的评价指标有 $Precision@N$、$Recall@N$、$F1@N$ 和 $MAP$.

CLEF 的数据集主要面向欧洲本土语种,例如法语、德语、意大利语等.其问答评测任务主要分为单语言任务和多语言任务,单语言任务是指用哪种语言提问,就用哪种语言回答,而多语言任务是指可以用 CLEF 所用语言中的任一种语言提问,系统须用英语回答.其常用的评价指标有 $MRR$、$CWS$ 等.

中国科学院创建的中文问答评测平台 EPCQA 构建的数据集[114]包含 4250 个基于事实的问题,其问题可以分为三大类,即事实类问题、列表类问题和描述类问题.

斯坦福大学于 2016 年推出了一个阅读理解数据集 SQuAD①.其有两个版本,SQuAD1.1 从 500 多篇文章中构建了 100 000 多对问答对;SQuAD2.0 在 SQuAD1.1 的基础上,增加了 50 000 多条看起来很像正确答案的回答.

(2) 研究者构建的数据集

除了使用标准数据集外,研究者会自主构建数据集用于自动问答.

对于事实类自动问答,一般有两种常用途径采集数据,一种是首先通过搜索引擎搜集问题,然后再从 Freebase、Wikipedia、雅虎问答、知乎、百度百科、百度知道、豆瓣等平台获取答案[44,100,120-122].例如 Yang 等人[120]通过 Bing 搜索引擎收集以"wh"开头的查询问题,答案则来源于 Wikipedia 的文章.另一种是直接从以上平台获取问答对,例如 Qiu 等人[59]除了从雅虎问答中获取英文问答对,还从百度知道中搜集了中文数据,构建了 423 000 对中文问答对.

阅读理解数据集的来源与事实类自动问答不同,其大多来源于真实考试,例如 Lai 等人[123]构建的大型阅读理解数据集 RACE 和 Sun 等人[124]构建的规模稍小的 DREAM,其中 RACE 包含 100 000 个问题和超过 28 000 篇文章,DREAM 包含 10 000 个问题和 6000 篇段落.除了真实考题外,Trischler 等人[125]从 CNN 新闻中挖掘数据,创建了 NewsQA,该数据集包含 119 633 个问题和 12 744 篇来源于 CNN 的新闻.

非事实类自动问答的数据来源于一些问答社区(雅虎问答、百度知道等),例如 SemEval-2016 发布的关于信息查找类问答的数据集[86],该数据集是从 Qatar Living 论坛爬取数据构建而成的;Fang 等

① https://rajpurkar.github.io/SQuAD-explorer/

人[99]从 Quora(美版知乎)上搜集了跨时一年的数据,共获得 444 138 个问题和 887 771 个答案.对于闲聊式对话,研究者通常喜欢从社交网络(Twitter、新浪微博等)中获取相关数据,例如,Wu 等人[84]通过新浪微博搜集了 100 多万个会话,每个会话包含 4 轮对话;Ritter 等人[55]从 Twitter 获取了 130 万个会话;Higashinaka 等人[126]收集了 120 多万个日语推特会话,共 250 万条 tweets.

表 3 展示了有关自动问答的数据集类型以及相关的评价方法.

**表 3 自动问答的数据集类型以及相关的评价方法**

| 数据集名称 | 来源 | 用途 | 评价方法 |
|---|---|---|---|
| 标准数据集 | TREC、CLEF、NTCIR、EPCQA、SQuAD | 主要用于事实类自动问答 | *Precision@N*、*Recall@N*、*F1@N*、*MAP*、*MRR*、*CWS*、*K*-Measure |
| 研究者构建的数据集 | Freebase、雅虎问答、Wikipedia、知乎、百度百科等 | 事实类自动问答 | *Precision@N*、*Recall@N*、*F1@N*、*MAP*、*MRR* |
| | 英语考试、CNN 等 | 阅读理解 | *Precision@N*、*Recall@N*、*F1@N*、*EM*、*MAP*、*MRR* |
| | 雅虎问答、百度知道 | 非事实类自动问答 | *F1@N*、*MAP*、*MRR*、*NDCG* |
| | Twitter、新浪微博等 | 检索式闲聊对话 | *Precision@N*、*MAP*、*MRR* |

## 5 展 望

自动问答的研究具有悠久的历史,它的发展从早期的限定域、面向特定任务、只接收自然语言输入,到后期的非限定域、无特定任务、可接收文本和语音输入,从以信息获取为目的,扩展到以闲聊为乐趣.检索式自动问答作为自动问答的重要技术路线,从人工订制规则,到机器自动学习,再到深度学习,取得了长足进展,但依然充满挑战和期望.以下是对检索式自动问答现存的挑战和未来应用发展趋势的思考.

(1)非事实类自动问答缺乏统一的数据集.TREC、CLEF、NTCIR 等发布的标准数据集均是面向事实类自动问答的,非事实类自动问答的数据集大多是研究者自行构建的.由于搜集时间、搜集方法等不同,即使数据来源一致(Twitter、新浪微博、雅虎问答、Wikipedia 等),但仍不相同,这不利于客观评价和对比各种非事实类问答模型.未来可以尝试规范数据的采集、选取以及答案构建等标准,从而降低来源相同的数据集之间的差异.

(2)非事实类自动问答数据的质量参差不齐.非事实类自动问答的数据大多是从公共平台(雅虎问答、知乎、百度知道等)获取的,提供候选答案的用户的水平和认知参差不齐,导致平台上提供的回答可能是无关的、有偏差的、不全面的、甚至是错误的,这极大地影响了数据的质量,阻碍了该类模型的发展.未来应该探索更多的办法来解决这一问题,例如完善各平台的奖惩激励机制,使其能更准确地反映用户贡献的真实价值.除此之外还可探索答案真实性的判断方法以及当答案出现歧义时,问答模型的处理办法等.

(3)问答模型的通用性有待提升.目前研究者对在各种应用环境中应用问答模型越来越感兴趣,例如教育、商业领域等,但问答模型的通用性欠佳.针对不同的具体应用,问答模型需要重新设计才能达到较好的效果.例如,我们尝试将经典的自动问答模型用在心理健康自动咨询任务上,但效果远不如其在其它领域中的应用效果.未来需要探索问答模型中检索匹配方法的通用性,在模型构建中尽可能地减少人工干预,从而提高模型的可移植性.

(4)模型方法的改进应进一步多样化.例如面向事实类的问答,除了简单引入知识库外,还应关注如何将知识推理、知识图谱等融入当前的问答模型,从而提升回答复杂问题的能力.

(5)检索式自动问答的评价体系需要进一步完善.本文介绍的客观评价指标主要是面向信息检索设计的,但被广泛应用于评价检索式问答模型,且能很好地满足问答模型的评价需求,但对于新应用领域的问答(例如面向心理健康的自动问答),现有的评价体系就无法很好地胜任.设计新的评价指标,完善现有自动问答的评价体系,以适应新的应用领域,这也是自动问答未来研究的一个方向.

(6)探索自动问答的新应用.尽管自动问答的应用领域在不断更新扩展,但还存在一些领域是其未涉及的.例如,我们在收集文献的过程中注意到,自动问答在心理健康服务方面的工作很少见.全球心理健康问题日益严峻,但是有限的医疗资源造成了医患比例的失衡,现有的使用计算机来辅助心理健康治疗的研究无法与当事人建立咨询同盟,仍然需要专业心理咨询师(心理医生)直接介入,无法成为心理健康服务系统的一个很好的补充.随着互联网的高速发展,基于互联网的社交网络已成为社会个体或组织进行情感表达、信息发布、传播与共享的重要渠道,这为使用计算机来辅助心理健康治疗带

来新的研究契机.同时 Tan 等人[127]的研究证实,超过一半的用户愿意阅读自己感兴趣或对自己有帮助,并且是信任的人发送的私信.这项研究表明基于问答模型的心理健康自动咨询具有可行性,可以是心理健康服务系统一个很好的辅助.

我们使用 CLPsych2017 workshop 上的 shared task 提供的英文论坛数据,利用自动问答中的经典基础模型尝试构建基于问答模型的心理健康自动咨询模型,包括使用 CNN、BM25、LDA 检索相关回复(其中 BM25 效果较好,$MAP=0.1509$,$MRR=0.2964$).对实验结果分析发现:基于问答模型的心理健康自动咨询任务,问题与答案多为"心理问题——行为或观点"的新关系模式匹配,并且该类任务的问题和答案文本都偏长,对现有基于检索的问答模型提出了新的挑战.

## 参 考 文 献


[1] Green Jr B F, Wolf A K, Chomsky C, et al. Baseball: An automatic question-answerer//Proceedings of the Western Joint Computer Conference: Papers Presented at the Joint IRE-AIEE-ACM Computer Conference. San Francisco, USA, 1961: 219-224

[2] Woods W A. Progress in natural language understanding: An application to lunar geology//Proceedings of the National Computer Conference and Exposition. New York, USA, 1973: 441-450

[3] Winograd T. Understanding natural language. Cognitive Psychology, 1972, 3(1): 1-191

[4] Diefenbach D, Lopez V, Singh K, et al. Core techniques of question answering systems over knowledge bases: A survey. Knowledge and Information systems, 2018, 55(3): 529-569

[5] Patra B. A survey of community question answering. https://arxiv.org/pdf/1705.04009.pdf, 2017

[6] Wang X, Huang C, Yao L, et al. A survey on expert recommendation in community question answering. Journal of Computer Science and Technology, 2018, 33(4): 625-653

[7] Srba I, Bielikova M. A comprehensive survey and classification of approaches for community question answering. ACM Transactions on the Web, 2016, 10(3): 1-63

[8] Pamungkas E W. Emotionally-aware chatbots: A survey. https://arxiv.org/pdf/1906.09774.pdf, 2019

[9] Almansor E H, Hussain F K. Survey on intelligent chatbots: State-of-the-art and future research directions//Proceedings of the 13th International Conference on Complex, Intelligent, and Software Intensive Systems. Sydney, Australia, 2019: 534-543

[10] Hussain S, Sianaki O A, Ababneh N. A survey on conversational agents/chatbots classification and design techniques//Proceedings of the Workshops of the International Conference on Advanced Information Networking and Applications. Matsue, Japan, 2019: 946-956

[11] Hendrix G G, Sacerdoti E D, Sagalowicz D, et al. Developing a natural language interface to complex data. ACM Transactions on Database Systems, 1978, 3(2): 105-147

[12] Weizenbaum J. ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM, 1966, 9(1): 36-45

[13] Suganuma S. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: Feasibility and acceptability pilot trial. JMIR Mental Health, 2018, 5(3): e10454

[14] He J, Wang B, Fu M, et al. Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems. IEEE Access, 2019, 7: 18871-18883

[15] Cincarek T, Shindo I, Toda T, et al. Development of preschool children subsystem for ASR and Q&A in a real-environment speech-oriented guidance task//Proceedings of the 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, 2007: 1469-1472

[16] Henderson M, Vulić I, Casanueva I, et al. PolyResponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Hong Kong, China, 2019: 181-186

[17] Zhou L, Gao J, Li D, et al. The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 2020, 46(1): 53-93

[18] Zhang R, Wang Z, Mai D. Building emotional conversation systems using multi-task seq2seq learning//Proceedings of the 6th CCF Conference on Natural Language Processing and Chinese Computing. Dalian, China, 2017: 612-621

[19] Colombo P, Witon W, Modi A, et al. Affect-driven dialog generation//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 3734-3743

[20] Sugiyama H, Meguro T, Higashinaka R, et al. Open-domain utterance generation for conversational dialogue systems using Web-scale dependency structures//Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Metz, France, 2013: 334-338

[21] Bateman J, Henschel R. From full generation to 'near-templates' without losing generality//Burgard W, Christaller T, Cremers A B, eds. KI-99: Advances in Artificial Intelligence 23rd Annual German Conference on Artificial Intelligence. Berlin,

Germany: Springer, 1999: 13-18

[22] Wallace R S. The anatomy of alice//Epstein R, Roberts G, Beber G, eds. Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Berlin, Germany: Springer Netherlands, 2008: 181-210

[23] Kwok C, Etzioni O, Weld D S. Scaling question answering to the Web. ACM Transactions on Information Systems, 2001, 19(3): 242-262

[24] Liu Z, Wang X, Chen Q, et al. A Chinese question answering system based on Web search//Proceedings of the 2014 International Conference on Machine Learning and Cybernetics. Lanzhou, China, 2014: 816-820

[25] Li X, Hu D, Li H, et al. Automatic question answering from Web documents. Wuhan University Journal of Natural Sciences, 2007, 12(5): 875-880

[26] Riloff E, Thelen M. A rule-based question answering system for reading comprehension tests//Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems-Volume 6. Seattle, USA, 2000: 13-19

[27] Akour M, Abufardeh S O, Magel K, et al. QArabPro: A rule based question answering system for reading comprehension tests in Arabic. American Journal of Applied Sciences, 2011, 8(6): 652-661

[28] Li X, Roth D. Learning question classifiers: The role of semantic information. Natural Language Engineering, 2016, 12(3): 229-249

[29] Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based Chinese question classification. Journal of Chinese Information Processing, 2006, 20(2): 35-41(in Chinese)
(文勖, 张宇, 刘挺等. 基于句法结构分析的中文问题分类. 中文信息学报, 2006, 20(2): 35-41)

[30] Zhang D, Lee W S. Question classification using support vector machines//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2003: 26-32

[31] Suzuki J, Taira H, Sasaki Y, et al. Question classification using HDAG kernel//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics Workshop on Multilingual Summarization and Question Answering. Sapporo, Japan, 2003: 61-68

[32] Moschitti A, Quarteroni S, Basili R, et al. Exploiting syntactic and shallow semantic kernels for question answer classification //Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, 2007: 776-783

[33] Pota M, Esposito M, De Pietro G. A forward-selection algorithm for SVM-based question classification in cognitive systems//De Pietro G, Gallo L, Howlett R J, et al, eds. Intelligent Interactive Multimedia Systems and Services 2016. Berlin, Germany: Springer, 2016: 587-598

[34] Moldovan D, Harabagiu S, Pasca M, et al. The structure and performance of an open-domain question answering system //Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, China, 2000: 563-570

[35] Xu J, Licuanan A, May J, et al. Answer selection and confidence estimation//Proceedings of the 2003 AAAI Spring Symposium. Palo Alto, USA, 2003: 134-137

[36] Yu Zheng-Tao, Fan Xiao-Zhong, Guo Jian-Yi, et al. Answer extracting for Chinese question-answering system based on latent semantic analysis. Chinese Journal of Computers, 2006, 29(10): 1889-1893(in Chinese)
(余正涛, 樊孝忠, 郭剑毅等. 基于潜在语义分析的汉语问答系统答案提取. 计算机学报, 2006, 29(10): 1889-1893)

[37] Yao X, Van Durme B, Callisonburch C, et al. Answer extraction as sequence tagging with tree edit distance// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 858-867

[38] Wang H, Bansal M, Gimpel K, et al. Machine comprehension with syntax, frames, and semantics//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 700-706

[39] Narasimhan K, Barzilay R. Machine comprehension with discourse relations//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 1253-1262

[40] Wang H, Lu Z, Li H, et al. A dataset for research on short-text conversations//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 935-945

[41] Ji Z, Lu Z, Li H. An information retrieval approach to short text conversation. https://arxiv. org/pdf/1408. 6988. pdf, 2014

[42] Bessho F, Harada T, Kuniyoshi Y. Dialog system using real-time crowdsourcing and twitter large-scale corpus// Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Seoul, South Korea, 2012: 227-231

[43] Higashinaka R, Imamura K, Meguro T, et al. Towards an open-domain conversational system fully based on natural language processing//Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, 2014: 928-939

[44] Sun H, Ma H, Yih W, et al. Open domain question answering via semantic enrichment//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 1045-1055

[45] Savenkov D, Agichtein E. When a knowledge base is not enough: Question answering over knowledge bases with external text data//Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 235-244

[46] Wu W, Lu Z, Li H. Learning bilinear model for matching queries and documents. The Journal of Machine Learning Research, 2013, 14(1): 2519-2548

[47] Deepak P, Garg D, Shevade S. Latent space embedding for retrieval in question-answer archives//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 855-865

[48] Yan Z, Duan N, Bao J, et al. DocChat: An information retrieval approach for chatbot engines using unstructured documents//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 516-525

[49] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 475-482

[50] Cai L, Zhou G, Liu K, et al. Learning the latent topics for question retrieval in community QA//Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand, 2011: 273-281

[51] Zhou G, Cai L, Zhao J, et al. Phrase-based translation model for question retrieval in community question answer archives//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 653-662

[52] Zhai K, Williams J D. Discovering latent structure in task-oriented dialogues//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 36-46

[53] Zhang K, Wu W, Wu H, et al. Question retrieval with high quality answers in community question answering//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. Shanghai, China, 2014: 371-380

[54] Ji Z, Xu F, Wang B, et al. Question-answer topic model for question retrieval in community question answering//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 2471-2474

[55] Ritter A, Cherry C, Dolan W B. Data-driven response generation in social media//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 583-593

[56] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation//Proceedings of the 16th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2002: 601-608

[57] Tan M, Santos C D, Xiang B, et al. LSTM-based deep learning models for non-factoid answer selection. https://arxiv.org/pdf/1511.04108.pdf, 2015

[58] Lu Z, Li H. A deep architecture for matching short texts//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 1367-1375

[59] Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1305-1311

[60] Yu L, Hermann K M, Blunsom P, et al. Deep learning for answer sentence selection. https://arxiv.org/pdf/1412.1632.pdf, 2014

[61] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks//Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 373-382

[62] Xu K, Reddy S, Feng Y, et al. Question answering on freebase via relation extraction and textual evidence//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 2326-2336

[63] Indurthi S R, Yu S, Back S, et al. Cut to the chase: A context zoom-in network for reading comprehension//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 570-575

[64] Yu A W, Dohan D, Luong M, et al. QANet: Combining local convolution with global self-attention for reading comprehension//Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, Canada, 2018

[65] Wu F, Lao N, Blitzer J, et al. Fast reading comprehension with ConvNets. https://arxiv.org/pdf/1711.04352.pdf, 2017

[66] Pan B, Li H, Zhao Z, et al. MEMEN: Multi-layer embedding with memory networks for machine comprehension. https://arxiv.org/pdf/1707.09098.pdf, 2017

[67] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension//Proceedings of 5th International Conference on Learning Representations (ICLR 2017). Toulon, France, 2017

[68] Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 707-712

[69] Tay Y, Phan M C, Tuan L A, et al. Learning to rank question answer pairs with holographic dual LSTM architecture//Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan, 2017: 695-704

[70] Dong L, Mallinson J, Reddy S, et al. Learning to paraphrase for question answering//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 875-886

[71] Hu M, Peng Y, Huang Z, et al. Reinforced mnemonic reader for machine reading comprehension//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18). Stockholm, Sweden, 2018: 4099-4106

[72] Kobayashi S, Tian R, Okazaki N, et al. Dynamic entity representation with max-pooling improves machine reading//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 850-855

[73] Hoang L, Wiseman S, Rush A M. Entity tracking improves cloze-style reading comprehension//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1049-1055

[74] Ghaeini R, Fern X Z, Shahbazi H, et al. Dependent gated reading for cloze-style question answering//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 3330-3345

[75] Gu J, Ling Z, Liu Q. Utterance-to-utterance interactive matching network for multi-turn response selection in retrieval-based chatbots. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 369-379

[76] Rao J, Upasani K, Balakrishnan A, et al. A tree-to-sequence model for neural NLG in task-oriented dialog//Proceedings of the 12th International Conference on Natural Language Generation. Tokyo, Japan, 2019: 95-100

[77] Chen D, Fisch A, Weston J, et al. Reading Wikipedia to answer open-domain questions//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1870-1879

[78] Dhingra B, Mazaitis K, Cohen W W. Quasar: Datasets for question answering by search and reading. https://arxiv.org/pdf/1707.03904.pdf, 2017

[79] Wang Z, Mi H, Hamza W, et al. Multi-perspective context matching for machine comprehension. https://arxiv.org/pdf/1612.04211.pdf, 2016

[80] Clark C, Gardner M. Simple and effective multi-paragraph reading comprehension//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 845-855

[81] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1832-1846

[82] Liu C, Zhao Y, Si Q, et al. Multi-perspective fusion network for commonsense reading comprehension//Sun M, Liu T, Wang X, et al, eds. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Berlin, Germany: Springer, 2018: 262-274

[83] Sharp R, Surdeanu M, Jansen P, et al. Tell me why: Using question answering as distant supervision for answer justification//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada, 2017: 69-79

[84] Wu B, Wang B, Xue H. Ranking responses oriented to conversational relevance in chat-bots//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 652-662

[85] Yin W, Schütze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association of Computational Linguistics, 2016, 4(1): 259-272

[86] Zhang X, Li S, Sha L, et al. Attentive interactive neural networks for answer selection in community question answering//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 3525-3531

[87] Rücklé A, Gurevych I. End-to-end non-factoid question answering with an interactive visualization of neural attention weights//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 19-24

[88] Romeo S, Da San Martino G, Barrón-Cedeno A, et al. Neural attention for learning to rank questions in community question answering//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 1734-1745

[89] Chen R, Yulianti E, Sanderson M, et al. On the benefit of incorporating external features in a neural architecture for answer sentence selection//Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan, 2017: 1017-1020

[90] Kundu S, Ng H T. A question-focused multi-factor attention network for question answering//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 5828-5835

[91] Kadlec R, Schmid M, Bajgar O, et al. Text understanding with the attention sum reader network//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 908-918

[92] Cui Y, Liu T, Chen Z, et al. Consensus attention-based neural networks for Chinese reading comprehension//Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016: 1777-1786

[93] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 593-602

[94] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 189-198

[95] Wang S, Jiang J. Machine comprehension using match-LSTM and answer pointer//Proceedings of the 5th International Conference on Learning Representations (ICLR 2017).

Toulon, France, 2017

[96] Hermann K M, Kočiský T, Grefenstette E, et al. Teaching machines to read and comprehend//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 1693-1701

[97] He X, Golub D. Character-level question answering with attention//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1598-1607

[98] Lu H, Kong M. Community-based question answering via contextual ranking metric network learning//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 4963-4964

[99] Fang H, Wu F, Zhao Z, et al. Community-based question answering via heterogeneous social network learning//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 122-128

[100] Savenkov D, Agichtein E. EviNets: Neural networks for combining evidence signals for factoid question answering//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 299-304

[101] Morales A, Premtoon V, Avery C, et al. Learning to answer questions from Wikipedia infoboxes//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1930-1935

[102] Yan R, Song Y, Wu H. Learning to respond with deep neural networks for retrieval-based human-computer conversation system//Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 55-64

[103] Oh J, Torisawa K, Kruengkrai C, et al. Multi-column convolutional neural networks with causality-attention for why-question answering//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. Cambridge, UK, 2017: 415-424

[104] Sordoni A, Bachman P, Trischler A, et al. Iterative alternating neural attention for machine reading. https://arxiv.org/pdf/1606.02245.pdf, 2016

[105] Du X, Shao J, Cardie C. Learning to ask: Neural question generation for reading comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1342-1352

[106] Gu Y, Gui X, Li D. TT-Net: Topic transfer-based neural network for conversational reading comprehension. IEEE Access, 2019, 7: 116696-116705

[107] Sukhbaatar S, Weston J, Fergus R. End-to-end memory networks//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 2440-2448

[108] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[109] Welbl J, Minervini P, Bartolo M, et al. Undersensitivity in neural reading comprehension//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2020: 1152-1165

[110] Ribeiro M T, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 856-865

[111] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2021-2031

[112] Wang Y, Bansal M. Robust machine comprehension models via adversarial training//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 575-581

[113] Ebrahimi J, Rao A, Lowd D, et al. HotFlip: White-box adversarial examples for text classification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 31-36

[114] Wu You-Zheng, Zhao Jun, Duan Xiang-Yu, et al. Research on question answering & evaluation: A survey. Journal of Chinese Information Processing, 2005, 19(3): 2-14(in Chinese)
(吴友政，赵军，段湘煜等. 问答式检索技术及评测研究综述. 中文信息学报，2005，19(3)：2-14)

[115] Turpin A, Scholer F. User performance versus precision measures for simple search tasks//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA, 2006: 11-18

[116] Chakrabarti S, Khanna R, Sawant U, et al. Structured learning for non-smooth ranking losses//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 88-96

[117] Yilmaz E, Kanoulas E, Aslam J A. A simple and efficient sampling method for estimating ap and NDCG//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 603-610

[118] Herrera J, Peñas A, Verdejo F. Question answering pilot task at CLEF 2004//Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Berlin, Germany: Springer, 2004: 581-590

[119] Baradaran R, Ghiasi R, Amirkhani H. A survey on machine reading comprehension systems. https://arxiv.org/ftp/arxiv/papers/2001/2001.01582.pdf, 2020

[120] Yang Y, Yih W, Meek C. WikiQA: A challenge dataset for open-domain question answering//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2013-2018

[121] Miller A H, Fisch A, Dodge J, et al. Key-value memory networks for directly reading documents//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1400-1409

[122] Berant J, Chou A K, Frostig R, et al. Semantic parsing on freebase from question-answer pairs//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1533-1544

[123] Lai G, Xie Q, Liu H, et al. RACE: Large-scale reading comprehension dataset from examinations//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 785-794

[124] Sun K, Yu D, Chen J, et al. DREAM: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 2019, 7: 217-231

[125] Trischler A, Wang T, Yuan X, et al. NewsQA: A machine comprehension dataset//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada, 2017: 191-200

[126] Higashinaka R, Kawamae N, Sadamitsu K, et al. Building a conversational model from two-tweets//Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. Hawaii, USA, 2011: 330-335

[127] Tan Z, Liu X, Liu X, et al. Designing microblog direct messages to engage social media users with suicide ideation: Interview and survey study on weibo. Journal of Medical Internet Research, 2017, 19(12): e381

**ZHAO Yun**, Ph. D. candidate. Her research interests include social media processing and natural language processing.

**LIU De-Xi**, Ph. D., professor, Ph. D. supervisor. His research interests include social media processing, information retrieval and natural language processing.

**WAN Chang-Xuan**, Ph. D., professor. His research interests include Web data management and sentiment analysis.

**LIU Xi-Ping**, Ph. D., professor. His research interests include Web data management and text mining.

**LIAO Guo-Qiong**, Ph. D., professor. His research interests include social network mining and IoT data management.



**Background**

Automatic question answer is a research hotspot in the field of artificial intelligence and natural language processing. It uses many techniques of artificial intelligence and natural language processing, such as language understanding, machine learning, deep learning, automatic speech recognition and so on. It was originally designed to meet the needs of people to obtain information quickly and accurately. With the development of technology, most of the existing retrieval-based Q&A models have no domain restrictions and can receive text and voice input. Although it has achieved some achievements in recent years, there are few papers to offer a comprehensive reviewing of retrieval-based Q&A.

In order to provide a systematic and comprehensive understanding of retrieval-based Q&A for researchers, it is necessary for us to perform a full survey on the technical methods and evaluation indicators of retrieval-based Q&A. In this paper, we first introduce the classification methods and several typical types of automatic Q&A, and summarize the characteristics of those types. Then, we discuss the three types of methods of the existing retrieval-based Q&A, and focus on analyzing the characteristics and existing problems of those methods. Specifically, we summarize the existing improved methods. Afterwards, we successively introduce the evaluation methods of retrieval-based Q&A. Finally, we discuss the possible future developing trends and challenges of retrieval-based Q&A.

This team has done some works about sentiment analysis and psychology analysis on financial text and social network text, which were published on or accepted by Chinese Journal of Computers, Journal of Chinese Computer Systems, Journal of Chinese Information Processing and Journal of Software etc.

This subject is supported by the National Natural Science Foundation of China (Nos. 61762042, 61972184, 62076112). These projects focus on mental health analysis for social network users, cause analysis of mental health, mental health oriented automatic question answer and financial text mining.