计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):1−15 [doi: 10.15888/j.cnki.csa.009038]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的智能问答系统综述①

姚元杰[1,2], 龚毅光[1,2], 刘　佳[1,2], 徐　闯[1,2], 朱栋梁[1,2]

[1](南京信息工程大学 自动化学院 江苏省大数据分析技术重点实验室, 南京 210044)
[2](江苏省大气环境与装备技术协同创新中心, 南京 210044)
通信作者: 龚毅光, E-mail: yiguang-gong@nuist.edu.cn

摘　要: 在科技发达和信息爆炸的时代, 如何从海量数据中准确地提取所需信息已成为人们研究的目标. 问答系统作为解决此问题的重要途径之一, 其主要通过对已有数据信息进行检索和分析, 并最终返回问题答案或其他相关信息. 近年来, 深度学习的革命性发展给问答系统带来了长足的进步, 序列到序列的模型, 端到端的模型以及最近流行的预训练, 都给问答系统留下无限的发展空间, 但其仍面临许多挑战. 本文首先对问答系统的发展进行简要介绍, 接着将问答系统按照 3 个不同角度进行分类, 并对相关数据集、评测指标和各类问答系统的主流技术进行阐述, 最后对问答系统面临的问题和未来的发展趋势进行讨论.

关键词: 问答系统; 智能问答; 自然语言处理; 深度学习

引用格式: 姚元杰,龚毅光,刘佳,徐闯,朱栋梁.基于深度学习的智能问答系统综述.计算机系统应用,2023,32(4):1–15. http://www.c-s-a.org.cn/1003-3254/9038.html

Survey on Intelligent Question Answering System Based on Deep Learning

YAO Yuan-Jie[1,2], GONG Yi-Guang[1,2], LIU Jia[1,2], XU Chuang[1,2], ZHU Dong-Liang[1,2]

[1](Jiangsu Key Laboratory of Big Data Analysis Technology, School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing 210044, China)

Abstract: In the era featuring advanced technology and information explosion, how to accurately extract the required information from massive data has become the study target. As one of the important ways to solve this problem, question-answering systems mainly retrieve and analyze existing data and information and finally return the answer to the question or other related information. In recent years, the revolutionary development of deep learning has brought considerable progress to question-answering systems. Sequence-to-sequence models, end-to-end models, and the recently popular pre-training have left unlimited development space for the question-answering systems, but these systems still face many challenges. This study first briefly introduces the development of the question-answering systems, then classifies these systems from three different perspectives, and expounds on the relevant data sets, evaluation indicators, and mainstream technologies of various question-answering systems. Finally, the study discusses the problems faced by question-answering systems and their future development trends.

Key words: question answering system; intelligent question answering; natural language processing (NLP); deep learning

问答 (question answering, QA) 是人工智能领域中一个快速发展的研究问题, 用户通过输入语音、文本、视频等多维信息, 通过问答系统处理并返回给用户所需答案. 在互联网和大数据时代, 如何精准而又快

① 基金项目: 国家重点研发计划 (2018YFC1405700); 国家自然科学基金 (61773219)
收稿时间: 2022-07-26; 修改时间: 2022-08-26; 采用时间: 2022-10-27; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-27

速获取所需信息, 一直未能得到有效解决. 虽然各类搜索引擎努力满足用户信息检索的需求, 但用户仍然只能通过关键词来搜索答案, 而且需要从大量的搜索结果中筛选答案. 随着大数据时代的来临, 传统的信息搜索方式已不能满足用户的需求.

关于问答系统, 一般认为输入应是以自然语言形式描述的问题, 输出则是一个简洁的答案或者候选答案的列表, 而不是一堆相关的文档[1]. 例如, 用户提问“明天天气如何?”, 问答系统根据用户问题应该返回一个精简的答案, 例如“晴天”. 问答系统主要需解决 3 个问题.

(1) 机器须理解用户输入的是什么问题.

(2) 根据这些问题关键点去检索并处理相关信息.

(3) 将答案返回给用户.

随着技术的发展, 各种各样的问答系统也纷纷出现. 对于问答系统的分类, 从不同的角度可以分为不同的种类. 本文从 3 个角度对问答系统进行分类.

(1) 按问题维度可分为单模态信息输入的问答系统和多模态信息输入的问答系统. 其中单模态问答系统又可分为开放域型和限定域型, 多模态则指的是以多维信息, 如文本、图像、语音等输入的问答系统. 目前多模态主要的研究热点集中在图像和语音信息输入的视觉问答系统, 本文也将对此进行介绍.

(2) 按数据信息来源进行分类, 可分为基于知识图谱、基于机器阅读理解和基于问答对的问答系统.

(3) 根据问答系统答案生成的方式, 可分为检索方法和生成方法的问答系统.

本文主要通过以下几个部分介绍问答系统: 首先介绍了问答系统的发展历程、主要数据集和相关评测指标, 接着对不同类型的问答系统及其相关技术进行简单介绍, 在数据集介绍中主要将数据集根据不同的问答系统进行划分并将一些指标进行对比, 方便研究者查询. 在问答系统介绍中, 将问答系统按照 3 种标准进行分类, 并介绍了各个问答系统主流方法. 最后, 对问答系统面临的问题和未来的发展趋势进行讨论.

1 问答系统的发展历程

1950 年, Turing 等人[2] 提出了一个图灵测试, 用来确定一台机器是否能思考, 这个测试被认为是问答系统的最早原型. 1960 年前后, 由于人工智能的发展, 研究人员将目光对准限定领域和处理结构数据的问答系统, 这一时期的大多数问答系统是人工智能系统和专家系统, 代表性的系统是 BASEBALL[3] 和 LUNAR[4].

20 世纪中后期, 计算语言学兴起, 大量研究者开始关注如何利用相关语言学技术去提高问答系统的性能, 且主要集中在限定领域和处理结构数据, 比较有名的系统是 Unix Consultant[5].

20 世纪末, 互联网的迅速发展带来了海量的电子文档, 问答系统进入了开放领域和基于文本的时代, 这时常见问题数据出现在互联网上. 21 世纪初, 社区问答数据源源不断地产生, 即有了大量的问题答案对数据.

近年来, 随着相关数据集和深度学习等领域的快速发展, 相关技术在图像处理、自然语言处理 (natural language processing, NLP) 等诸多领域都取得了令人满意的效果, 已被证明能够有效地捕获大数据中的复杂信息. 其中循环神经网络 (recurrent neural networks, RNN)、卷积神经网络 (convolutional neural network, CNN)、长短期记忆网络 (long short-term memory, LSTM)、Attention、Transformer 结构等很多深度学习方法都被运用在问答系统的相关任务中.

问答系统的核心之一在于如何更好地建模语言. 传统的词嵌入方法主要有 Word2Vec[6]、GloVe[7] 模型等. Word2Vec 是 Mikolov 等人[6] 于 2013 年提出的一种词嵌入方法, 它的特点是将所有的词向量化, 这样词与词之间就可以定量的去度量它们之间的关系, 进而挖掘出词与词之间的联系, Word2Vec 主要有 CBOW 和 skip-gram 两种模型. 相比于 Word2Vec 关注单词同时出现的概率分布, GloVe 更加关注单词同时出现的概率比率, 其不需要计算那些共现次数为 0 的词汇, 因此, 可以较大地减少计算量和数据的存储空间.

与传统的语言建模方法相比, 预训练语言模型利用大规模的文本数据进行训练, 能够更好地进行语义表征, 且能解决传统模型存在一词多义的问题. 目前比较流行的是 Google 于 2018 年提出的 BERT[8], 随后几年 BERT 的变体层出不穷, 如图 1.

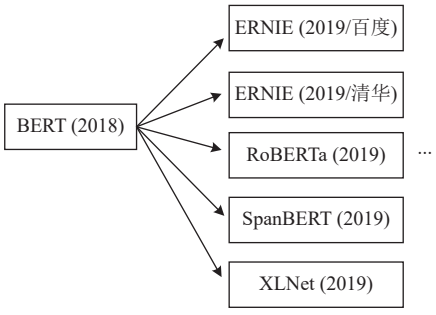

图 1 BERT 部分变体

BERT 采用生成式的掩码语言模型 (masked language model, MLM), 该模型是一种双向 Transformer 结构, 其是第一次采取预训练和微调的方法, 即首先在 BooksCorpus 语料库[9] 和英语维基百科进行预训练, 而微调则是将特定任务的输入、输出插入 BERT, 并微调参数. 该模型在 11 个 NLP 任务上都取得了很好的结果.

表 1 为部分常见预训练模型的对比, 可以看出预训练模型大都采用 Transformer 结构[10], Transformer 拥有更好捕捉信息和并行计算的能力, 但其对于位置信息的处理还需进一步的改善.

表 1 部分常见预训练模型对比

模型	语言建模	模型结构	创新点
ELMo[11]	LM	双向LSTM	加入动态词向量, 使用2个单向语言模型拼接获取语义信息
GPT1.0[12]	LM	单向Transformer	统一框架处理下游任务
BERT[8]	MLM	双向Transformer	其采用新的MLM模型, 以致能生成深度的双向语言表征
ALBERT[13]	MLM	双向Transformer	相比于BERT, ALBERT用更少的参数, 取得了较好效果
SpanBERT[14]	MLM+SBO	双向Transformer	更好span mask方案, 加入SBO训练目标, 不再依赖分词内单个字词表示
XLNet[15]	PLM	双向Transformer-XL	Transformer改为Transformer-XL, 来满足长句子的信息获取并且加入双注意力流机制
RoBERTa[16]	MLM	双向Transformer	采用更大的batch、动态的掩码方式和训练数据更多
ERNIE1.0[17]	MLM	双向Transformer	3种掩码策略预测短语和实体
ERNIE (THU)[18]	MLM	双向Transformer	将实体向量和文本表示融合
ELECTRA[19]	RTD	Generator-discriminator	采用类似于GAN的模型, 将MLM改为RTD模式
K-BERT[20]	MLM+KG	双向Transformer	加入图谱, 采用软位置和可见矩阵来解决知识噪音问题

2 数据集与评测

2.1 数据集

2015 年前后, 几个大规模 QA 数据集的发布, 极大地推动了问答系统的发展. 其中比较有名的数据集有斯坦福大学发布的 SQuAD 数据集, 此数据集是从维基百科中衍生出来的问答集, 其问题的正确答案是给定文本中的任何序列的标记. SQuAD 1.1[21] 包含 536 篇文章, 其中有 107 785 个问答对. 版本 SQuAD 2.0[22] 中将 SQuAD 1.1 中的 100 000 个问题与超过 50 000 个不可回答的问题结合在一起, 加入了对抗性的问题.

QuAC[23] 是 2018 年发布的数据集, 其是一个基于上下文语境的问答数据集, 包含了 100k 个问题. QuAC 有两个比较显著的特点, 提问者只知道问答的主题并没看过文章, 这样就避免了提问者根据文章关键字或近义词提出问题. 第 2 点是提问的问题之间存在语义关联, 可以通过训练模型来提取上下文语义.

CoQA[24] 是一个用来衡量机器对话式问答能力的数据集. 数据集包含 127 个问题和答案, 这些数据包含 8k 个文本段落, 这些段落又分属于 7 个不同的领域. 与传统机器阅读数据集不同, CoQA 中的问题和答案更加简洁自然, 且和人们日常对话更加相似.

在 CoQA、SQuAD 2.0 和 QuAC 中都包含了无法回答的问题, 从数据集中各取 50 个上下文无法回答的问题进行比较, 发现 SQuAD 2.0 包含不可回答问题的种类最多[25].

COCO-QA[26] 和 VQA-real[27] 是两个常见的视觉问答系统数据集. COCO-QA: 问答对是由 NLP 算法生成的, 图像来自 COCO 数据集, 一共有 78 736 个训练问答对和 38 948 个测试问答对. VQA-real: 数据集共分为 v1 和 v2 两个版本, 图像来自 MSCOCO 数据集, 问答包括开放式问题和封闭式问题. v1 包含 614 163 个问题, v2 包含 110 万条问题.

其他还有一些数据集. TREC-QA[28] 数据集有两个版本, 这两个版本有相同数量的训练和测试问题数, TREC-5 比 TREC-6 拥有更多主题的问题. 微软 2016 年发布了 MS MARCO 数据集[29], 与所有问题都是由编辑产生的 SQuAD 不同, MS MARCO 所有的问题都来自于用户的搜寻样例和使用 Bing 搜索引擎得到的真实的网页文件, 且其中有的为生成式的答案, 便于研究生成式的问答系统. WikiQA 数据集[30] 由一系列问答对组成, 为开放领域的问答系统研究提供便利, 并且数据集还包含了一些无法回答的问题, 方便研究者对答案触发模型的研究. NewsQA 数据集[31] 是用于机器阅读理解的数据集, 它为研究者提供了超过 10 万个经过人工标注的问答对, 其问题答案主要来自于美国有线电视新闻网的文章. QAConv 数据集[32] 由香港科技大学于 2021 年提出, 专注于提供信息对话, 与开放领域和面向任务的对话不同, 这些对话通常是长时间

的、复杂的、异步的, 并且涉及不同的领域知识.

除此之外还有一些中文数据集. 哈工大讯飞联合实验室在 2016 年发布了第一个中文完形填空阅读理解数据集 PD&CFT[33], 其中包括了《人民日报》新闻数据集和“儿童读物”数据集. 同年, 百度发布了 WebQA[34] 数据集, 2017 年又发布 Dureader[35] 数据集, 它们是一种大规模开放领域的中文机器理解数据集, 数据源于百度搜索和百度知道 (包含了约 20 万题, 40 万答案和 1M 的文档), 其答案是手工生成的, 且 Dureader 数据集中还包含大量的是非和观点类的问题. 接着, 哈工大讯飞联合实验室在 2018 年前后, 先后颁布了数据集 CMRC 2017[36]、2018[37]、2019[38] 三个版本, 并以此举办了“讯飞杯”中文机器阅读理解的比赛.

表 2 列出了部分数据集的数据来源、类型、文档数、问题数、评价指标和所属语种. 我们主要搜集了英语和中文两个语种的数据集. 从表 2 可知, 数据集的数据来源包含维基百科、对话文章、新闻、文学作品、电子邮件、日志、报纸、故事等, 各个数据集多采用不同的数据源. 从文档数和问题数来看, 大部分数据集的数量在几十到几百 k, 这种量级的数据集既能够覆盖较多的问题, 又便于问答系统处理. 表 2 中的“类型”列, 给出了数据集的类别, 这些类型主要按照数据集的信息特征进行划分, 主要包括自由文本、无法回答、完型填空、区域预测、命名实体识别 (named entity recognition, NER). 由表 2 可知, 部分数据集可以准确地归属到某一类型, 而有些数据集较复杂, 包含两个或两个以上类型的信息特征.

“问答系统”列给出了数据集适合的问答系统类型. 对于基于知识图谱的问答系统来说, 处理实体、属性、关系的三元组信息是系统的首要任务. 提供了这些信息的数据集有: MSRA 数据集[39]、Weibo 数据集[40]、人民日报数据集 (http://s3.bmio.net/kashgari/china-people-daily-ner-corpus.tar.gz)、NLPCC2016KBQA 数据集 (http://tcci.ccf.org.cn/conference/2017/taskdata.php) 等.

表 2 部分数据集及其相关信息

数据集	数据来源	类型	文档数	问题数	评价指标	语种	问答系统
SQuAD 1.1	维基百科	区域预测	536	100k	*F*1、*EM*	英语	机器阅读理解
QuAC	对话、文章	自由文本、无法回答	14k	100k	*F*1	英语	问答对
CoQA	新闻、文学作品	自由文本、无法回答	8k	127k	*F*1	英语	问答对
SQuAD 2.0	维基百科	区域预测、无法回答	—	—	*F*1、*EM*	英语	机器阅读理解
NewsQA	CNN新闻	区域预测	12k	120k	*F*1	英语	机器阅读理解
WikiQA	维基百科	区域预测	20k	3k	*F*1	英语	问答对
MS MARCO	用户日志	自由文本、无法回答	1M passage, 200k+ doc	100k	*BLEU*, *ROUGE-L*	英语	机器阅读理解
QAConv	电子邮件、工作等	自由文本、无法回答	—	34 204	*F*1、*EM*	英语	问答对
CMRC (2019)	故事	完形填空	10k	100k	QAC、PAC	中文	机器阅读理解
PD&CFT	人民日报、儿童故事	完形填空	28k	28k	准确率	中文	机器阅读理解
WebQA	百度知道	区域预测	—	42k	*F*1、*EM*	中文	问答对
Dureader	社区、百度百科	自由文本	1M	200k	*BLEU*, *ROUGE-L*	中文	机器阅读理解
Resume NER	新浪金融	NER	3.8k+ sentence	—	*F*1	中文	知识图谱
Weibo NER	新浪微博	NER	1.4k+ sentence	—	*F*1	中文	知识图谱
VQA-real v1	MSCOCO	—	—	614 163	*MRR*	—	视觉问答

对于基于机器阅读理解的问答系统而言, 系统需根据给定的上下文来回答问题, 常见的任务可以分为 4 种类型: 完形填空、多项选择、片段抽取、自由回答, 适合的数据集有: SQuAD、CMRC (2019)、MS MARCO、Dureader 等.

对于基于问答对的问答系统, 系统不仅要实现对话历史的回答, 而且要保证答案的自然性, 在某些方面还需对对话者的意图进行识别, 常见的数据集有: CoQA 数据集、ATIS 数据集[41] 和 SNIPS 数据集[42] 等.

2.2 相关评价指标

关于评测指标, 常见的有 exact match (*EM*)、*F*1、mean reciprocal rank (*MRR*)、*BLEU*[43] 和 *ROUGE*[44] 等指标.

EM 用来评价预测中匹配到正确答案的百分比, 如式 (1). 其常用于 SQuAD 数据集任务之中.

$$EM = \frac{Num_{\text{right}}}{Num_{\text{total}}} \tag{1}$$

*F*1、*Pre*、*Rec* 常用于命名实体识别等任务相关评

测中. $F1$ 值表示答案之间的重合度, 如式 (2):

$$F1=\frac{2\times Pre\times Rec}{Pre+Rec} \tag{2}$$

其中, Pre 为精确率, 如式 (3), Rec 为召回率, 如式 (4):

$$Pre=\frac{TP}{TP+FP} \tag{3}$$

$$Rec=\frac{TP}{TP+FN} \tag{4}$$

其中, TP 为被模型预测为正类的正样本; FP 为被模型预测为正类的负样本; FN 为被模型预测为负类的正样本.

MRR 用于评估 NLP 任务, 例如查询文档排名和 QA 中排名算法的性能. MRR 定义如式 (5), 其中 Q 为查询个数, $rank_i$ 为查询的序列.

$$MRR=\frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{5}$$

$BLEU$ 和 $ROUGE$ 这两种指标能够评价语言生成的质量, 常用于机器翻译和文章摘要评价. 不同的是 $BLEU$ 通过计算与参考语句的相似度和计算语句流畅性来衡量生成的质量, $ROUGE$ 则主要是基于召回率的计算, 包含 $ROUGE\text{-}L$ 和 $ROUGE\text{-}N$ 等. 其中, $BLEU$ 公式如式 (6)、$ROUGE\text{-}N$ 公式如式 (7):

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right) \tag{6}$$

其中, BP 为长度惩罚因子, w_n是针对不同的 n-gram 的权重, p_n 为修正的 n-gram 精度, n-gram 则指一个语句里面连续的 n 个单词组成的片段, N 为 n-gram 的长度.

$$ROUGE\text{-}N=\frac{\sum_{S\in\{\text{ReferenceSummaries}\}}\sum_{gram_n\in S}Count_{\text{match}}(gram_n)}{\sum_{S\in\{\text{ReferenceSummaries}\}}\sum_{gram_n\in S}Count(gram_n)} \tag{7}$$

其中, $Count_{\text{match}}(gram_n)$ 为表示同时出现在一篇候选摘要和参考摘要的 n-gram 个数; $Count(gram_n)$ 为参考摘要里的 n-gram 个数.

其他还有一些评测指标: 问题准确率 QAC, 篇章准确率 PAC 和曲线下面积 AUC 等.

3 问答系统分类

问答系统想要满足用户需求, 主要需处理 3 个问题: 问题分析、信息检索和答案生成. 根据不同的维度和不同的处理方法,

本文对问答系统进行分类, 如图 2.

3.1 根据问题领域分类

根据问题的知识领域, 问答系统可分为单模态信息问答系统和多模态信息问答系统. 其中单模态信息的问答系统又可分为开放域 (问题可以是任意领域) 和限定域型 (问题只局限于某一领域)[1]. 限定域问答系统可以形象表示为提供问题 (question, Q) 以及包含答案内容的文档 (document, D), 通过问答系统处理后得到答案 (answer, A), 如式 (8) 所示:

$$A=F(D,Q) \tag{8}$$

限定域自动问答系统应用的代表有: 科技馆导游机器人、商场导购机器人等.

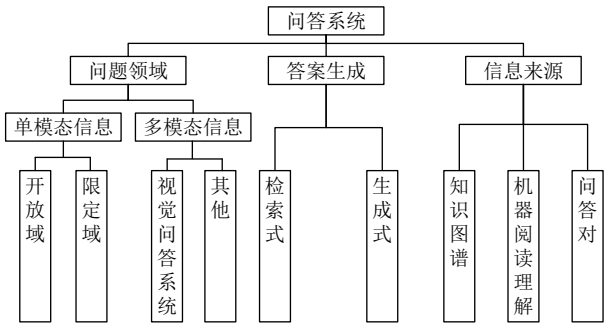

图 2 问答系统分类

开放域问答系统最初被定义为在非结构化的文档中寻找答案, 代表应用包括微软小冰, 小米小爱等. 由于开放域问题种类和领域的多样性, 人工创建答案的模式已满足不了需求, 对于此问题, 基于深度学习技术的智能问答系统应运而生. 目前, 开放域比较流行利用神经网络来构建人和机器的问答渠道.

问答系统整体框架如图 3.

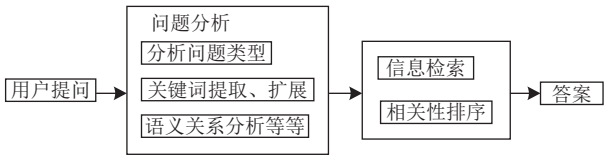

图 3 问答系统整体框架

多模态指的是以多维信息, 如文本、图像、语音等输入的问答系统. 目前多模态主要的研究热点集中在图像和语音信息输入的视觉问答系统. 如图 4 所示, 视觉问答系统与以往仅用于处理单模态信息的 QA 系

统不同, 需要计算机同时处理来自视觉和语音两个模态的输入[45], 即同时处理图形信息和语音信息. 在对话任务中, 视觉问答系统还需结合对话历史的上下文信息进行推理.

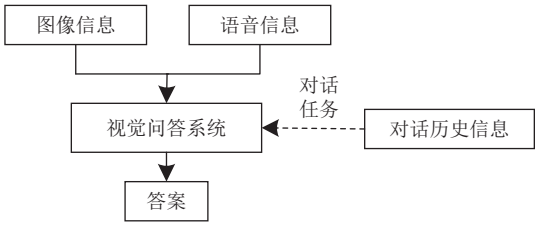

图 4 视觉问答系统

3.2 根据信息来源分类

根据问答系统信息源的数据类型的不同, 可将问答系统分为: (1) 数据来源于结构化知识图谱的问答系统; (2) 数据来源于对话、问答对的基于问答对的问答系统; (3) 数据来源于自由文本的基于机器阅读理解的问答系统.

3.2.1 基于知识图谱问答系统

1977 年, 知识工程的概念在第五届国际人工智能大会上被提出[46], 作为代表的专家系统[47] 得到了广泛的研究和应用. 进入 21 世纪, 随着信息化时代的发展与大数据时代的来临, Google 于 2012 年提出一种大规模语义网络并定义为知识图谱, 即将现实中的实体关系, 利用语义网描述成可以被计算机处理的结构, 用图的形式将现实世界中复杂的关系简单化、直观化.

作为问答系统的重要组成部分, 知识图谱主要通过实体抽取、关系抽取以及属性抽取等自然语言处理技术, 实现文本知识的抽取、融合和存储工作, 构建一个完整的关系网络图. 构造流程图如图 5.

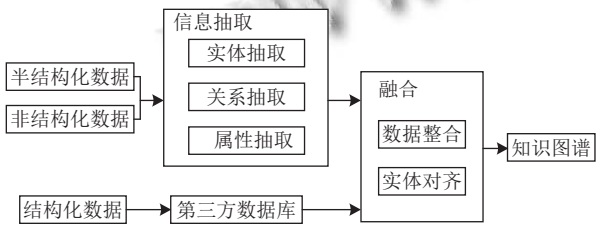

图 5 知识图谱构造流程图

在基于知识图谱的问答系统 (knowledge graph question answering, KGQA) 中, 知识图谱的本质是一个具有特定数据格式的数据库. 存储的数据有其自身的特点, 这些数据以 RDF 三元组[48] 的形式存在. 在进行图谱构建时, 根据数据收集和模型建立的方式不同, 形成了自顶向下和自底向上两种构建知识图谱的方法. 构建过程中, 主要涉及实体识别、关系抽取、知识融合以及知识存储等技术[49]. 本文介绍知识图谱问答系统中较为热点且重要的实体抽取、关系抽取的研究现状.

命名实体识别 (NER) 即实体抽取技术, 它可以识别原始数据语料库中的命名实体, 并对它们进行分类. 早期实体提取的关键目标是原始数据中的时间、地点、名称和专业名称, 识别任务通常面向特定的行业和领域. 1991 年 Rau[50] 首次提出 NER 任务, 自此拉开了 NER 任务在 NLP 的序幕. 起初, 命名实体识别大多是基于规则、统计或是两者结合的混合方法. Sheffield 大学提出的一种基于规则的命名题识别系统[51], 此方法鲁棒性不高. Bikel 等人[52]在 1999 年提出了基于统计法的隐马尔可夫方法, 该方法及其变种后来被广泛应用. Ratnaparkhi[53] 提出用最大熵求解文本分类的问题. 近年来, 由于深度学习能够从自由文本等非结构化数据中提取相关特征, 且相关模型取得不错的效果, 因此采用深度学习来识别命名实体的方法已然成为一种趋势. Bordes 等人将 DNN 模型[54] 应用在命名实体识别和词性标注中, 取得了较好的效果. 也有人将传统机器学习方法与深度学习结合, 取得了不错的效果. 孙娟娟等人[55] 采用 LSTM-CRF 模型的方法来识别渔业领域命名实体, 其识别结果的准确率、召回率以及 $F1$ 值比采用单一LSTM 或 CRF 模型的识别结果提高了 3% 左右. 陈鹏等人[56] 利用融合多特征的 BERT 模型来从电力规章制度中识别相关特征实体, 取得了较高的识别率. Huang 等人[57] 在 BiLSTM-CRF 模型中添加注意力机制, 在识别疾病相关名称时, 一致率高达 0.87. 陈彦妤[58] 则采用预训练模型 BERT的字向量作为 BiLSTM-CRF 模型的输入, 在保险行业相关的数据集上得到了较好的准确率和召回率. 也有不少研究者将字形作为特征信息之一进行提取, 例如文献 [59,60]. 命名实体识别过程如图 6, 数据输入后, 系统首先对这些数据进行预处理, 然后再采用具体的方法对处理后的信息进行分析, 最终得出识别结果.

对语料库采用命名实体识别后, 会得到一系列离散的命名实体. 但要想得到正确的语义信息, 只有实体信息是不够的, 还需要提取这些实体之间的关联关系. 关系抽取是在实体识别的基础上抽取实体之间的语义关系的技术.

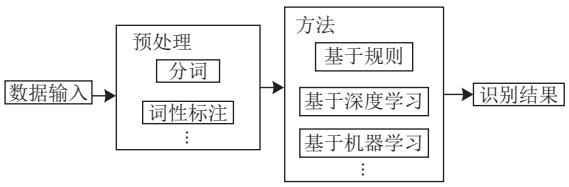

图 6 命名实体识别过程

如果把自动识别实体对间的语义关系形式化, 可将该过程定义为: 对于句子 S, 将其所含的实体定义为集合, 其中下标代表句子中所含实体的序号, 实体的关系则定义为 R_{ij}, 具体如式 (9) 所示. 其中 E_i、E_j 为实体, N^*为实体序号:

$$\left(E_i, E_j, S\right) \to R_{ij}(i \neq j \ \&\ i, j \in N^*) \tag{9}$$

随着互联网技术的发展和数据化时代的来临, 越来越多的问答系统需要在大量非结构的文本里获取对用户有价值的信息. 为此, 一些重要的评测会议得以举办, 如 ACE、MUC 会议和 SemEval 评测会议, 这些会议极大地推动了关系抽取技术的发展. 目前, 关系抽取技术常见的有两种方法: 一种是基于规则的方法, 另一种是基于机器学习的方法. 基于规则的方法主要是按照人工定义的相关规则等去抽取相应的关系实例, 而机器学习方法则是先对相关数据进行标注, 再训练出符合当前关系抽取的模型, 最后采用训练好的模型实施关系抽取. 近些年, 深度学习火遍各个领域, 不少学者也将其应用到关系抽取的任务中来.

1988 年, Millller 等人[61] 开发了第一代实体关系抽取系统, 他们采取相同的概率去实现词性标注和实体关系抽取. 2004 年, Kambhatla[62] 利用最大熵模型实现了实体关系的抽取. Socher 等人[63] 采用深度学习技术: RNN 实施关系抽取任务. Zeng 等人[64] 采用 CNN 对对象的层次特征进行关系抽取. Nguyen 等人[65] 在 CNN 卷积层中采用不同大小的卷积核进行提取, 以此提出了多尺寸卷积神经网络的关系抽取. Zheng 等人[66] 提出共享神经网络的方法, 即用 CNN 实现关系提取, 用双向 LSTM 的编码器-解码器实现实体提取, 然后通过底层神经网络模型进行参数共享, 实现实体和关系模块的联系. Miwa 等人[67] 在此基础上通过神经网络来完成解码, 在 BiLSTM 基础上结合树最短路径来进行关系分类.

基于不同方法的抽取技术比较如表 3. 在大数据和智能化发展的时代下, 相较于笨重的基于规则、统计和传统机器学习的方法来说, 基于深度学习方法具有人工干预少、智能化程度高等特点. 且多数情况下, 效果要好于传统方法.

表 3 部分抽取技术比较

抽取技术	方法	优点	缺点
实体抽取	基于规则、统计	表示直观	成本高、涉及范围低
	基于传统机器学习	鲁棒性较好	需要大量标注数据、存在数据稀疏问题
	基于深度学习	人工依赖少	对算法要求高、针对领域的模型不方便移植
关系抽取	基于统计	召回率结果较好	成本高, 较适合特定领域
	基于深度学习	人工干预不高	对算法要求高、模型迁移性低

关于基于知识图谱问答系统的应用, 大多集中在特定领域. 常见的领域有: 医疗领域、金融投资领域、电商领域、聊天机器人领域等.

医疗领域方面: Open Phacts 整合并构建了药理学的知识图谱, 来支持该领域的发展.

金融投资领域: Kensho 通过各种数据源搜集信息, 构建的图谱能帮助用户在特定行情下做出较正确的判断.

电商领域: 阿里巴巴构建了百亿级别的商品知识图谱, 用于各项服务.

聊天机器人领域: 琥珀·虚颜 (全球第一个具有人工智能的虚拟偶像)、天猫精灵等, 背后均有大规模知识图谱的支持. 在琥珀·虚颜中, 除了通用百科知识图谱, 还包含例如动漫知识图谱的众多子领域.

3.2.2 基于机器阅读理解的问答系统

基于机器阅读理解的问答系统是由计算机自动根据给定的语料资料来回答用户所提出的问题, 是人工智能领域最困难的挑战之一. 机器阅读理解主要任务有: 完型填空[68]、单项选择[21] 以及答案抽取[69] 等.

机器阅读理解任务可类似为一个监督学习的问题: 对给定的样本集合, 训练出一个模型 F. 输入问题 c 和文章 d, 输出答案 a, 如式 (10):

$$F(c, d) \to a \tag{10}$$

目前机器阅读理解研究有两种主流的方法: 一种是基于循环神经网络与卷积神经网络的方法, 例如: BiDAF[70]、Match-LSTM[71] 和 QANet[72]. 另一种是基于预训练语言模型的方法, 如 ELMo、BERT、ALBERT、RoBERTa、XLNet 和 ELECTRA.

基于循环神经网络与卷积神经网络的机器阅读理

解模型是一个端到端的深度神经网络模型, 其中包含若干不同功能的神经网络层, 一般有 4 个主要层: 词向量转换层、编码层、注意力层和答案预测层.

除了 BERT 及其变体的预训练模型外, 下面介绍几种基于循环神经网络与卷积神经网络的机器阅读理解的模型.

BiDAF 是一种双向注意流网络, 包含多个阶段架构, 分别用于从不同级别上对上下文段落进行建模, 并使用双向注意力机制获得感知上下文表示. BiDAF 模型由六层组成, 分别为字符嵌入层、单词嵌入层、上下文嵌入层、注意力机制层、采用递归神经网络的建模层和提供答案的输出层. 与以往不同的是, 模型计算每个时间步长的关注度而非将上下文总结成固定的向量, 这样能减少信息损失; 其次, 使用无记忆力机制, 使得后面注意力的计算不会受到之前错误信息的影响, 并且计算 query-to-context 和 context-to-query 两个方向的注意力机制, 彼此提供互补信息. 该模型在 Cable News Network/DailyMail 完形填空任务中, 取得当时最先进的结果.

QANet 由 Yu 等人[72] 在 2018 年提出, 其是一个前馈模型, 只包含卷积和自我注意力. QANet 模型主要包含 5 个组件: 嵌入层、嵌入编码层、上下文关注层、模型编码层和输出层. 但不同的是在嵌入和建模编码层中, 模型只利用卷积和自我注意机制, 不需要递归网络进而减少了训练和推理的时间. 在 SQuAD 数据集上, 其 *F*1 指标达到 84.6%.

ROaD-ELECTRA for MRC&NLI[73] 由 ElFadeel 等人在 2021 年提出. 文章在多任务学习 (multi-task learning, MTL) 中引入知识提炼 (knowledge distillation, KD) 来作为预处理的第 2 个步骤, 以改善 NLU 任务中的通用语言表征、下游表现和概括效果. 将 robustly optimized and distilled 与 ELECTA 模型结合, 与 BERT 不同的是 ELECTA 并没有采取 MLM 任务, 而是改成了替换令牌检测 (replaced token detection, RTD) 任务. 此方法在 MRC 和 NLI 任务中产生明显强基线改进, 其在 SQuAD 2.0 和 MNLI[74] 的相关任务中取得当时最好的结果.

Retrospective Reader for MRC[75] 由 Zhang 等人于 2020 年提出. 其模型是一种回顾式阅读器, 包含两个平行的模块组: 粗略阅读模块和精读阅读模块. 粗略阅读模块对问题可答性做出大致判断, 接着同精读模块共同预测候选答案, 最后将其回答的正确性判断和粗略读的判断得分相结合, 得出最终的答案. 回顾式阅读器不是简单叠加答案验证程序, 实验结果也表明, 验证机制的选择对 MRC 性能有很大影响, 该模型在 SQuAD 2.0 和 NewsQA 两个数据集的相关任务中, 取得当时的最好成绩.

由表 4 可直观看出, 不同于以上基于 DNN 和 RNN 的模型, 基于预训练加微调的模型在处理效率和答案呈现的准确率上基本上都有一定的优势.

表 4 部分模型在 SQuAD 1.1 数据集上 *EM* 和 *F*1 指标 (%)

模型	BiDAF	Match-LSTM	QANet	FastQA	BERT	ALBERT	XLNet	RoBERTa
EM	68.0	64.7	76.2	68.4	87.4	89.3	89.7	88.9
*F*1	77.3	73.7	84.6	77.1	93.2	94.8	95.1	94.6

3.2.3 基于问答对的问答系统

基于问答对的问答系统可以看成是多轮短对话任务. 短对话任务作为聊天机器人实现自然语言对话的第一步, 其在 NTCIR-12 中首次被提出. 多轮对话大体上根据内容可以分为两个方向: 一是对于一个问题, 涉及多个实体或信息. 由于初次提供的信息不完整无法回答, 需要反复询问以补充完整, 如图 7.

甲: 你要买什么
乙: 我想喝饮料
甲: 你想喝什么饮料
乙: 可乐
甲: 好的, 一瓶可乐 3 元
...

图 7 问答示例

第 2 个方向是对于一个话题, 不断进行问答对话, 类似于采访, 每次的问题都会被回答, 随着对话轮次的增加, 话题不断深入. 这类对话的特点是每次问答之间有高度相关性, 但均可以独立作为一个问答对, 没有很强的依赖性.

对话式问答的问题和答案基本都是短语或句子, 没有其他多余的内容. 问答系统在识别后, 应去除与主题无关的语句, 且当系统没有很好解决用户问题时, 用户可以进行多轮提问. 以传统问答为基础的交互式问答系统是在传统的问答系统基础上加入交互功能, 如加入连续问句的处理, 以及上下文信息处理, 在系统中, 上下文的关系要素等信息是不可或缺的. 根据对话涉及的人数不同, 可以分为双人对话和多人对话, 多人对

话也可看作是双人对话的组合.

基于问答对的问答系统使用很普遍. 早在 1992 年, 美国 DARPA 就开发了 SLS 项目[76], 为用户提供航班信息. 欧盟也在 20 世纪开发了多个项目, 比如支持 3 种语言的列车时刻信息系统[77] 和提供标准的保险合约查询电话呼叫中心[78]. 当下, 各大 IT 公司也开发了各种聊天机器人, 如苹果的 Siri、微软小冰、小米小爱等.

关于多轮短对话问答系统, 传统方法有基于隐马尔可夫模型[79]、朴素贝叶斯[80] 和条件随机场[81]. 近年来, 学者们将深度学习的方法引入进来. Lee等人[82] 提出了基于 CNN 和 RNN 的模型, 用向量表示每一个短文本, 并对当前文本进行分类. Cherng等人[83] 提出用两种分层多栈模型去完成 dialogue quality 和 nugget detection 子任务, 除了 CNN 和 Word2Vec 算法, 还加入注意力机制和门控制机制并尝试引入 BERT 模型. Song 等人[84] 提出了一种用于多意图识别的多标签分类方法, 可解决高标签成本问题. Liu 等人[85] 提出了一种基于 BERT 和 BiLSTM模型, 可用来检测用户表达的意图, 其模型的意图检测准确率高达 92.39%.

3.3 根据答案生成方式分类

根据答案生成的不同方式, 可将问答系统分为检索式和生成式. 检索式问答系统主要通过检索抽取文本库或数据里的词句, 以此作为答案呈现给用户; 而生成式主要是通过一定规则来生产答案后, 给用户合适的回复.

3.3.1 检索式

20 世纪 90 年代以后, 随着信息检索技术的不断发展, 基于检索的问答系统已成为问答系统领域的一个研究热点. START 系统[86] 是世界上第一个检索式问答系统, 由手动构建的知识库和互联网数据作为答案来源. Answerbus 系统[87] 是 Michigan 大学开发的基于信息检索式的问答系统, 该系统返回的是包含答案的句子和对应的网页.

检索式问答系统, 主要通过计算语句相似度, 并将候选答案进行权重计算和排序. 早期的相关度算法源自于信息检索, 其目的在于如何在大量的数据中快速地定位出相关的文档集合. 典型的如 Salton 等人提出的向量空间模型 (VSM)[88], 该模型使用了词袋表示方法[89]. Ramos[90] 提出使用词频-逆文档频率 (term frequency-inverse document frequency, TF-IDF) 的方法, 来更好地体现关键词在文档中的重要性. 近些年, 随着深度学习在 NLP 领域各个方面崭露头角, 研究者们也关注采用基于深度学习的方法去计算相关度, 取得了显著的进展. Zhou 等人[91] 为了得到问题和答案的匹配关系, 采取有监督的方式利用 CNN 将问题和答案转换为向量形式进行训练. Tan 等人[92] 将 CNN 和 RNN 结合, CNN 关注局部信息, RNN 关注全局信息. Huang 等人[93] 提出的深度结构化语义模型 (deep structured semantic model, DSSM), 并在此基础上利用卷积网络实现了语义匹配模型 (convolutional DSSM, CDSSM)[94]. Mitra 等人[95] 提出 DUET 模型, 它是采用局部和分布式表示的匹配模型. Wan 等人提出了基于多位置句子表示的语义匹配 MV-LSTM 模型[96] 和 Match Pyramid 模型[97].

检索式问答系统回答相对精确且回复灵活, 但很依赖于事先构建或选择的数据库和检索算法, 数据库的好坏直接影响后续答案的呈现质量.

3.3.2 生成式

生成式问答系统通过理解问题, 使用答案生成的方式为用户提供合适的回复. 早期的答案生成方式多采用人造模板, 即根据提问内容, 选择合适的答案模板. 这种方法虽然合成的答案合乎基本语法和逻辑, 但是人工干预太多且无法应对所有语言表达模式.

目前, 生成式的问答系统大都基于传统 Seq2Seq 和 Transformer 框架. Seq2Seq 最早被提于 2014 年, 属于 encoder-decoder 的一种[98,99], 经典 Seq2Seq 的基本思想是利用两个 LSTM, 分别作为 encoder 和 decoder. Bahdanau 等人[100] 引入注意力机制来解决长句子任务中信息丢失的问题, 取得了很好的效果. Sordoni 等人[101] 将具体的 context 加入模型, 提升了系统的智能性. Kumar 等人[102] 引入动态记忆网络 (dynamic memory network, DMN) 处理输入序列和问题, 形成情景记忆并生成相关答案. Seq2Seq 会产生一些无意义的回答, 例如“I don’t know”, 对于这个问题, 哈尔滨工业大学 Xu 等人[103] 在 2017 提出一种基于生成对抗网络 (generative adversarial networks, GAN) 的模型, 其在 Seq2Seq 框架基础上加入具有对抗功能的鉴别器, 来判断人与机器回答的区别, 以此提高回答的多样性. 近些年, 随着 BERT 等预训练模型的出现, NLP 领域开始关注语言模型预训练和少量数据下游任务调优的开发模式, 在答案生成的相关任务方面微软于 2019–2020 年期间, 相继提出 MASS[104]、UNILM[105,106] 和 DialoGPT 模型[107],

在相关任务中都取得了非常好的成绩.

生成式问答系统优点是不依赖数据库, 比较灵活. 缺点也很明显, 不能完全满足用户的提问需求, 可能生成一些无意义和重复的回答等.

4 结束语

从限定域到开放域; 从单数据源到多数据源; 从单维信息输入到多维信息输入, 随着科学技术不断发展, 问答系统也愈发的成熟. 本文大体介绍了问答系统的发展轨迹、相关数据集和部分评测指标, 且将问答系统分类并简述各类问答系统的相关技术.

近些年, 深度学习的革命性发展给问答系统带来了长足的进步, 序列到序列的模型, 端到端的模型以及最近流行的预训练, 都给问答系统留下无限的发展空间. BERT 等一系列预训练模型刷榜了多项 NLP 任务, 也为问答系统扩展了新的发展方向. 当前问答系统大多数都只用于单一的特定领域, 如“医疗”等, 如何处理多领域多语言的问答系统也是问答系统走向智能的一大趋势[108].

除此之外, 单一模态的问答系统具有局限性, 现实场景中大多数都是多模态的, 因此越来越多的研究者关注将视觉和语音融合等多模态问答系统. 在问答系统的一些基础性 NLP 任务中, 结合字形[59,60]、拼音[109]等多源信息已成为提高识别准确率的方法之一. 不少研究者在字词嵌入向量的基础上, 结合 POS 标签等语法特征来提取不同粒度的信息[110], 以此来提高模型提取文本语义信息的能力.

尽管深度学习技术获得了快速的发展, 但目前仍然存在着一些挑战: 如何避免对话前后不一致; 如何实现长期多轮对话; 如何对多模态信息进行合理融合; 如何产生丰富灵活的回复、避免产生无意义的回复等. 这些挑战都是当下问答系统亟待解决的问题和未来需努力的方向. 总之, 问答系统越来越智能化, 但离问答系统像人一样解答问题或是自由地进行对话, 还是有很长一段路要走.

参考文献

1 毛先领, 李晓明. 问答系统研究综述. 计算机科学与探索, 2012, 6(3): 193–207. [doi: 10.3778/j.issn.1673-9418.2012.03.001]

2 Turing AM, Haugeland J. Computing machinery and intelligence. The Turing Test: Verbal Behavior as the Hallmark of Intelligence, 1950: 29–56.

3 Green Jr BF, Wolf AK, Chomsky C, *et al*. Baseball: An automatic question-answerer. Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference. Los Angeles: ACM, 1961. 219–224.

4 Woods W A. Lunar rocks in natural English: Explorations in natural language question answering. In: Zampolli A, ed. Linguistic Structures Processing. New York: North Holland, 1977.

5 Wilensky R. The Berkeley UNIX consultant project. In: Brauer W, Wahlster W, eds. Wissensbasierte Systeme. Berlin: Springer, 1987. 286–296.

6 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2013. 3111–3119.

7 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543.

8 Devlin J, Chang MW, 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: Association for Computational Linguistics, 2019. 4171–4186.

9 Zhu YK, Kiros R, Zemel R, *et al*. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 19–27.

10 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

11 Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 2227–2237.

12 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. https://

cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. 2018.

13 Lan ZZ, Chen MD, Goodman S, *et al*. ALBERT: A lite BERT for self-supervised learning of language representations. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

14 Joshi M, Chen DQ, Liu YH, *et al*. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 2020, 8: 64–77. [doi: 10.1162/tacl_a_00300]

15 Yang ZL, Dai ZH, Yang YM, *et al*. XLNet: Generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: ACM, 2019. 5753–5763.

16 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

17 Sun Y, Wang SH, Li YK, *et al*. ERNIE: Enhanced representation through knowledge integration. arXiv:1904.09223, 2019.

18 Zhang ZY, Han X, Liu ZY, *et al*. ERNIE: Enhanced language representation with informative entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1441–1451.

19 Clark K, Luong MT, Le QV, *et al*. Electra: Pre-training text encoders as discriminators rather than generators. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

20 Liu WJ, Zhou P, Zhao Z, *et al*. K-BERT: Enabling language representation with knowledge graph. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2901–2908. [doi: 10.1609/aaai.v34i03.5681]

21 Rajpurkar P, Zhang J, Lopyrev K, *et al*. SQuAD: 100000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 2382–2392.

22 Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 784–789.

23 Choi E, He H, Iyyer M, *et al*. QuAC: Question answering in context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2174–2184.

24 Reddy S, Chen DQ, Manning CD. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7: 249–266. [doi: 10.1162/tacl_a_00266]

25 Yatskar M. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 2318–2323.

26 Ren MY, Kiros R, Zemel RS. Exploring models and data for image question answering. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015. 2953–2961.

27 Antol S, Agrawal A, Lu JS, *et al*. VQA: Visual question answering. Proceedings of the 2015 IEEE international Conference on Computer Vision. Santiago: IEEE. 2015. 2425–2433.

28 Minaee S, Kalchbrenner N, Cambria E, *et al*. Deep learning-based text classification: A Comprehensive Review. ACM Computing Surveys, 2022, 54(3): 62.

29 Nguyen T, Rosenberg M, Song X, *et al*. MS MARCO: A human generated machine reading comprehension dataset. Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016). Barcelona: CEUR-WS.org, 2016.

30 Yang Y, Yih WT, Meek C. WikiQA: A challenge dataset for open-domain question answering. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 2013–2018.

31 Trischler A, Wang T, Yuan XD, *et al*. NewsQA: A machine comprehension dataset. Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver: Association for Computational Linguistics, 2017. 191–200.

32 Wu CS, Madotto A, Liu WH, *et al*. QAConv: Question answering on informative conversations. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational

Linguistics, 2022. 5389–5411.
33 Cui YM, Liu T, Chen ZP, *et al*. Consensus attention-based neural networks for Chinese reading comprehension. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 1777–1786.
34 Li P, Li W, He ZY, *et al*. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. arXiv:1607.06275, 2016.
35 He W, Liu K, Liu J, *et al*. DuReader: A Chinese machine reading comprehension dataset from real-world applications. Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne: Association for Computational Linguistics, 2018. 37–46.
36 Cui YM, Liu T, Chen ZP, *et al*. Dataset for the first evaluation on Chinese machine reading comprehension. Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki: ELRA, 2018.
37 Cui YM, Liu T, Che WX, *et al*. A span-extraction dataset for Chinese machine reading comprehension. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 5883–5889.
38 Cui YM, Liu T, Yang ZQ, *et al*. A sentence cloze dataset for Chinese machine reading comprehension. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 6717–6723.
39 Levow GA. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics, 2006. 108–117.
40 Peng NY, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 548–554.
41 Shivakumar PG, Yang M, Georgiou PG. Spoken language intent detection using confusion2Vec. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 819–823.
42 Coucke A, Saade A, Ball A, *et al*. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. arXiv:1805.10190, 2018.
43 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACM, 2002. 311–318.
44 Lin CY. ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.
45 牛玉磊, 张含望. 视觉问答与对话综述. 计算机科学, 2021, 48(3): 87–96.
46 李涛, 王次臣, 李华康. 知识图谱的发展与构建. 南京理工大学学报, 2017, 41(1): 22–34. [doi: 10.14177/j.cnki.32-1397n.2017.41.01.004]
47 袁国铭, 李洪奇, 樊波. 关于知识工程的发展综述. 计算技术与自动化, 2011, 30(1): 138–143.
48 Li HY, Qu YZ. KREAG: Keyword query approach over RDF data based on entity-triple association graph. Chinese Journal of Computers, 2011, 34(5): 825–835. [doi: 10.3724/SP.J.1016.2011.00825]
49 邢立栋. 面向特定领域的知识图谱构建技术研究与应用 [硕士学位论文]. 北京: 北京化工大学, 2018.
50 Rau LF. Extracting company names from text. Proceedings of the 7th IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991. 29–30.
51 Humphreys K, Gaizauskas R, Azzam S, *et al*. Description of the LaSIE-II system as used for MUC-7. Proceedings of the 7th Message Understanding Conference (MUC-7). Fairfax, 1998.
52 Bikel DM, Schwartz R, Weischedel RM. An algorithm that learns what's in a name. Machine Learning, 1999, 34(1): 211–231.
53 Ratnaparkhi A. Maximum entropy models for natural language ambiguity resolution [Ph.D. Thesis]. Philadelphia: University of Pennsylvania, 1998.
54 Bordes A, Weston J, Collobert R, *et al*. Learning structured embeddings of knowledge bases. Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2011. 301–306.
55 孙娟娟, 于红, 冯艳红, 等. 基于深度学习的渔业领域命名实体识别. 大连海洋大学学报, 2018, 33(2): 265–269. [doi: 10.16535/j.cnki.dlhyxb.2018.02.020]
56 陈鹏, 蔡冰, 何晓勇, 等. 面向电力规章制度的命名实体识别. 计算机系统应用, 2022, 31(6): 210–216. [doi: 10.15888/j.cnki.csa.008525]
57 Huang CY, Chen YG, Liang QC. Attention-based

bidirectional long short-term memory networks for Chinese named entity recognition. Proceedings of the 2019 4th International Conference on Machine Learning Technologies. Nanchang: ACM, 2019. 53–57.

58 陈彦妤. 健康保险智能问答问句理解和答案检索的研究与实现 [博士学位论文]. 上海: 东华大学, 2018.

59 Xuan ZY, Bao R, Jiang SY. FGN: Fusion glyph network for Chinese named entity recognition. Proceedings of the 5th China Conference on Knowledge Graph and Semantic Computing. Nanchang: Springer, 2020. 28–40.

60 Meng YX, Wu W, Wang F, *et al*. Glyce: Glyph-vectors for Chinese character representations. Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver: ACM, 2019. 32.

61 Miller S, Crystal M, Fox H, *et al*. BBN: Description of the SIFT system as used for MUC-7. Proceedings of the 7th Message Understanding Conference (MUC-7). Fairfax, 1998.

62 Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. Proceedings of the ACL Interactive Poster and Demonstration Sessions. Barcelona: Association for Computational Linguistics, 2004. 178–181.

63 Socher R, Huval B, Manning CD, *et al*. Semantic compositionality through recursive matrix-vector spaces. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012. 1201–1211.

64 Zeng DJ, Liu K, Lai SW, *et al*. Relation classification via convolutional deep neural network. Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: Dublin City University and Association for Computational Linguistics, 2014. 2335–2344.

65 Nguyen TH, Grishman R. Relation extraction: Perspective from convolutional neural networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: Association for Computational Linguistics, 2015. 39–48.

66 Zheng SC, Hao YW, Lu DY, *et al*. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing, 2017, 257: 59–66. [doi: 10.1016/j.neucom.2016.12.075]

67 Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1105–1116.

68 Hermann KM, Kočiský T, Grefenstette E, *et al*. Teaching machines to read and comprehend. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015. 1693–1701.

69 Joshi M, Choi E, Weld DS, *et al*. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 1601–1611.

70 Seo MJ, Kembhavi A, Farhadi A, *et al*. Bidirectional attention flow for machine comprehension. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

71 Wang S, Jiang J. Machine comprehension using match-LSTM and answer pointer. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

72 Yu AW, Dohan D, Luong MT, *et al*. QANet: Combining local convolution with global self-attention for reading comprehension. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

73 ElFadeel H, Peshterliev S. Robustly optimized and distilled training for natural language understanding. arXiv:2103.08809, 2021.

74 Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 1112–1122.

75 Zhang ZS, Yang JJ, Zhao H. Retrospective reader for machine reading comprehension. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14506–14514. [doi: 10.1609/aaai.v35i16.17705]

76 Shriberg E, Wade E, Price P. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. Proceedings of Speech and Natural Language. Harriman: HLT, 1992. 49–54.

77 Den Os E, Boves L, Lamel L, *et al*. Overview of the ARISE project. Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest: ISCA, 1999. 1527–1530.

78 Ehrlich U, Hanrieder G, Hitzenberger L, *et al*. Access-automated call center through speech understanding system. Proceedings of the 5th European Conference on Speech Communication and Technology. Rhodes: ISCA, 1997.

79 Stolcke A, Ries K, Coccaro N, *et al*. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 2000, 26(3): 339–373. [doi: 10.1162/089120100561737]

80 Lendvai P, Geertzen J. Token-based chunking of turn-internal dialogue act sequences. Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue. Antwerp: Association for Computational Linguistics, 2007. 174–181.

81 Zimmermann M. Joint segmentation and classification of dialog acts using conditional random fields. Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton: ISCA, 2009. 864–867.

82 Lee JY, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 512–520.

83 Cherng HE, Chang CH. Dialogue quality and nugget detection for short text conversation (STC-3) based on hierarchical multi-stack model with memory enhance structure. Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo, 2019.

84 Song J, Luo QF, Nie JC. Research and application of multi-round dialogue intent recognition method. Proceedings of the 2020 16th International Conference on Computational Intelligence and Security (CIS). Guangxi: IEEE, 2020. 131–135.

85 Liu D, Zhao Z, Gan LD. Intention detection based on BERT-BiLSTM in taskoriented dialogue system. Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing. Chengdu: IEEE, 2019. 187–191.

86 Katz B, Felshin S, Lin J, *et al*. Viewing the web as a virtual database for question answering. New Directions in Question Answering. Cambridge: MIT Press, 2004. 215–226.

87 Zheng ZP. AnswerBus question answering system. Proceedings of the 2nd International Conference on Human Language Technology Research. San Diego: ACM, 2002. 399–404.

88 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620. [doi: 10.1145/361219.361220]

89 Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 2010, 1(1–4): 43–52.

90 Ramos J. Using TF-IDF to determine word relevance in document queries. Proceedings of the 1st Instructional Conference on Machine Learning. 2003. 29–48.

91 Zhou XQ, Hu BT, Chen QC, *et al*. Answer sequence learning with neural networks for answer selection in community 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: Association for Computational Linguistics, 2015. 713–718.

92 Tan M, Dos Santos C, Xiang B, *et al*. Improved representation learning for question answer matching. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin: Association for Computational Linguistics, 2016. 464–473.

93 Huang PS, He XD, Gao JF, *et al*. Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco: ACM, 2013. 2333–2338.

94 Shen YL, He XD, Gao JF, *et al*. Learning semantic representations using convolutional neural networks for web search. Proceedings of the 23rd International Conference on World Wide Web. Seoul: ACM, 2014. 373–374.

95 Mitra B, Diaz F, Craswell N. Learning to match using local and distributed representations of text for Web search. Proceedings of the 26th International Conference on World Wide Web. Perth: ACM, 2017. 1291–1299.

96 Wan SX, Lan YY, Guo JF, *et al*. A deep architecture for semantic matching with multiple positional sentence representations. Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016. 2835–2841.

97 Wan SX, Lan YY, Xu J, *et al*. Match-SRNN: Modeling the

recursive matching structure with spatial RNN. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: IJCAI/AAAI Press, 2016.

98 Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1724–1734.

99 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: ACM, 2014. 3104–3112.

100 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2014.

101 Sordoni A, Galley M, Auli M, *et al*. A neural network approach to context-sensitive generation of conversational responses. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver: Association for Computational Linguistics, 2015. 196–205.

102 Kumar A, Irsoy O, Ondruska P, *et al*. Ask me anything: Dynamic memory networks for natural language processing. Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, 2016. 1378–1387.

103 Xu Z, Liu BQ, Wang BX, *et al*. Neural response generation via GAN with an approximate embedding layer. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 617–626.

104 Song KT, Tan X, Qin T, *et al*. MASS: Masked sequence to sequence pre-training for language generation. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 5926–5936.

105 Dong L, Yang N, Wang WH, *et al*. Unified language model pre-training for natural language understanding and generation. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: ACM, 2019. 13063–13075.

106 Bao HB, Dong L, Wei FR, *et al*. UniLMv2: Pseudo-masked language models for unified language model pre-training. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 642–652.

107 Zhang YZ, Sun SQ, Galley M, *et al*. DialoGPT: Large-scale generative pre-training for conversational response generation. arXiv:1911.00536, 2020.

108 Soares MAC, Parreiras FS. A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and Information Sciences, 2020, 32(6): 635–646. [doi: 10.1016/j.jksuci.2018.08.005]

109 Sun ZJ, Li XY, Sun XF, *et al*. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. 2065–2075.

110 Nie YY, Tian YH, Song Y, *et al*. Improving named entity recognition with attentive ensemble of syntactic information. Proceedings of the Findings of the Association for Computational Linguistics (EMNLP 2020). Association for Computational Linguistics, 2020. 4231–4245.

(校对责编: 孙君艳)