智能对话系统研究综述

贾熹滨[1]，李 让[1]，胡长建[2]，陈军成[1]
(1. 北京工业大学信息学部，北京 100124; 2. 联想集团北京研究院，北京 100085)

摘 要: 智能对话系统作为人工智能领域的核心技术，即将成为新的和谐人机交互方式，具有重大的研究意义和应用价值，因此，较为全面、深入地总结了深度学习模型在对话生成中的应用及对话系统领域的研究进展和现状. 首先，阐述了智能对话系统的类型划分，介绍系统的模块框架构成，包括各模块的主要研究问题与关键技术的主流思路和研究现状; 然后，从理论模型、研究进展及可用性等角度深度剖析了现有的对话生成技术解决方案，重点分析了应用于自然语言生成的序列到序列的神经网络结构及搭建原理; 最后，对存在的问题进行总结，并展望了未来的研究方向.

关键词: 对话系统; 深度学习; 人工智能; 人机交互
中图分类号: TP 391 **文献标志码:** A **文章编号:** 0254 - 0037(2017)09 - 1344 - 13
doi: 10.11936/bjutxb2016090023

Review of Intelligent Dialogue System

JIA Xibin[1], LI Rang[1], HU Changjian[2], CHEN Juncheng[1]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Lenovo Corporate Beijing Research & Development, Beijing 100085, China)

Abstract: Intelligent dialogue system is the core technology in the field of artificial intelligence, and it will act as a new harmonious human-computer interaction. The research has great theory significance and application value. The status and advances of dialogue system and the application of deep learning model in dialogue generation were comprehensively summarized in this paper. First, the type of intelligent dialogue system was elaborated. The module framework of the system was introduced, including the research contents of the modules and key components of the main technology ideas and their research status. Then, the various solutions to dialogue generation were systematically analyzed from the perspective of theoretical models, advances and usability. The neural network structures and the construction principle of sequence-to-sequence applied to natural language generation were analyzed emphatically. Finally, the problems and prospects for future research directions were summarized.

Key words: dialogue system; deep learning; artificial intelligence; human-computer interaction

对话系统并不是一个新兴的概念，自Weizenbaum于20世纪60年代在麻省理工学院开发Eliza[1]开始，它被认为是第一个对话系统，可以和人进行简单对话，但得到的回复大部分是“What are you saying about…”. 微软在20世纪90年代为Office软件配备的虚拟助手Clippy可以认为是最早

收稿日期: 2016-09-08
基金项目: 国家自然科学基金资助项目(61370113); 北京市自然科学基金资助项目(4152005); 北京市教育委员会重点资助项目(KZ201610005009)
作者简介: 贾熹滨(1969—)，女，教授，主要从事视觉图像处理、认知方面的研究，E-mail: jiaxibin@bjut.edu.cn
通信作者: 陈军成(1980—)，男，讲师，主要从事大数据、软件测试与分析方面的研究，E-mail: juncheng@bjut.edu.cn

被大规模推向市场的对话系统，它以对话的方式为用户在使用 Office 工具的过程中提供帮助，但收效甚微，而且很多用户评价它“冒冒失失令人讨厌”，致使微软不得不将其下线. 进入 21 世纪，Wallace 通过一种纯模式匹配的方法，利用启发式的会话规则并嵌入 AIML（artificial intelligence markup language）[2] 创造出了 A. L. I. C. E（artificial linguistic internet computer entity）[3]，这个对话系统在对话质量上与以往的研究工作相比有大幅提高，并 3 次获得了罗纳奖（Leobner prize）. 但这样的进步并不足以使对话系统通过图灵测试[4]，任何人通过几轮对话就能发现其中的破绽，或答非所问，或前后矛盾. 所以过去智能对话系统仅仅是作为一项有趣的、半科幻的且不太成熟的玩具而存在.

而进入 2016 年后，短短几个月的时间内，科技行业巨头微软、Facebook、亚马逊、Google 和苹果纷纷发布了各自在智能对话领域的战略和相关产品，创业公司也层出不穷，对话系统迅速成为了科技媒体和开发者社区讨论的焦点. 究其原因，是网络生态的发展和技术的变革让智能对话变得炽手可热.

过去几年中，消息服务类应用迅速壮大，国内的微信，国外的 WhatsApp、Facebook Messenger 等，几乎占领了用户所有的碎片时间，活跃用户数以亿计，在事实上成为了移动互联网时代的“浏览器”入口. 用户只需使用一个应用就能获取大部分信息，不需要把注意力分散到其他类应用上，下载移动应用所带来的流量红利正在慢慢消失. 这时对话系统的优点被体现出来，开发成本低，又可以依附大的软件平台，忽略应用下载量和活跃量的问题，成为消息服务上最自然的应用，而且在移动时代成长起来的用户易接受即时消息通讯的方式，进入门槛低、黏性高. 因此，对话系统或许可以取代移动应用构建自己的生态环境.

在技术方面，2006 年以来深度学习的飞速进展让多层神经网络得以在计算机视觉和语音识别领域取得突破性进展，让人工智能在原本人类擅长的领域中表现得更为优异. 深度学习的成果将拓展到更多的领域研究中，包括自然语言处理（natural language processing，NLP），对话系统可以借此突破现在的一些瓶颈.

市场生态和技术水平沿着各自的轨道向前发展，在这个时间点上交汇到一起，使得智能对话系统领域引来巨头们争相投入. 也许在不久的未来，自然语言会代替输入设备和触摸屏成为最广泛使用的人机交互界面.

1　对话系统类型

对话系统可以根据应用场景的不同分为开放域（open-domain）问题和封闭域（closed-domain）问题 2 种类型.

开放域对话系统没有任何限定的主题或明确的目标，用户和系统之间可以进行任何话题的自由对话，这要求系统具备丰富的知识，能完成多项任务，同时具有社会性（友好度、自觉性、幽默感等），这在技术上短期内难以实现. 已经实现的这类对话系统更多地应用在聊天机器人、虚拟形象等泛娱乐领域，用户基数大且容易传播，但由于对话质量不高或内容深度不够等问题，使得用户黏性并不高，应用前景也比较模糊.

封闭域的对话系统是面向具体任务（task-oriented）的，具有明确的目标和限定的知识范围，只需专注完成一项工作，输入和输出有限，实现起来相对简单，在垂直使用场景中更有助于节省人力成本或提升人工效率，但是在具体任务中对话内容的容错率更低，且对话数据规模很小，难以通过数据驱动（data-driven）的方式训练模型，需要人为整理知识库或对话模板，耗费大量人力的同时又难以向平行任务领域迁移；所以，建立这样的对话系统需要从商业、产品、运营、数据知识积累和模型调优等各方面综合考量.

2　对话系统模块框架

对话系统是一个综合性问题，主要涉及自动语音识别（automatic speech recognition，ASR）、自然语言理解（natural language understanding，NLU）、对话状态跟踪（dialog state tracking，DST）、自然语言生成（natural language generation，NLG）和语音合成（text to speech，TTS）五部分，一个完整的人机对话流程如图 1 所示.

系统中从输入到输出的每个模块都是一个独立的有价值的重要研究课题，综合了多种理论和技术[5-6]. 这些理论和技术一直在发展和进步，但还远没有达到成熟的地步，仍具有广阔的研究空间. 在实际开发中，一个完整的系统并不是各个模块之间的简单拼凑，而是需要将各个部分有效地结合在一起，最大程度地提升整体性能.

2.1　自动语音识别

自动语音识别的功能是将用户语音中的连续时

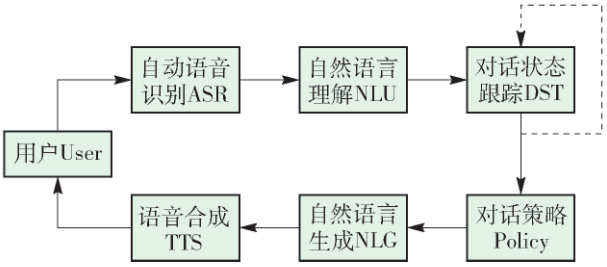

图 1 对话系统模块框架

Fig. 1 Dialogue system module framework

间信号转变为离散的音节单元或单词. 在口语对话系统中,用户语音存在大量的口语现象,通常还伴随着环境噪声,这些为识别算法的特征提取、模型训练等环节增加了更多难度. 作为系统输入点,语音识别的准确与否将直接影响后续语言理解和整个系统的性能优劣. 提高语音识别的准确率是研发对话系统的关键问题之一.

2.2 自然语言理解

自然语言理解的功能是利用语义和语法分析将语音识别的结果转化为计算机能够理解的结果化表现形式. 对话系统中的语言理解最常用的 2 个关键方法是文本分类(text classification)和序列标注(sequence labeling).

2.2.1 文本分类

文本分类的目的是根据预先定义的主题类别,按照一定的规则为未知类别的文本自动确定一个类别. 对话系统中,通过文本分类的方法,将用户的自然语言根据涉及领域的不同分为几类,以判断用户的意图.

目前应用在文本分类领域传统的机器学习模型有 k 近邻(k-nearest neighbor,kNN)[7]、朴素贝叶斯(naive Bayesian)[8]、支持向量机(support vector machines,SVM)[9]等. 神经网络兴起后,一些基于深度学习模型的研究也取得了不俗的效果,主要是运用了卷积神经网络(convolutional neural networks,CNN)模型[10-11],以及 CNN 和循环神经网络(recurrent neural networks,RNN)相结合的模型[12-13].

2.2.2 序列标注

序列标注模型被广泛应用于文本处理相关领域,以得到自然语言序列对应的标签序列. 对话系统中利用序列标注的方法对自然语言序列进行分词(word segmentation)、词性标注(part-of-speech tagging)、命名实体识别(named entity recognition,NER)等工作,得到标签序列后生成结构化的数据,便于对整个句子进行理解.

序列标注领域中常用的模型有隐马尔科夫(hidden Markov model,HMM)[14]、最大熵马尔科夫(maximum entropy Markov model,MEMM)[15]、条件随机场(conditional random fields,CRF)[16-17]等. 与传统的方法相比,深度学习中常用的长短期记忆网络(long-short term memory,LSTM)模型[18-19]也取得了相近的效果. 另外,将神经网络模型自动提取特征的优点与机器学习模型计算联合概率的方式结合在一起,得到了 LSTM + CRF 模型[20-22],这也是目前学术界常用的做法.

口语对话中句式灵活多变,且用户有不同的语言习惯,再加上语音识别本身出现的错误,语言理解模块同样面临着重大挑战.

2.3 对话状态跟踪

对话状态跟踪的作用在于通过语言理解生成的结构化数据理解或者捕捉用户的意图或目标,目前应用在这个领域的模型很多,包括有限状态机、填槽法、基于实例方法、基于规划方法、贝叶斯网络等,文献[23-24]对其中一部分方法进行了分析和比较.

对话状态跟踪的思想是将系统和用户交互时的行为看作是在填写一张记录用户当前对话状态的表格. 以订机票为例,将这张表格预先设定好状态,比如目的地、出发地、出发时间等,与系统背后的业务数据表中的属性相关联,不断地从对话中抽取相应的值来填充这个表格. 这是一个利用监督学习(supervised learning)完成的多分类任务,根据对话的分类结果判断这句话中包括哪些状态和值. 往往从一句对话中获取所有的状态只是理想情况,当状态表中的信息存在空白时,系统会根据空白的状态来提问并获取对应的值,直到获取到足够的状态,给出对用户的建议,或者进行相应的服务. 对话状态跟踪如图 2 所示.

图 2 对话状态跟踪

Fig. 2 Dialogue state tracker

2.4　自然语言生成

自然语言生成的作用是组织适当的应答语句，将系统的答复转换成用户能够理解的自然语言，通常有 3 种解决方案：基于人工模板（rule-based）、基于知识库检索（query-based）和基于深度学习的序列到序列（Sequence-to-Sequence）生成模型．语言生成方案的优缺点和适用场景总结如表 1 所示．

表 1　语言生成方案特点

Table 1　Characteristics of language generation solutions

方案	优点	缺点	适用场景
rule-based	具体领域内回答精准	可移植性和可拓展性差	个人助理
query-based	知识库易扩充，答案没有语法错误	对话连续性差，出现所答非所问	娱乐聊天、问答系统
Sequence-to-Sequence	数据驱动，省去语言理解等过程	需要大量语料支持	虚拟形象、智能聊天机器人

自然语言生成是对话系统的核心内容，本文将在第 3 节中对其中不同的解决方法进行详细介绍．

2.5　语音合成

语音合成的功能是将系统答复的自然语言文本合成应答语音反馈给用户．其主要难点在于如何使生成的语音更加自然生动、需要的语音数据库更小及成本更低．

3　自然语言生成主要解决方案

3.1　基于人工模板

基于人工模板的技术通过人工设定对话场景，并对每个场景编写针对性的对话模板，回复的最终形式是填充一个大多数的内容已经给定的模板，只有一些具体的参数需要填充．这个技术路线的好处是回答精准，缺点是过多的人工标注和模板编写工作导致移植性差，需要逐个场景去扩展．

2010 年，苹果公司在 iPhone4S 上推出的个人语音助理 Siri，在 NLG 环节采用了基于人工模板的技术路线．Siri 里包含的众多数据、模型和计算模块，可以划分为输入系统、活跃本体、执行系统、服务系统和输出系统 5 个子系统．Siri 整体架构如图 3 所示．

活跃实体内存放的数据和模型包括：领域模型、用户个性化信息、语言模式、词汇表和领域实体数据库等．其中领域模型包括某个垂直领域内的概念、实体、关系、属性和实例的内部表示，即本体（ontology）．词汇表用于维护 Siri 中的表层单词到领域模型或者任务模型中定义的概念、关系、属性的映射关系，即人工编写的特定领域的对话模板．执行系统将用户的文本表示解析为内部用户意图之后调用活跃本体中的数据拼装对话，引导用户输入和生成输出结果．

通过这种方法，Siri 根据某几个垂直领域的领

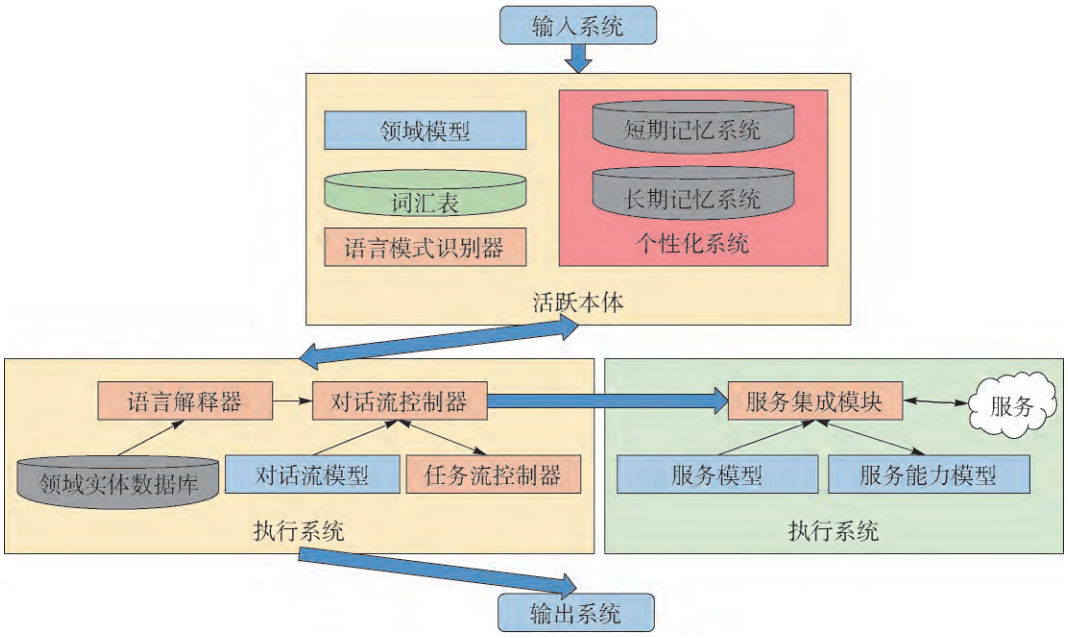

图 3　Siri 整体架构

Fig. 3　Overall architecture of Siri

域模型和人工编写的对话模板完成固定的任务，通过引导用户输入能得到很高的成功率，而对于解决不了的问题，Siri 会直接调用搜索引擎.

3.2 基于知识库检索

基于知识库检索的技术路线与搜索引擎类似，预先准备好一个称为知识库的数据库，里面包含丰富的对话资料，对其中的问题建立索引，然后以 NLP 技术对用户提出的问题进行分析，通过关键词提取、倒排索引、文档排序等方法与定义好的知识库进行模糊匹配，找到最合适的应答内容. 这类解决方案的核心技术在于找更多的数据来丰富和清洗知识库，但数据量过大时难以监督，通常找来的数据杂乱无章，使对话连续性很差.

2011 年，IBM 推出了电脑问答(Q&A)系统 Watson [25-26]，在 NLG 部分采用了以知识库检索技术为基础，集高级自然语言处理、知识图谱、自动推理、机器学习等开放式问答技术 [27-28] 为一体的技术思路，通过假设认知与大规模的证据搜集、分析和评价得出最终答案.

Watson 的 DeepQA 架构以处理流程的形式定义了分析问题的各个步骤，如环形办公室走廊，每间办公室都负责一项特殊的工作，允许多重实现来产生多个结果. 每一间办公室的工作都当作大规模并行计算的一部分而单独进行，如图 4 所示.

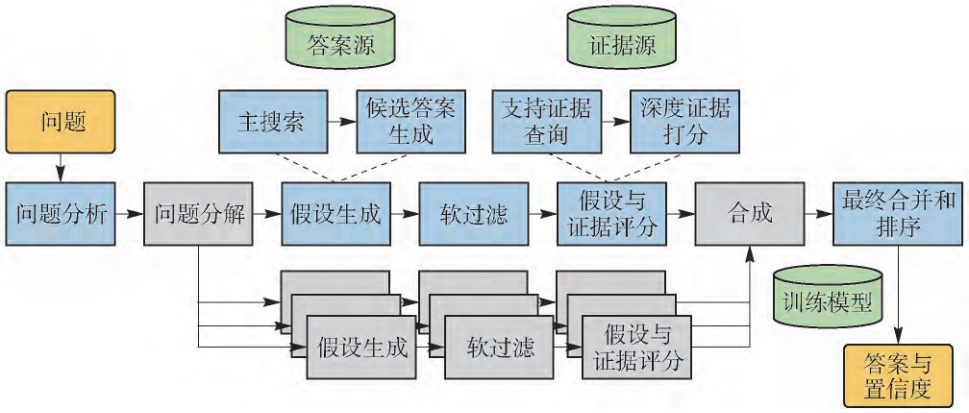

图 4 Watson 工作流程

Fig. 4 Workflow of Watson

系统在基于对问题和类型的不同理解上对多个不同的资源进行检索，返回多种候选答案 [29-30]. 任何答案都不会立即被确定，因为随着时间推移系统会收集到越来越多的证据来分析每一个答案和每一条不同的道路. 之后系统用几百种不同的算法从不同的角度分析证据得出上百种特征值或得分，这代表着在某一特定维度上一些证据支持一个答案程度，每个答案的所有特征值或得分综合为一个得分，表示该答案正确的概率. 系统通过统计学机器学习方法对大量数据集进行学习来确定各个特征值的权重 [31]，最终将得分排名最高的答案输出.

DeepQA 技术根据一个问题通过搜索和量化评估给出一个确定的答案，通过知识库的扩张和切换可以很好地完成 Q&A 任务，但这种形式结构还是无法高效地跟上源知识的增长和领域的切换，也没能与用户进行有效的互动，无法在大量的非结构化内容支持下为用户提供决策.

3.3 基于深度学习

基于深度学习的技术通常不依赖于特定的答案库或模板，而是依据从大量语料中习得的语言能力来进行对话，根据问题内容直接生成回答的方法被定义为基于某个条件下的生成模型. 深度学习的 Sequence-to-Sequence 技术 [32-33] 可以非常好地实现生成模型的框架，其最诱人的优势就是可以避免人为特征工程的端到端(End-to-End)框架，即利用强大的计算和抽象能力自动从海量的数据源中归纳、抽取对解决问题有价值的知识和特征，使这一过程对于问题的解决者来说透明化，从而规避人为特征工程所带来的不确定性和繁重的工作量.

目前，基于深度学习技术的对话系统多数采用了 Encoder-Decoder 框架，本节将首先描述 Encoder-Decoder 框架技术原理，然后分别针对深度学习应用在开发实践时需要特殊考虑的主要问题及其对应的解决方案进行阐述，模型改进思路如图 5 所示.

一些较早的文献 [34-35] 采用 Encoder-Decoder 模型来建立对话系统，收集 Twitter 或者微博中评论里的聊天信息来作为训练数据，得到了效果不错的对话系统. 之后其他文献在 Encoder [36-38]、目标函

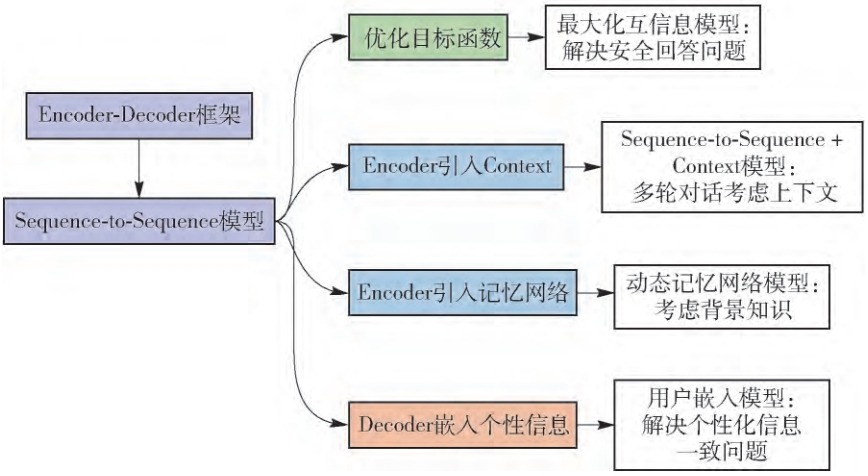

图 5　模型改进思路

Fig. 5　Solutions of model improvement

数[39]和 Decoder[40]部分不断改进模型，有效解决了在应用深度学习技术时遇到的一些问题.

3.3.1　Encoder-Decoder 框架

Encoder-Decoder 框架是一种文本领域的研究模式，图 6 是文本处理领域里常用的 Encoder-Decoder 框架的抽象表示.

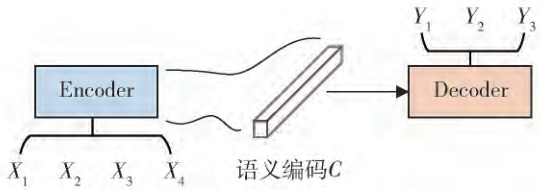

图 6　Encoder-Decoder 框架

Fig. 6　Encoder-Decoder framework

Encoder-Decoder 框架适合处理由一个句子(或篇章)生成另外一个句子(或篇章)的通用处理模型. 对于句子对 $\langle X, Y\rangle$，目标是给定输入句子 X 生成目标句子 Y. X 和 Y 可以是同一种语言，也可以是 2 种不同的语言. 而 X 和 Y 分别由各自的单词序列构成，即

$$X = \langle x_1, x_2, \cdots, x_n\rangle$$

$$Y = \langle y_1, y_2, \cdots, y_m\rangle$$

Encoder 顾名思义就是对输入句子 X 进行编码，将输入句子通过非线性变换转化为中间语义表示，即

$$C = \mathcal{A}(y_1, y_2, \cdots, y_m)$$

对于解码器 Decoder 来说，其任务是根据句子 X 的中间语义表示 C 和之前已经生成的历史信息 y_1，$y_2, \cdots, y_{i-1}$ 来生成 i 时刻要生成的单词，即

$$y_i = g(C, y_1, y_2, \cdots, y_{i-1})$$

对话系统使用 Encoder-Decoder 框架来解决自然语言生成问题时，X 指的是用户输入语句，一般称作 Message，而 Y 一般指的是聊天机器人的应答语句，一般称作 Response. 其含义是当用户输入 Message 后，经过 Encoder-Decoder 框架计算，首先由 Encoder 对 Message 进行语义编码，形成中间语义表示 C，Decoder 根据 C 生成了聊天机器人的应答 Response. 这样，用户反复输入不同的 Message，聊天机器人每次都形成新的应答 Response，形成了一个实际的对话系统.

实现开发时，对话系统的 Sequence-to-Sequence 常采用 RNN 模型和 RNN 的改进模型 LSTM 来实现. LSTM 解决了 RNN 的长期依赖(long-term dependencies)问题，对于较长的线性序列会取得更好的效果.

3.3.2　Sequence-to-Sequence 模型

文献[34]中，Google 收集了一个 IT 问题解答领域的问答数据集和大量有噪音的电影字幕数据集，采用完全 data-driven 的方式训练 2 个 LSTM 中对应的神经网络连接参数，得到了一个对话系统，模型抽象结构如图 7 所示.

图中 A、B、C 是 Message 中的单词序列，用来生成 Response 的单词序列 W、X、Y、Z，为方便计算，单词用分布式表示(distributed representations)的方式转换为词向量(word embedding)[41-45].

Google 用此系统与一个 rule-based 对话系统 CleverBot 进行了人工评价的对比测试，CleverBot 在垂直领域解决问题的效果更好，而在 open-domain 的对话中生成模型给出的答案得到了更多人的认

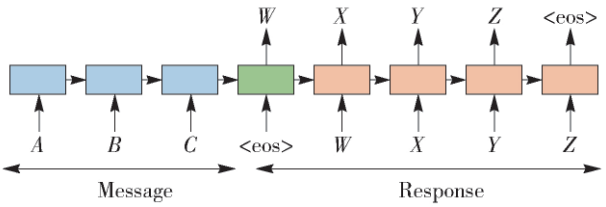

图 7 基于 Sequence-to-Sequence 框架的模型抽象结构

Fig. 7 Using Sequence-to-Sequence framework for modeling conversations

可. 一些测试结果如表 2 所示.

表 2 CleverBot 与 Neural Conversational Model 生成结果

Table 2 Results generated by CleverBot and Neural Conversational Model

CleverBot
Human: What do you think about tesla?
CleverBot: I am tesla.
Human: What do you think about cleopatra?
CleverBot: Cool is cool.
Human: What do you think about england during the reign of elizabeth?
CleverBot: I like him. Do you?
Human: What do you think about bill gates?
CleverBot: Lets talk about Link.
Neural Conversational Model
Human: What do you think about tesla?
NCM: He's a good conductor.
Human: What do you think about bill gates?
NCM: He's a good man.
Human: What do you think about messi?
NCM: He's a great player.
Human: What do you think about cleopatra?
NCM: Oh, she's very regal.
Human: What do you think about england during the reign of elizabeth?
NCM: It was a great place.

这种生成模型可以应对各种不在预设的问题库中的问题,表现形式更加灵活,但是需要巨大规模的训练语料,并且,每回合对话中没有考虑到上下文关系,只根据 Message 生成 Response.

3.3.3 Sequence-to-Sequence + Context 模型

一般聊天中并不是简单的一问一答,回答时往往要参考上下文聊天信息 Context 的内容,所以应该把 Context 引入到 Encoder 中,因为这是除了当前输入 Message 外的额外信息,有助于 Decoder 生成更好的会话应答 Response 内容. 很直接的想法是把 Context 和 Message 拼接起来形成一个长的输入提供给 Encoder,这样就把上下文信息融入模型中了,但是直接拼接起来形成的输入非常长,而对于 RNN 模型来说,输入的线型序列长度越长,模型效果越差.

文献[36]提出在 Encoder 部分采用多层向前神经网络代替 RNN 模型,这样既能将 Context 和 Message 通过多层前向神经网络编码成 Encoder-Decoder 模型的中间语义表达,又避免了 RNN 对于过长输入敏感的问题,图 8、图 9 是文中提到的 2 种融合方法.

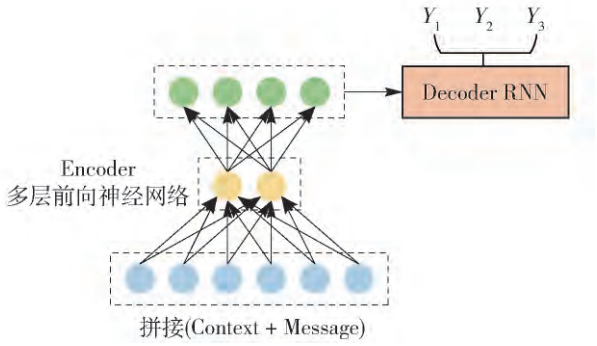

图 8 融合方法 1

Fig. 8 Fusion model 1

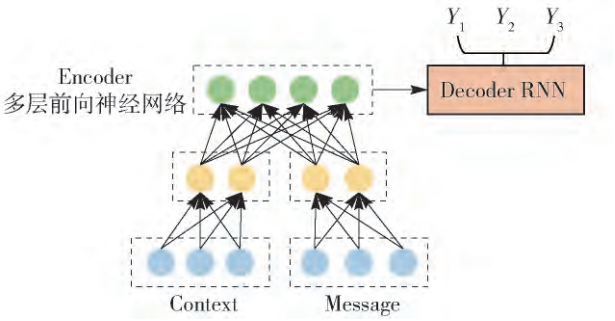

图 9 融合方法 2

Fig. 9 Fusion model 2

方法 1 对 Context 和 Message 不做明显区分,直接拼接成一个输入;而方法 2 则明确区分了 Context 和 Message,在前向神经网络的第 1 层分别对其进行编码,拼接结果作为深层网络后续隐层的输入,核心思想是强调 Message 的作用.

文献[37]提出采用层级神经网络(hierarchical neural network,HNN),如图 10 所示,将 Context 中每个句子首先用"句子 RNN(Sentence RNN)"对每个单词进行编码形成每个句子的中间表示(表示一个句子或一种思想,称作 Thought Vector[46]),而

第二级的 RNN 则将第一级句子 RNN 的中间表示结果按照上下文中句子出现先后顺序序列进行编码，这级 RNN 模型被称作“上下文 RNN(Context RNN)”. 这样尾节点处隐层节点状态信息就是所有 Context 以及当前输入 Message 的语义编码，以这个信息作为 Decoder 产生每个单词的输入之一，这样就可以在生成 Response 的单词时把上下文信息考虑进来.

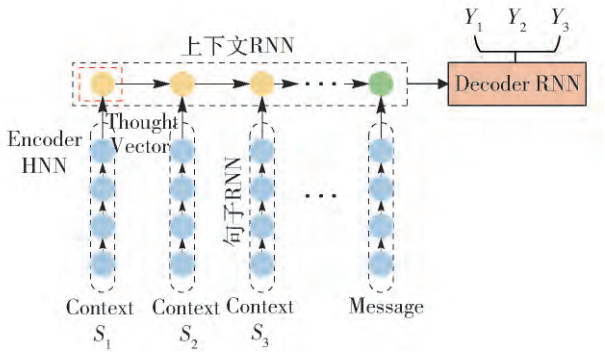

图 10　HNN 结构

Fig. 10　HNN architecture

虽然引入了 Context，但这些方法往往会将距离很远的或本身没有意义的 Thought Vector 也加入 Encoder 部分，生成 Response 的效果并不是很好.

3.3.4　动态记忆网络模型

在 NIPS 2015 Deep Learning Symposium 中文献[8]基于 HNN 的思路提出了动态记忆网络(dynamic memory network，DMN)，结构如图 11 所示.

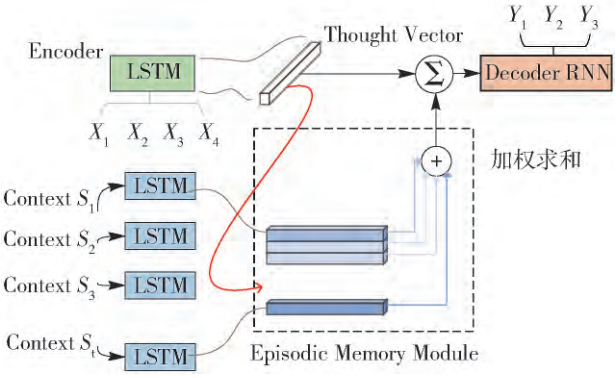

图 11　DMN 网络结构

Fig. 11　DMN network architecture

DMN 的核心思想是在输入模块将 Context 和背景知识编码为 Thought Vector 后存入情景记忆模块(Episodic Memory Module)中，新的 Message 进入时，在 Episodic Memory Module 中检索和推理得到涉及的 Thought Vectors 并激活，放入 Decoder 中生成更好的 Response.

这种方法让系统不仅考虑 Context，还能在一定的背景知识下产生回答，解决好如何检索和推理出相关 Thought Vectors 的问题对于是否能得到高质量的 Response 至关重要.

3.3.5　最大化互信息模型

采用经典的 Encoder-Decoder 模型构建的生成式对话系统比较容易产生“安全回答”问题，用户不论说什么内容，系统都回答少数非常安全、符合语法(grammatical)但没有实际意义的 Response，比如“I don't know!”之类. 原因在于传统的 Sequence-to-Sequence 在 Decoding 过程中都是以极大似然估计法(maximum likelihood estimate，MLE)为目标函数，公式为

$$\hat{\boldsymbol{R}} = \arg\max_{\boldsymbol{R}} \{\log_p(\boldsymbol{R}|\boldsymbol{M})\}$$

式中：$\boldsymbol{M}$ 为 Message；$\boldsymbol{R}$ 为 Response，即生成最符合语法的话，而不是最有用的话，这些安全的句子大量地出现在训练语料中，模型学习了之后，无可避免地总是生成这样的 Response.

文献[9]提出了改进的优化目标函数：最大化互信息(maximum mutual information，MMI)，其目标函数为

$$\hat{\boldsymbol{R}} = \arg\max_{\boldsymbol{R}} \{(1-\lambda)\log_p(\boldsymbol{R}|\boldsymbol{M}) + \lambda\log_p(\boldsymbol{M}|\boldsymbol{R})\}$$

可以从公式差异中看出，MMI 的优化目标除了最大化从 Message 生成应答 Response 的概率，同时加入了反向优化目标，即最大化应答 Response 产生 Message 的概率，其中 λ 是控制两者哪个更重要的调节超参数. 通过具体公式可以看出，这个优化目标函数要求应答 Response 和 Message 内容密切相关而不仅考虑哪个 Response 以更高概率出现，所以降低了那些非常常见的回答的生成概率，使得应答 Response 更多样化且跟 Message 语义更加相关.

采用 MMI 作为目标函数明显解决了很多“安全回答”问题，表 3 是文献中给出的 2 个不同优化目标函数产生的应答 Response 的示例，其中 Message 列代表用户输入语句 Message，S2S Response 代表 MLE 优化目标产生的应答，MMI Response 代表 MMI 优化目标产生的应答.

3.3.6　Speaker Embedding 模型

对话系统作为一个虚拟人物，相关的个性化信息比如年龄、性别、爱好、语言风格等应该维护回答的一致性. 而利用经典的 Encoder-Decoder 模型训练出的对话系统往往很难保持这种一致性，示例如表 4 所示. 这是因为 Sequence-to-Sequence 模型训

表 3 不同优化目标函数产生的 Response

Table 3 Responses generated with the different objective functions

Message	S2S Response	MMI Response
I did not get the report from an mi-6 agent.	I don't think that's a good idea.	You did the right thing, did you?
You haven't been given an assignment in this case.	I don't know what you are talking about.	I've been looking all over for you.
I'm losing my grip.	I don't know what you are talking about.	I'm the only one in the world.
I am ready to help.	Come on, come on.	Ihave something we need to talk about.

练的都是单句 Message 对单句 Response 的映射关系,内在并没有统一维护聊天助手个性信息的场所,不能保证每次相同的问题能够产生完全相同的应答.

文献[40]中提出在模型中考虑个性化信息(比如背景信息、用户画像、年龄等),构建出一个个性化的 Sequence-to-Sequence 模型,称为 Speaker Model,为不同的用户,以及同一个用户对不同的对象自然语言生成不同风格的 Response,架构如图 12 所示.

其整体技术框架仍然采用 Sequence-to-Sequence 架构,基本思路是把聊天助手的个性信息导入到 Decoder 的输出过程中,因为无法对用户的信息显式地进行建模,所以用了一种Embedding的

表 4 利用 Twitter 2 500 万训练数据经过 Sequence-to-Sequence 模型训练后产生的结果

Table 4 Inconsistent responses generated by a 4-layer SEQ2SEQ model trained on 25 million Twitter conversation snippets

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

方法,通过训练来得到 Speaker 向量. 图 12 中右下方是 Speaker 向量在二维平面上的表示,与词向量一样,具有相似背景信息的用户会很接近. Decoding 部分通过计算 LSTM 中的门(gate)向量 $[i_t, f_t, q_t, l_t]$,将 Speaker 向量的信息加载,计算公式为

$$\begin{bmatrix} i_t \\ f_t \\ q_t \\ l_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \boldsymbol{W} \cdot \begin{bmatrix} h_{t-1} \\ e_t^s \\ v_i \end{bmatrix}$$

其中:v_i 表示 Speaker Embedding 的个性信息;σ、tanh

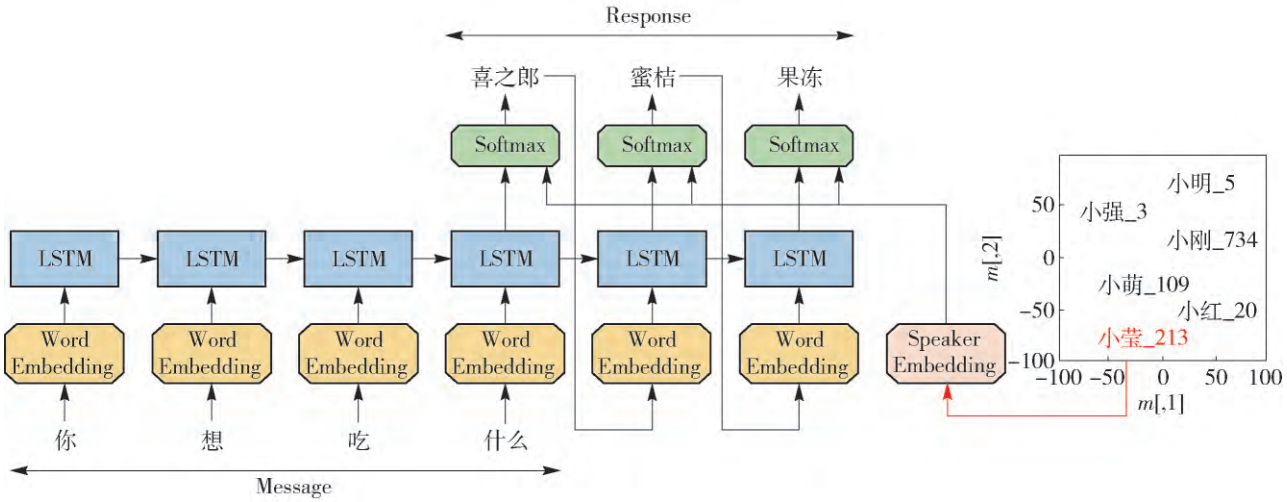

图 12 Speaker Model 架构

Fig. 12 Speaker Model architecture

分别为激活函数；W 为权重矩阵；h_{t-1} 为上一时刻的隐藏单元的输出；e_t^s 为个体文本单元的输入. 在 LSTM 的 Decoder 生成 Response 时，在 t 时刻神经网络节点除了 LSTM 标准的输入外，也将选定身份的个性化 Speaker Embedding 信息一并作为输入，这样就可以引导系统在输出时倾向于输出符合身份特征的个性化信息. 表 5 展示了同样依靠 Twitter 数据训练得到的模型，在解决多轮对话一致性问题上具有突出表现，其中 User1 采用了 Speaker Model，User2 是普通的 Sequence-to-Sequence 模型.

表 5　Speaker Model 产生的结果

Table 5　Results generated by Speaker Model

User 1
message Where is your hometown? response I'm from England. message Where are you from? response I'm from England. message In which city do you live now? response I live in London. message In which country do you live now? response I live in England.
message What is your major? response Business. You? message What did you study in college? response I did business studies.
message How old are you? response I'm 18. message What is your age? response I'm 18.
User 2
message Where is your hometown? response I live in Indonesia. message Where are you from? response England, you? message In which city do you live now? response I live in Jakarta. message In which country do you live now? response I live in Indonesia.
message What is your major? response Business, you? message What did you study in college? response Psychology, you?
message How old are you? response I'm 18. message What is your age? response I'm 16.

Google 的最新智能聊天应用 Allo 也采用类似的方法使用户可以进行个人定制，在神经网络中加入用户的个人风格，将这个神经网络用于生成回复，Google 称之为用户嵌入(user embedding).

可以看出，深度学习中解决此类问题的技术思路都是类似的，核心思想是把对话系统的个性信息在 Decoder 阶段能够体现出来，以此达到维护个性一致的目的.

4　问题与展望

4.1　发展趋势

从目前的技术水平和数据积累程度来看，制作一个开放域的对话系统通过图灵测试更像是一个科幻的梦想，而实现一个面向具体任务的对话系统更加具体、实用，可以解决好垂直领域的问题. 有很多研究，针对具体的业务提供了一些解决方案，虽然通用性或者扩展性还不够强，但目前来看这是对话系统发展的趋势.

4.2　深度学习的应用

基于人工模板或基于知识库检索打造的对话系统在垂直领域中往往有更好的表现，根据目标领域人工编写形式逻辑模型会把对话限制在较窄的范围，易于机器理解，且 NLG 部分也不会产生语法错误的回答，但这种形式结构无法高效地跟上源知识的增长和领域的切换. 相反，基于深度学习打造的对话系统构建过程是端到端数据驱动的，只要给定训练数据即可训练出效果不错的系统，省去了很多特征抽取以及各种复杂的中间步骤的处理，使得系统开发效率大幅提高. 而且对于开发不同语言的对话系统来说，采用 Encoder-Decoder 技术框架，只需要使用不同语言的聊天数据进行训练即可，不需要专门针对某种语言做语言相关的特定优化措施，这使得系统可扩展性大大加强. 但是制作面向具体任务的对话系统时，在具体的任务中拿不到海量的数据，标准化的特大规模的人与人对话数据相对缺乏，很多研究都是通过 Twitter 或者微博评论等高成本的采集方式来收集对话训练数据，或者使用电影字幕等比较间接的方式来积累训练数据.

从上述问题来看，深度学习的可塑性非常强，还有更多的研究致力于让其为对话系统增加更多的能力，比如如何让系统主动引导话题的能力或通过自我学习举一反三的能力等. 所以如何将端到端应用在局部，而非整体上，配合信息抽取和知识图谱等技术，实现一个高度可用的框架体系，这个应该是面向具体任务的对话系统的发展方向.

4.3 对话系统的评价

对话系统 NLG 效果质量的评价标准对于持续提升系统是至关重要的,因为只有这样才能目标明确地有针对性地设计技术方案并进行改进. 对话系统在评价标准方面还有待深入研究,目前深度学习 NLG 中常用的标准包括机器翻译的评价指标 BLEU[47]、语言模型评价标准困惑度等[48-50],这些标准只能评价生成的句子与标准答案间的相似度,而对话过程中一句话的标准答案可能根据上下文的不同而有很大差别,无法用来判断是否真正符合对话语境,所以很多工作是通过人工来进行效果评价的. 没有特别合适地专用于聊天机器人的评价标准,这是阻碍聊天机器人技术持续发展的一个障碍.

5 结论与展望

1) 本文主要讨论了智能对话系统中的自然语言生成领域的研究进展,重点介绍了基于深度学习的 Encoder-Decoder 框架技术原理,然后阐述了深度学习应用在开发实践时需要特殊考虑的几个主要问题,以及在应对这些问题时模型在 Encoder、目标函数和 Decoder 部分的改进方案.

2) 目前对话系统能做的事情十分有限,整体用户体验甚至还没有达到一个合格的应用程序的标准,也并未体现出以自然语言作为交互界面的优势. 由于相关技术还处于发展初期,还有很大的进步空间,所以,智能对话系统的效果也将随之提高,这就需要该领域的研究从解决具体的问题入手,在行程规划、个人助理、售前咨询、客户服务等领域深入研究用户需求,搭建技术基础设施,开发相关产品,不断探索和尝试. 最终,自然语言会作为门槛更低的人机交互界面,使很多烦琐、重复的文字类人工劳动可以被自动化的机器取代.

参考文献:

[1] WEIZENBAUM J. ELIZA-a computer program for the study of natural language communication between man and machine [J]. Communications of the ACM, 1983, 26(1): 23-28.

[2] WALLACE R S. Artificial intelligence markup language [EB/OL]. [2016-08-27]. http://www.alicebot.org/documentation.

[3] WALLACE R S. The anatomy of A. L. I. C. E [M] // Parsing the Turing Test. Berlin: Springer, 2009: 181-210.

[4] TURINGA M. Computing machinery and intelligence [J]. Mind, 1950, 59(236): 433-460.

[5] MCTEAR M F. Spoken dialogue technology - toward the conversational user interface [J]. ACM Computing Surveys, 2004, 34(1): 90-169.

[6] MINKER W, NAKAMURA S, MARIANI J, et al. Spoken dialogue systems technology and design [M]. New York: Springer, 2011.

[7] YANG Y, LIU X. A re-examination of text categorization methods [C] // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, August 15-19, 1999. New York: ACM, 1999: 42-49.

[8] LEWIS D D. Naive (Bayes) at forty: the independence assumption in information retrieval [C] // European Conference on Machine Learning. Berlin: Springer-Verlag, 1998: 4-15.

[9] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features [M] // Machine Learning: ECML-98. Berlin: Springer-Verlag, 1998: 137-142.

[10] KIM Y. Convolutional neural networks for sentence classification [J/OL]. arXiv: 1408.5882, 2014 [2016-08-27]. https://arxiv.org/abs/1408.5882.

[11] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences [J/OL]. arXiv: 1404.2188, 2014 [2016-08-27]. https://arxiv.org/abs/1404.2188.

[12] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification [J]. Computer Science, 2015, 1(4): 39-44.

[13] WEN Y, ZHANG W, LUO R, et al. Learning text representation using recurrent convolutional neural network with highway layers [J/OL]. arXiv: 1606.06905, 2016 [2016-08-27]. https://arxiv.org/abs/1606.06905.

[14] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[15] SUN G L, GUAN Y, WANG X L, et al. A maximum entropy Markov model for chunking [C] // International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18 Aug - 21 Aug. Piscataway: IEEE, 2005: 3761-3765.

[16] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C] // International Conference on Machine Learning. Burlington: Morgan

Kaufmann Publishers Inc，2001：282-289.

[17] PENG F，MCCALLUM A. Information extraction from research papers using conditional random fields [J]. Information Processing & Management，2006，42(4)：963-979.

[18] BENGIO Y，SIMARD P，FRASCONI P. Learning long-term dependencies with gradient descent is difficult. [J]. IEEE Transactions on Neural Networks，1994，5(2)：157-166.

[19] DYER C，BALLESTEROS M，WANG L，et al. Transition-based dependency parsing with stack long short-term memory [J/OL]. arXiv：1505. 08075，2015 [2016-08-27]. https：//arxiv. org/abs/1505. 08075.

[20] MA X，HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [J/OL]. arXiv：1603. 01354，2016 [2016-08-27]. https：// arxiv. org/abs/1505. 08075.

[21] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF Models for sequence tagging [J/OL]. arXiv：1508. 01991，2015 [2016-08-27]. https：// arxiv. org/abs/1508. 01991.

[22] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural architectures for named entity recognition [C] // Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，the Sheraton San Diego Hotel & Marina，June 12 to June 17，2016. Stroudsburg：ACL，2016：260-270.

[23] **王菁华，钟义信，王枞，等. 口语对话管理综述 [J]. 计算机应用研究，**2005，22(10)：5-8.
WANG J H，ZHONG Y X，WANG C，et al. Overview of dialogue management in spoken dialogue system [J]. Application Research of Computers，2005，22(10)：5-8. (in Chinese)

[24] **拜战胜，蓝岚，彭佳红，等. 对话系统中控制模型的比较研究 [J]. 郑州大学学报(理学版)，**2006，38(4)：112-116.
BAI Z S，LAN L，PENG J H，et al. Comparison of dialogue control model in question answering system [J]. Journal of Zhengzhou University (Natural Science Edition)，2006，38(4)：112-116. (in Chinese)

[25] FERRUCCI D A，BROWN E W，CHU-CARROLL J，et al. Building watson：an overview of the DeepQA Project [J]. Ai Magazine，2010，31(3)：59-79.

[26] FERRUCCI D，LEVAS A，BAGCHI S，et al. Watson：beyond jeopardy! [J]. Artificial Intelligence，2013，199/200(3)：93-105.

[27] MURDOCK J W，TESAURO G. Statistical approaches to question answering in watson [J]. Abgerufen AM，2012，11：2013.

[28] GLIOZZO A，BIRAN O，PATWARDHAN S，et al. Semantic technologies in IBM watson [C] // The 51st Annual Meeting of the Association for Computational Linguistics (ACL2013)，Sofia，August 4-9，2013. Stroudsburg：ACL，2013：85-92.

[29] KALYANPUR A，PATWARDHAN S，BOGURAEV B K，et al. Fact-based question decomposition in DeepQA [J]. IBM Journal of Research & Development，2012，56(3)：13-1-13-11.

[30] KALYANPUR A，BOGURAEV B K，PATWARDHAN S，et al. Structured data and inference in DeepQA [J]. IBM Journal of Research & Development，2012，56(3/4)：10-1-10-14.

[31] GONDEK D C，LALLY A，KALYANPUR A，et al. A framework for merging and ranking of answers in DeepQA [J]. IBM Journal of Research & Development，2012，56(3)：399-410.

[32] SUTSKEVER I O，LE Q V. Sequence to sequence learning with neural networks [J]. Advances in Neural Information Processing Systems，2014，4：3104-3112.

[33] GRAVES A. Long short-term memory [M] // Supervised Sequence Labelling with Recurrent Neural Networks. Berlin：Springer-Verlag，2012：1735-1780.

[34] VINYALS O，LE Q. A neural conversational model [J/OL]. arXiv：1506. 05869，2015 [2016-08-27]. https：//arxiv. org/abs/1506. 05869.

[35] SHANG L，LU Z，LI H. Neural responding machine for short-text conversation [J/OL]. arXiv：1503. 02364，2015 [2016-08-27]. https：// arxiv. org/abs/1503. 02364.

[36] PASCUAL B，GURRUCHAGA M，GINEBRA M P，et al. A neural network approach to context-sensitive generation of conversational responses [J]. Transactions of the Royal Society of Tropical Medicine & Hygiene，2015，51(6)：502-504.

[37] SERBAN I V，SORDONI A，BENGIO Y，et al. Building end-to-end dialogue systems using generative hierarchical neural network models [J/OL]. arXiv：1507. 04808，2015 [2016-08-27]. https：// arxiv. org/abs/1507. 04808.

[38] KUMAR A，IRSOY O，ONDRUSKA P，et al. Ask me anything：dynamic memory networks for natural language processing [J/OL]. arXiv：1506. 07285，2015 [2016-08-27]. https：//arxiv. org/abs/1506. 07285.

[39] LI J，GALLEY M，BROCKETT C，et al. A diversity-promoting objective function for neural conversation

models [J/OL]. arXiv: 1510. 03055, 2015 [2016-08-27]. https://arxiv. org/abs/1510. 03055.

[40] LI J, GALLEY M, BROCKETT C, et al. A persona-based neural conversation model [J/OL]. arXiv: 1603. 06155, 2016 [2016-08-27]. https: // arxiv. org/abs/1603. 06155.

[41] XU W, RUDNICKY A. Can artificial neural networks learn language models? [C] // International Conference on Spoken Language Processing, Beijing, China, October, 2000. Amsterdam, the Netherlands: Speech Communication, 2000: 202-205.

[42] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[43] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning [C] // Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, July 05 -09, 2008. New York: ACM, 2008: 160-167.

[44] MNIH A, HINTON G. Three new graphical models for statistical language modelling [C] // Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, June 20 -24, 2007. New York: ACM, 2007: 641-648.

[45] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model [C] // INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September, 2010. Amsterdam, the Netherlands: Speech Communication, 2010: 1045-1048.

[46] KIROS R, ZHU Y, SALAKHUTDINOV R, et al. Skip-thought vectors [C] // Advances in Neural Information Processing Systems, Palais des Congrès de Montréal, December 07-12, 2015. California: NIPS, 2015: 3294-3302.

[47] PAPINENI K, ROUKOS S, WARD T, et al. IBM research report bleu: a method for automatic evaluation of machine translation [J]. ACL Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002, 30(2): 311-318.

[48] LIU C W, LOWE R, SERBAN I V, et al. How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation [J/OL]. arXiv: 1603. 08023, 2016 [2016-08-27]. https: // arxiv. org/abs/1603. 08023.

[49] ROSEN-ZVI M, CHEMUDUGUNTA C, GRIFFITHS T, et al. Learning author-topic models from text corpora [J]. ACM Transactions on Information Systems, 2010, 28(1): 312-324.

[50] HEINRICH G. Parameter estimation for text analysis [R]. Leipzig: University of Leipzig, 2008.

(责任编辑 梁 洁)