文章编号：2096-1472(2021)-02-02-06
DOI:10.19644/j.cnki.issn2096-1472.2021.02.001

聊天机器人的分类标准和评估标准综述

王艳秋[1,2]，管浩言[1,2]，张 彤[1,2]

(1.大连东软信息学院，辽宁 大连 116023；
2.大连东软教育科技集团有限公司研究院，辽宁 大连 116023)
✉wangyanqiu@neuedu.com; guanhaoyan@neuedu.com; zhangtong@neuedu.com

摘 要：近年来，人工智能和大数据技术的发展极大地推动了聊天机器人产业的发展。如今，聊天机器人种类众多，但质量参差不齐，对其进行评估成为当下的重要问题之一。本文首先通过功能和技术实现方式方面的分析，对当前的聊天机器人进行了归纳分类。然后从多方面对聊天机器人的评估方式进行了系统的整理与总结，并详细介绍分析了其中各种评估指标。最后探讨了当前聊天机器人的研究难点与评估难点，并在此基础上对聊天机器人未来的研究发展方向进行了展望。

关键词：聊天机器人；分类标准；评估标准
中图分类号：TP391 **文献标识码**：A

Research on Taxonomy and Evaluation Criteria of Chatbots

WANG Yanqiu[1,2], GUAN Haoyan[1,2], ZHANG Tong[1,2]

(1. *Dalian Neusoft University of Information*, *Dalian* 116023, *China*;
2. *Research Institute*, *Dalian Neusoft Education Technology Group Co. Limited*, *Dalian* 116023, *China*)
✉wangyanqiu@neuedu.com; guanhaoyan@neuedu.com; zhangtong@neuedu.com

Abstract: In recent years, the development of artificial intelligence and big data technology has greatly promoted chatbot industry. Currently, there are many types of chatbots, but the quality is uneven, and the evaluation criteria are becoming an important issue. This paper first summarizes and categorizes the current chatbots technology based on their functions and technical implementation methods. Then, it systematically proposes an evaluation approach for chatbots quality via different aspects, and introduces various evaluation indicators in detail. Finally, current research issues and evaluation difficulties of chatbots are discussed, and on this basis, future research and development directions of chatbots are prospected.

Keywords: chatbot; classification criteria; evaluation criteria

1 引言(Introduction)

随着人工智能和大数据技术的发展，聊天机器人已经不再是个新鲜的词汇，并且早已慢慢渗入人们的日常生活中，如苹果的Siri、阿里巴巴的阿里小蜜、百度的小度、微软的Cortana和小冰、亚马孙的Alexa、IBM的Watson等。这些聊天机器人应用于不同场景，有着不同的定位与功能，但其中都使用了自然语言处理(Natural Language Processing，NLP)相关技术，使机器人能够使用文本或语音与人类进行对话。如今的聊天机器人并不完善，时常会出现答非所问、语句不通顺等问题，因此聊天机器人需要能够反映其真实智能水平的评估标准来促进其优化改进。本文针对不同功能与技术实现方式对聊天机器人进行了分类，同时对现有的所有聊天机器人评价指标进行了分析、分类与总结，并指出了当下聊天机器人发展的困境以及未来的发展方向。

2　聊天机器人分类(Chatbot taxonomy)

2.1　任务导向型与闲聊型

根据功能的不同，可分为任务导向型聊天机器人和闲聊型聊天机器人。任务导向型聊天机器人是指以任务驱动来完成多轮对话的对话系统，通常针对封闭专业领域知识，机器人需要在对话过程中理解、澄清并生成对话，其构建方式主要为Pipeline和End-to-end。Pipeline的构建采用模块化结构，包含四个主要模块：自然语言理解(Natural Language Understanding，NLU)、对话状态追踪(Dialogue State Tracking，DST)、对话策略学习(Dialogue Policy Learning，DPL)、自然语言生成(Natural Language Generation，NLG)。这种构建方式容易实现，可解释性强，但模块之间误差会逐层积累，又因各模块之间相互独立导致无法联合调优。End-to-end即基于深度学习的端到端系统，使用大量标注数据进行训练得到一个深度学习模型，用户从输入端输入语句便可从输出端得到相应回复。这种方法可扩展性强，但需要大量且高质量的标注数据，目前仍处于探索阶段。闲聊型聊天机器人主要与用户进行面向开放域的闲聊，目标是与用户进行有意义的自然多轮对话。相比于任务导向型聊天机器人，闲聊型聊天机器人涉及的领域范围更大，用户意图更模糊难识别，因此要求更高，实现更难。

2.2　检索式与生成式

聊天机器人需要对用户的输入做出自然的语言回复，这涉及自然语言生成技术。根据对话生成的不同技术，可将聊天机器人分为检索式与生成式两种[1]。检索式模型基于现成的数据库进行基于规则的模式匹配，或应用较为复杂的深度学习算法进行模式匹配，但并不生成全新的回复。因此检索式模型产生的回复具有流畅自然、信息量丰富的优点，但同时也具有无法进行上下文关联的不足[2]。最早出现的模拟心理医生的聊天机器人ELIZA便是完全基于规则手工建立的，虽然能够生成较好的回复，但构建过程过于烦琐，耗费了大量人力。生成式模型则会产生全新的语句回复，通过将大量人类真实语料输入深度学习模型中进行特征提取与特征学习，再使用模型对用户的输入做出回复。生成式模型会有记忆功能，即可利用历史对话信息形成对话的上下文关联，但生成的回复可能会有不符合语法规则、语句不通顺、逃避复杂问题进行无意义回复等情况出现。目前生成效果较好的模型有微软DialoGPT、谷歌Meena、Facebook Blender、百度PLATO & PLATO-2等，这些均使用了超大规模文本数据进行模型训练，模型参数都在亿量级。

3　评价指标分类(Evaluation index taxonomy)

聊天机器人评价通常是指对机器人对话回复质量的评估，但也有聊天机器人能够识别用户发出的图片并进行回复与评论，这时则涉及图像描述生成的相关评估。此外，对于一些产品化的任务导向型聊天机器人，也需要进行一些产品层面的评估。本文分别对这几类聊天机器人的评价指标进行了总结，图1是所有评价指标的分类图。下文将对所有评价指标进行详细介绍。

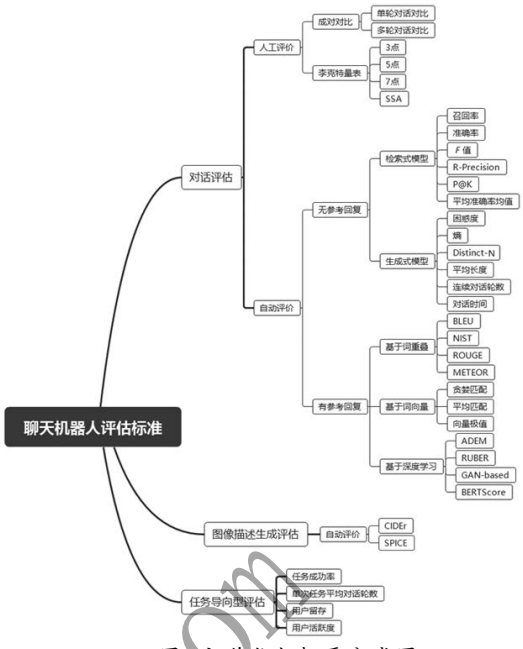

图1 评价指标概要分类图

Fig.1 Evaluation index taxonomy

3.1　对话评估

3.1.1　人工评价

人工评价是目前最准确、最有效地对话质量评价指标，但存在耗费人力、耗时长的问题，主要包含成对对比和李克特量表评价两种评价方式。

成对对比：即对两个系统产生的回复就不同的侧重点进行人工评价，如图2所示的ACUTE-EVAL评估界面，它要求人们比较两个多轮对话，在对话1(浅蓝色)和对话2(深蓝色)之间进行选择。同样还有基于单轮对话的成对对比评估。

图2 ACUTE-EVAL成对对比人工评价界面[3]

Fig.2 ACUTE-EVAL multi-turn dialogues pairwise human evaluation interface

李克特量表：在聊天机器人的人工评价中，李克特量表指的是李克特量表形式的人工评分，评分可设置为3、5、7等级，如对聊天机器人的回复是否连贯进行5等级评分，将分数范围设置为[0,1,2,3,4]，再由人工针对回复的连贯性在

分数范围内选择合适的分数进行评价。可以针对对话质量的多方面进行评价，如对话的信息量(Informativeness)、连贯性(Coherence)、新颖性(Engagingness)、人性(Humanness)等。还有一种谷歌在其Meena聊天机器人中提出的评价指标SSA(Sensibleness and Specificity Average)，指的是敏感性和特异性平均值。特异性表示是否是针对上句对话的特定的具体的回答，敏感性表示聊天机器人的对话是否有意义。单纯以敏感性作为唯一指标，会导致回答趋向模糊无聊的安全回答，因此将敏感性与特异性结合来进行综合评价更能体现回复的质量。实验显示，SSA与自动评价指标困惑度成正相关关系。

尽管人们一直在探索能够代替人工评价的自动评价方法，但至今没有自动评价方法能够代替人工评价，人工评价仍是所有聊天机器人都必须进行的评价。人工评价尽管必不可少，但也有一些弊端，例如不同模型的评价者背景条件、人群规模往往不尽相同，在不同模型之间很难做到完全客观的对比评价。

3.1.2 自动评价

自动评价可以分为两部分：一部分不需要参考回复即可进行评价，其中包含针对检索式模型和生成式模型的评价指标；另一部分是需要参考回复的评价指标，且基本都是针对生成式模型所生成对话的质量的评价。而根据评价原理又可分为基于词重叠、基于词向量以及基于深度学习的各种评价指标。

(1)不需要参考回复——检索式模型

检索式聊天机器人的关键点在于匹配答案时候选回复的排列顺序，所以其评价指标一般使用传统信息检索系统常用的评价指标。

召回率(Recall)：又称查全率，指检索出的相关回复占所有相关回复总数的比例，表示是否查全。

准确率(Precision)：又称查准率，指检索出的相关回复占所有检索出的回复总数的比例，表示是否查准。

*F*值(*F*−measure)：指召回率和准确率的调和平均值，它综合了两者的评价效果。

$$F=\frac{\left(\alpha^2+1\right)P*R}{\alpha^2\left(P+R\right)} \tag{1}$$

通常取$\alpha=1$，即为F_1值，其取值范围为[0,1]，值越接近1，说明对算法的评价越好。

$$F_1=2\frac{precision\cdot recall}{precision+recall} \tag{2}$$

R−Precision：指的是每个查询检索出*R*个相关回复时的准确率。

P@K：指的是单个查询中检索出的前*K*个回复的准确率。

平均准确率均值(Mean Average Precision，MAP)：平均准确率(Average Precision，AP)将准确率与候选回复的排列顺序相结合，如公式(3)所示，其中*i*指第*i*个候选回复；*s*表示第*i*个回复的位置，指的是一个查询中检索出的相关回复的P@K的平均值。MAP则是对所有查询的平均准确率再求均值，其值越高说明检索出的相关回复排列顺序越靠前。

$$AP=\frac{1}{r}\sum_{i=1}^{r}\frac{i}{s} \tag{3}$$

$$MAP=\frac{1}{Q}\sum_{q=1}^{Q}AP(q) \tag{4}$$

(2)不需要参考回复——生成式模型

生成式模型主要评价的是生成回复的质量，其评价一方面聚焦于回复本身的信息量和生成概率，另一方面则由用户与其交互的时间来侧面反映。

熵(Entropy)：指的是回复中N−gram包含的信息量，用来衡量回复多样性[4]。

困惑度(Perplexity)：语言模型的标准度量指标[5]，可用来评价对话模型中回复的生成质量，一定程度上可体现多样性，是目前常用的聊天机器人回复质量评价指标。语言模型实际上是计算语句概率的模型，概率值越高，语言模型越好，困惑度越小。

$$PP(W)=P\left(w_1w_2\cdots w_N\right)^{-\frac{1}{N}}=\sqrt[N]{\frac{1}{P\left(w_1w_2\cdots w_N\right)}} \tag{5}$$

Distinct−N：回复多样性指标[6]，其值越大说明生成回复越具有多样性。分母表示回复中的N−gram总数，分子表示回复中不重复的N−gram数目。*N*通常取1和2，即取Uni−grams和Bi−grams。

$$Distinct(N)=\frac{Count\left(unique\ gram_N\right)}{Count(word)} \tag{6}$$

平均长度(Average Length)：指的是生成回复的平均长度，普遍认为生成长句子的对话生成模型相对质量更高。

单次平均对话轮数(Conversation−turns Per Session，CPS)：指的是聊天机器人和用户之间的每次对话中所含对话轮数的平均值[7]。一般用来对闲聊型聊天机器人进行评价，CPS越大，说明聊天机器人的社交参与程度越高。

对话时间：指用户与聊天机器人的对话所持续的时间。

(3)需要参考回复——基于词重叠

基于词语重叠的评价方法需要有参考回复，主要是根据参考回复与生成回复之间词语的重叠程度来进行度量。

BLEU：全称为BiLingual Evaluation Understudy[8]，最早用于机器翻译任务，评价前提是需要语料库中有高质量的参考回复，核心思想是比较生成回复文本和参考回复文本中N−gram的重合程度，重合程度越高则认为文本质量越高。*N*

一般取1—4，然后进行加权平均，$N=1$时用于衡量单词翻译的准确性，$N>1$时用于衡量句子的流畅性。随后优化改进出了多种新的评价指标。虽然近年来BLEU被证明与人工判断的相关性不高[9]，但目前仍然是聊天机器人评估常用的指标。

NIST：全称是National Institute of Standards and Technology[10]，改进自BLEU方法，引入了每个N-gram的信息量的概念，定义见公式(7)。公式中分母表示N-gram在参考回复中出现的次数，分子表示对应的(N-1)-gram在参考回复中出现的次数，当$N=1$时，分子取值为整个参考回复的长度。由此，将一些出现较少的重点词的权重增大。

$$Info(w_1\cdots w_N)=\log_2\left(\frac{the\ \#\ of\ occurrences\ of\ w_1\cdots w_{N-1}}{the\ \#\ of\ occurrences\ of\ w_1\cdots w_N}\right) \tag{7}$$

ROUGE：全称是Recall-Oriented Understudy for Gisting Evaluation，改进自BLEU方法，不同于BLEU，它专注于衡量N-gram的召回率，而不是准确率。通常使用的有ROUGE-N[11]和ROUGE-L[12]。ROUGE-N通过统计参考回复中N-gram的个数与参考回复和生成回复中共有的N-gram个数来计算召回率。

$$ROUGE-N=\frac{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_N\in S}Count_{match}(gram_N)}{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_N\in S}Count(gram_N)} \tag{8}$$

ROUGE-L中L表示最长公共子序列(Longest Common Subsequence，LCS)，使用参考回复和生成回复的最长公共子序列进行计算。

$$R_{lcs}=\frac{LCS(X,Y)}{m} \tag{9}$$

$$P_{lcs}=\frac{LCS(X,Y)}{n} \tag{10}$$

$$F_{lcs}=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}} \tag{11}$$

其中，公式(10)中m表示参考回复的长度；公式(11)中n表示生成回复的长度。

METEOR：全称为Metric for Evaluation of Translation with Explicit ORdering[13]，该指标同时考虑了准确率和召回率，其中召回率的权重更高。将生成回复与参考回复之间的Uni-grams通过简单的映射进行对齐，可进行词干提取和精确的单词匹配，从而计算得到特定的匹配关系，与人类判断有较好的相关性。

(4)需要参考回复——基于词向量

不同于基于词重叠(即利用N-gram计算生成回复和参考回复之间的重合程度)的方式，基于词向量的评价方式则是利用Word2Vec、Sent2Vec等方法把回复表示为句向量，再通过余弦相似性等方法计算生成回复与参考回复之间的相似程度。

贪婪匹配(Greedy Matching)：本质是计算两个语句的相似性。该方法分别将生成回复和参考回复中的每个词转换为词向量，然后对参考回复中每个词向量，计算其在生成回复中与每个词向量的余弦相似度，取最高的余弦相似度将其相加并求平均，最后再对生成回复进行相同流程的计算，取两者平均值。

平均匹配(Embedding Average)：使用句向量计算生成回复和参考回复的余弦相似度。句向量由语句中每个词向量相加再取平均值得到。

向量极值(Vector Extrema)：同样基于句向量计算两个语句的相似性，但句向量由词向量每个维度中极值最大的一维构成，然后再计算余弦相似度。这种方法可以忽略语句中的常见表达，保留特殊的重要语义词语[14]。

(5)需要参考回复——基于深度学习

近几年，深度学习快速发展，针对基于深度学习进行生成回复评价的研究也逐渐增多。以下是几种典型的应用深度学习进行生成回复评价的方法。

ADEM：全称为Automatic Dialogue Evaluation Model[15]，即对话系统自动评价模型，它将对话系统的评价问题转换为预测回复语句的人工评分问题，收集人类对对话语料进行评分的数据集，训练使用循环神经网络(RNN)构建自动评价模型。虽然文章指出这种方法效果要好于BLEU、ROUGE，但后续研究表明ADEM存在明显的缺陷，其分配给各种回复的分值分布在2.75 ± 0.34范围内，分辨力较低，无法为多个回复提供合适的评分，仍需要改进[16]。

RUBER：全称为Referenced metric and Unreferenced metric Blended Evaluation Routine[17]，是一种针对开放域对话系统的无监督自动评估方法，不需要人工评分数据。其主要思想是将有参考回复评估和无参考回复评估以不同的策略结合起来以提高评估性能。有参考回复评估采用词向量池化的方法，选择词向量每个维度的最大值和最小值来代表语句，然后计算余弦相似度；无参考回复评估通过训练神经网络模型来衡量生成回复和对应查询之间的匹配程度。实验表明，RUBER可扩展到不同数据集中，且与人工评价具有一定的相关性。

GAN-based：生成式对抗网络(Generative Adversarial Network，GAN)通常应用于图像生成任务中，受其启发产生了基于GAN结构的对话系统评价模型，使用生成器生成回复，判别器区分生成回复和参考回复。

BERTScore：一种基于Bert的生成回复评估方法[18]。给定一个参考回复和生成回复，使用Bert来提取输入每个单词的上下文特征，表示为带有上下文信息的词向量，然后使用余弦相似度计算每两个词向量之间的匹配相似度。使用贪婪匹

配来最大化匹配相似度得分，选择性地使用逆文档频率分数对词向量进行重要性加权。实验表明，BERTScore取得了比一般指标更好的相关性，并且对于模型选择有一定效果，但是没有一种BERTScore配置明显优于其他所有配置。

3.2 图像描述生成评估

人们在社交聊天中经常会围绕图片展开交流和讨论，图片中所体现的事物、事件、氛围或感情通常是人们讨论的主要内容。图像描述生成技术便是为了能够自动生成能真实全面地表现图片中发生事件以及反映出的感情的描述，运用到的技术实际上是计算机视觉(Computer Vision，CV)和自然语言处理的结合，通过CV技术分析图像内容，利用NLP技术生成相对应的文字来描述图像中明显的特征。生成对话的评估方法大多数能直接用于图像描述生成的评估，除此以外，CIDEr和SPICE是专门用于图像描述生成的评估方式。

CIDEr：全称是Consensus-based Image Description Evaluation[19]，即基于共识的图像描述评估。其主要思想是利用TF-IDF计算得到生成回复和参考回复的不同N-gram的权重，将在数据集中比较常见、包含较小信息量的N-gram权重调低，然后计算生成回复与参考回复的余弦相似度，再对每个N-gram的相似度加和求平均值，得到最终的CIDEr评估值。

SPICE：全称是Semantic Propositional Image Caption Evaluation[20]，即语义命题图像描述评估。不同于CIDEr利用词语重叠进行评估，SPICE通过建立场景图(Scene Graphs)来对图像描述中的对象、属性和关系进行编码。首先利用PCFG依赖解析器把要评估的图像描述转换为语法依赖树；然后根据九种简单的语言规则把生成的语法依赖树映射到场景图；再把场景图中的语义关系看作对象、属性和关系构成的元组，计算生成回复和参考回复的元组之间的F值作为最终的SPICE评估值。

3.3 任务导向型评估

任务导向型聊天机器人通常应用于特定的情景和场所中，面向特定领域，主要是一些为用户提供信息或任务导览等服务来满足用户明确需求的机器人。目前这类机器人在订餐、订票、订酒店、商品咨询、业务办理等方面应用较多。虽然任务导向型聊天机器人也可以用准确率、召回率等评价对话质量的标准来评估，但更多地需要从整体来对产品进行评价。

任务成功率：指成功解决用户问题的对话所占比例，如票务系统为用户成功订票次数占全部订票需求数量的比例。

单次任务平均对话轮数：与前文中的单次平均对话轮数(CPS)不同，任务导向型聊天机器人讲求效率，需要在尽可能少的对话轮数内解决问题，所以对话越简洁、越明确，越能为用户提供更好的服务。

用户留存：好的产品需要不断地迭代更新，与此同时，用户往往是流动的。用户留存率可以让开发者更清晰地看到更新前后一段时间内的用户留存状态，从而对产品的优化提供反馈。

用户活跃度：用户活跃度指的是频繁使用产品的用户所占比例，即会频繁地使用任务导向型聊天机器人进行相关服务的用户所占比例，用户活跃度越高，侧面说明机器人的任务完成得越好，越能满足用户要求。

4 研究难点与未来发展方向(Research difficulties and future development direction)

4.1 研究难点

随着近几年相关技术的发展，尤其是深度学习的逐渐成熟，聊天机器人技术也在快速发展，但仍存在着诸多难点。

(1)对话技术依旧不成熟

目前在某些封闭域方面，聊天机器人可以很好地与用户进行沟通，比如购票系统等。但当聊天范围逐渐扩大到开放领域，即用户希望与聊天机器人闲聊时，聊天机器人的回答就会变得粗糙。这就是目前技术的瓶颈，即如何让聊天机器人在与用户进行无特定范围的开放域聊天时，能做出合理回复。聊天机器人需要数据集来反复训练，一旦用户期望的对话内容没有在训练数据集中体现，聊天机器人就无法给出合理的回答，然后给出“我不知道”等搪塞用户的敷衍回答。

(2)人类和聊天机器人对话的心理问题

恐怖谷理论说明，当机器人的外貌和人类极其相似的时候，人类会对它产生非常强烈的厌恶情绪。在对话方面，人类也有类似心理，即当聊天机器人的回答内容过于真实或表现出过于透彻的了解时，会使用户产生隐私被窥视的感受，用户可能会产生厌恶心理。这种现象是十分矛盾的，算法的设计需要聊天机器人的回答内容趋向于真实自然，并且以对用户信息的了解为基础才能生成个性化对话内容；但是表现得过于真实与了解就可能使用户产生反感，甚至出现侵犯隐私问题。

(3)聊天机器人的个性选取

对于同一个问题，不同的人会有不同的回答，这取决于每个人的个性，聊天机器人也一样。目前主流的聊天机器人个性设置都是温柔、耐心等，但由于暴力、色情等不良内容很容易出现在聊天机器人的训练数据集中，导致聊天机器人的个性并不能完全被控制。另一方面，某些用户在与聊天机器人对话的过程中可能表现出一些心理问题，聊天机器人如何疏导用户，帮助其调整心态，而不是加重其心理问题是目前技术暂时无法突破的难点。

(4)聊天机器人所需计算资源较大

深度学习让聊天机器人的鲁棒性有了很大的飞跃，但

同时也带来了巨大的计算资源的需求。尤其是现在聊天功能的需求广泛，网页端、移动端等没有太多计算资源的边缘设备，都需要后台服务器辅助计算。对此问题，轻量化聊天机器人的算法、对算法的蒸馏等，仍需要更多的研究和应用。

(5)需要“大规模”和“有质量”的语料库

语料库，即聊天机器人的训练数据集，是机器人学习说话的来源，对于回答的质量非常关键。“大规模”指的是语料库内容要多，涉及方方面面，才能让机器人无所不知；“有质量”指语料库的内容要可靠，不能有不良信息，也不能有答非所问的内容，这样的语料库才能训练出优秀的聊天机器人。而现实是，一方面高效获得语料库是一个难点问题；另一方面即使找到现有的语料库，目前最多的训练用语料库都是以成亿计，语料的内容也是良莠不齐，高质量语料筛选工作也是一个难点问题。

(6)自动评估与人工评估相关性较差

生成回复的自动评价一直是聊天机器人评估领域探索的重点内容，也是难点内容。由于自动评价与人工评价的相关性一直不高，尤其是现有的自动评价方法很多都来源于机器翻译等其他领域，对生成回复的语义多样性能否进行评价，以及对模型的有效性和优化反馈能否起到作用等问题一直存在争议。

4.2 发展方向

未来聊天机器人的发展方向将趋向于成熟的对话生成模型训练和模型轻量化。目前聊天机器人的回答依然存在答非所问等问题，未来的发展方向必然需要向增强对话生成的鲁棒性和合理性前进。另一方面，计算轻量化的需求也日益增长，即能够在计算能力较弱的机器人中部署需求，这是当今聊天机器人应用场景与应用设备日益扩张的必然要求。

5 结论(Conclusion)

目前，进入市场并产品化的聊天机器人主要是功能导向型聊天机器人，产品形式主要是嵌入PC端与手机端应用的问询功能模块、实体化的问询功能机器人和智能语音音箱等智能家居。当前相关产业已经较为成熟，产品也逐渐趋同，评价精度方面并无较大进展。处于研究阶段的大规模开放域的训练模型，训练参数逐渐增多，模型体量逐渐增大，发展空间与潜力较大。但这些模型质量参差不齐，对其进行有效精准的评价十分重要。本文在实现功能和实现技术两方面对聊天机器人进行了分类，从多方面对评价标准进行了较为系统的介绍、分析与总结，提出了目前聊天机器人技术的研究难点与未来的发展方向。希望能够为目前聊天机器人的分类和评价标准构建出一个较为完整的全局概览图，为相关研究人员提供一定参考和借鉴。

参考文献(References)

[1] 陈晨,朱晴晴,严睿,等.基于深度学习的开放领域对话系统研究综述[J].计算机学报,2019,042(007):1439-1466.

[2] 戴怡琳,刘功申.智能聊天机器人的技术综述[J].计算机科学与应用,2018,8(6):918-929.

[3] Li M, Weston J, Roller S. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons[DB/OL]. [2019-09-06]. https://arxiv.org/pdf/1909.03087.pdf.

[4] Zhang Y, Galley M, Gao J, et al. Generating informative and diverse conversational responses via adversarial information maximization[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 1815-1825.

[5] Tevet G, Berant J. Evaluating the evaluation of diversity in natural language generation[DB/OL]. [2020-04-26]. https://arxiv.org/pdf/2004.02990v2.pdf.

[6] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016:110-119.

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

[8] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002:311-318.

[9] 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[DB/OL]. [2017-01-03]. https://arxiv.org/pdf/1603.08023v2.pdf.

[10] Doddington G. Automatic evaluation of machine translation quality using N-gram co-occurence statistics[C]. Proceedings of the second international conference on Human Language Technology Research, 2002:138-145.

[11] Lin C Y, Hovy E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003:71-78.

[12] Lin C Y, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram

statistics[C]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004: 605-612.

[13] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005:65-72.

[14] 张伟男,张杨子,刘挺.对话系统评价方法综述[J].中国科学:信息科学,2017,47(08):953-966.

[15] Lowe R, Noseworthy M, Serban I V, et al. Towards an automatic turing test: Learning to evaluate dialogue responses[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017:1116-1126.

[16] Sai A B, Gupta M D, Khapra M M, et al. Re-evaluating ADEM: A deeper look at scoring dialogue responses[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33:6220-6227.

[17] Tao C Y, Mou L, Zhao D Y, et al. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems[C]. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018, 32(1):722-729.

[18] Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating text generation with BERT[DB/OL]. [2020-02-24]. https://arxiv.org/pdf/1904.09675.pdf.

[19] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based Image Description Evaluation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:4566-4575.

[20] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic Propositional Image Caption Evaluation[J]. Adaptive Behavior, 2016, 11(4):382-398.

作者简介：

王艳秋(1993-)，女，硕士，初级研究员.研究领域：人工智能，数据挖掘.

管浩言(1994-)，男，硕士，初级研究员.研究领域：人工智能，计算机视觉.

张　彤(1994-)，女，硕士，初级研究员.研究领域：人工智能，图像处理.

(上接第18页)

[15] Lin Z, Xu P, Winata GI, et al. CAiRE: An End-to-End Empathetic Chatbot[C]. In AAAI, 2020:13622-13623.

[16] Li Y, Su H, Shen X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset[DB/OL]. [2017-10-11]. https://arxiv.org/pdf/1710.03957.pdf.

[17] Chen SY, Hsu CC, Kuo CC, et al. Emotionlines: An emotion corpus of multi-party conversations[DB/OL]. [2018-02-23]. https://arxiv.org/pdf/1802.08379.pdf.

[18] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation[DB/OL]. [2015-03-09]. https://arxiv.org/pdf/1503.02364.pdf.

[19] Zhou X, Wang WY. Mojitalk: Generating emotional responses at scale[DB/OL]. [2017-11-11]. https://arxiv.org/pdf/1711.04090.pdf.

[20] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[DB/OL]. [2014-09-08]. https://arxiv.org/pdf/1409.2329.pdf.

[21] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[22] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[DB/OL]. [2014-06-03]. https://arxiv.org/pdf/1406.1078.pdf.

[23] Hochschild AR. Emotion work, feeling rules, and social structure[J]. American journal of sociology, 1979, 85(3):551-75.

[24] Song Z, Zheng X, Liu L, et al. Generating responses with a specific emotion in dialog[C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019:3685-3695.

[25] Shen L, Feng Y. CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation[DB/OL]. [2020-05-01]. https://arxiv.org/pdf/2005.00329.pdf.

[26] Liu CW, Lowe R, Serban IV, et al. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation[DB/OL]. [2016-03-25]. https://arxiv.org/pdf/1603.08023.pdf.

作者简介：

肖　鹏(1993-)，男，硕士，初级研究员.研究领域：计算机视觉，自然语言处理.

于　丹(1976-)，女，博士，研究员.研究领域：数据分析与挖掘，人工智能.

王建超(1989-)，男，硕士，中级研究员.研究领域：人工智能，图像处理.

来关军(1984-)，男，硕士，中级研究员.研究领域：大数据分析，人工智能.