对话系统评价技术进展及展望

张伟男 车万翔
哈尔滨工业大学

关键词：人机对话系统 开放域对话系统评价 任务型对话系统评价

人机对话系统的研究最早可以追溯到1950年图灵(Alan M. Turing)在*Mind*上发表的文章“Computing Machinery and Intelligence”。文章开篇提出了“机器能思考吗？”(Can machines think?)的设问，并通过让机器参与一个模仿游戏(Imitation Game)来验证“机器”能否“思考”，这就是后来被人们广泛熟知的图灵测试(Turing Test)。值得注意的是，图灵测试的实现及操作方式是以人机对话的形式进行的，如图1所示，测试者借助某种装置以对话的方式与人类和对话系统进行交谈，当测试结束后，如果有30%以上的测试者不能正确区分对话系统和人的回复（即将对话系统的回复误判成人的回复），那么则称该对话系统通过了图灵测试，拥有了人的智能。

尽管近年来人机对话的研究和应用如火如荼，但是，如何评价对话系统，尤其是开放域对话系统（聊天机器人），仍然是一个开放性问题。

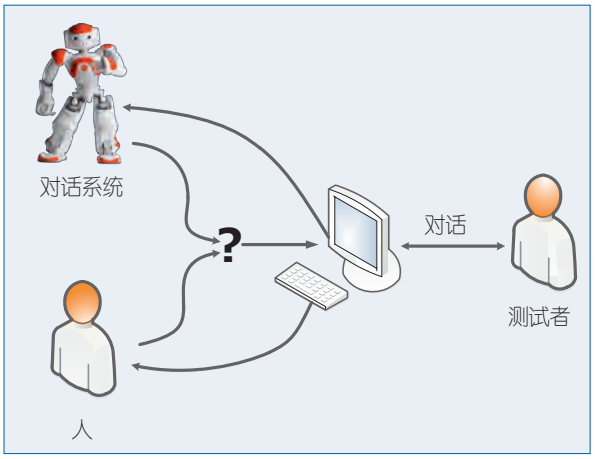

图1 图灵测试示意

我们将人机对话系统的功能主要分为四个方面，如表1所示。

表1 人机对话系统的四类主要功能

	聊天	知识问答	任务执行	推荐
目的	闲聊	知识获取	完成任务或动作	信息推荐
领域	开放域	开放域	特定域（垂类）	特定域
对话轮数评价	越多越好	越少越好	越少越好	越少越好
应用场景	娱乐、情感陪护、营销沟通	客服、教育	虚拟个人助理	个性化推荐
典型系统	微软小冰	IBM Watson、Wolfram Alpha	Siri、Cortana、Google Allo、度秘、灵犀	Quartz、今日头条

在这四个功能中，知识问答和推荐功能的评价指标较为客观，评价方式也相对成熟。而对于聊天功能，给定相同的输入，其回复形式可以多种多样，如表2所示。对于用户的一个输入，通常有多种合

表2 聊天输入及回复示例

输入	回复
嗨，你今天怎么样？	非常好啊
	一般般
	今天真的糟透了
变形金刚5真烂	不能同意更多
	就为了音响和视觉效果去的
	是你没品味吧？

理的回复，且数目不固定，这使得聊天中的回复评价很难通过一种客观的机制进行，而必须加入人的主观判断作为评价依据。类似地，在任务执行类对话中，除了任务完成度这类客观指标外，也需要综合考虑用户满意度等主观评价指标。

开放域对话系统评价技术

当前，开放域对话系统评价的主流方法有两种，客观评价与模拟人工评价。其中客观评价指标又包括两类，一是以 BLEU[1]、METEOR[2] 和 ROUGE[3] 为代表的词重叠评价矩阵，二是以贪婪匹配方法 (Greedy Matching)[4]、向量均值法 (Embedding Average)[5]、向量极值法 (Vector Extrema)[6] 为代表的基于词向量的评价矩阵。模拟人工评价主要包括三种前沿的利用神经网络模拟人工评分的方法，即类生成式对抗网络 (Generative Adversarial Networks, GAN) 结构的对抗评价模型、基于循环神经网络 (RNN) 的 ADEM(Automatic Dialogue Evaluation Model) 对话评价系统和基于人工神经网络 (ANN) 结构的对话评价系统。

客观评价

对于评价开放域对话系统回复的质量，一类方法是计算对话系统生成的回复与标准答案之间的词重叠率。其中，BLEU 和 METEOR 是在机器翻译任务中取得很好效果的两种评价方法，ROUGE 也在文本的自动摘要任务中取得了不错的评价效果，业界普遍认为这些指标可以准确反映生成结果的部分特征。

1.BLEU 是指对模型输出和参考答案的 n-gram[1] 进行比较并计算匹配片段个数的方法。这些匹配片段与它们在上下文 (Context) 中存在的位置无关，这里仅认为匹配片段数越多，模型输出的质量越好。

2.METEOR 矩阵会在候选答案与目标回复之间产生一个明确的分界线（这个分界线是基于一定优先级顺序确定的，优先级从高到低依次是：特定的序列匹配、同义词、词根和词缀、释义）。有了分界线，METEOR 可以把参考答案与模型输出的精度 (Precision) 与召回率 (Recall) 的调和平均值作为结果进行评价。

3.ROUGE 是一系列用于自动生成文本摘要的评价矩阵，记为 ROUGE-L。它是通过对候选句与目标句之间的最长相同子序列 (Longest Common Subsequence, LCS) 计算 F 值 (F-measure) 得到的，LCS 是在两句话中都按相同次序出现的一组词序列，与 n-gram 不同的是，LCS 不需要保持连续（即在 LCS 中间可以出现其他的词）。

虽然这三个指标还没有彻底适应对话系统评价的任务，但从评价词重叠率的角度看，这类方法仍然值得尝试、改进和关注。

另一种评价对话生成效果的思路是通过了解每一个词的语义来判断回复的相关性，词向量是实现这种评价方法的基础。通常采用 Word2Vec[7] 等方法，给每一个词分配一个向量用于表示这个词，通过计算每个词在语料库中出现的频率来近似地表示这个词所表达的语义。所有的词向量矩阵通过向量连接就可以近似地表示句子级的语义向量，通过这种方法可以分别得到候选回复句与目标回复句的句向量，再通过句向量的余弦距离计算二者的相似度。根据词向量相似性的计算方式，有三种评价方法。

1. **贪婪匹配方法。**这是基于词级别的一种矩阵匹配方法。给定两个句子，句子间每两个词的余弦相似度形成一个矩阵。其计算思想是将该矩阵每行（列）的最大值累加后再除以行（列）数，最终对行和列计算得出的数值再取平均值。

2. **向量均值法。**该方法通过对句子中每一个

[1] 在计算机语言学和概率论范畴内，n-gram是指给定的一段文本或语音中N个项目(item)的序列。项目(item)可以是音节、字母、单词（从序列的角度看甚至可以是碱基对）。通常n-gram取自文本或语料库。人们基于一定的语料库，可以利用n-gram来预计或者评估一个句子是否合理。n-gram还用来评估两个字符串之间的差异程度，这是模糊匹配中常用的一种手段。

词的向量求均值来计算句子的向量。除对话系统之外的很多自然语言处理任务都应用过这种方法，例如计算文本相似度的任务。

3. 向量极值法。这是另一种在句子级向量上计算相似度的方法，筛选词向量的每一维然后选择整句话中极值最大的一维作为这个句子的向量表示。

想要更准确地表达两个回复的相似度，仅计算向量极值是不够的，还需要计算回复之间的余弦距离，才能更好地表示它们之间的相似程度。直观上看，在某个文本中具有特殊意义的词应当具有比常用表达更高的优先级，但由于常用表达往往会出现在更多的文本中，这种计算方法会使得它们在向量空间中离得更近，计算相似度之后常用表达就会占据输出向量排序更靠前的位置，使具有重要语义信息的词被“挤”到靠后的位置。因此，在采用向量极值法时需有意识地忽略常用表达 [8]。

模拟人工评价

模拟人工评价，顾名思义即利用模型来模拟人工评价的指标，提高模型评价与人类评价的相关度。目前，模拟人工评价的方法主要有两种，一种是类 GAN 模型，另一种是 RNN 模型。

1. 类 GAN 模型

GAN 模型在 2014 年由伊恩·古德费洛 (Ian Goodfellow) 提出，至今不仅催生了很多论文，也带来了许多应用，成为人工智能学界的一个热门研究方向。Google Brain 团队的安居里·卡南 (Anjuli Kanan) 和 Google DeepMind 团队的奥里奥尔·维亚尔斯 (Oriol Vinyals) 等人于 2017 年 1 月提出了类 GAN 结构的对抗评价模型，并设计了一种类 GAN 的网络结构，用于直观评价生成器 (Generator) 产生的回复结果与人类回复的相似程度。受到 GAN 在图像生成任务上的成果启发，安居里等人的工作 [9] 也采用了 GAN 的基本生成器 - 判别器 (Generator-Discriminator) 结构，通过训练得到的生成器用于生成回复，判别器用于区分人的回复与生成器生成的结果。与传统的 GAN 的基本结构不同，模型的生成器是一个序列对序列 (Sequence to Sequence, Seq2Seq) 模型，包含一个完整的 RNN 编码解码结构，而判别器虽然也是一个 RNN，但采用的是编码器加一个二元分类器的结构。实验结果表明，判别器的区分效果与回复的长短有很大关系，随着回复长度的增加，判别器的区分能力越强。

2. RNN 模型

2016 年 6 月，麦吉尔大学 (McGill University) 的瑞恩 (Ryan) 等人研究发现，传统的客观评价指标都具有一定的局限性，无法很完整地表示评分与人类评价的相关度，于是尝试使用 RNN 的方法进行自动评分模型的训练，并提出了 ADEM 用于预测回复的人工评价结果，同时也将 ADEM 的评分结果与传统指标 BLEU 和 ROUGE 进行了对比，证明了自动评价系统的可行性。

ADEM 是一个通过半监督性学习方法训练得到的多层 RNN 结构的评价模型，使用了多层编码器 (Encoder) 将训练语料中文本转化为向量，训练阶段的输入为对话文本、生成回复及参考回复。ADEM 中的编码器将这些语料分别转化为向量，然后通过对这些向量进行线性变换得到一个分数。考虑到人工标注数据费时费力，希望训练过程能够用更少的标注数据而达到更准确的预测效果，所以采用了预训练的方法学习编码器的参数，将原模型中的编码器产生的结果输入到一个独立的 RNN，然后经过对这个 RNN 的训练产生特定条件下对特定上文的回复，并把这些数据当作原 RNN 的训练数据。这样，同样的上文就可以产生许多句不同的回复，从而得到更多的训练数据。实验使用了 Twitter 数据集，实验之前，作者先通过亚马逊土耳其机器人 (Amazon Mechanical Turk, AMT) 的志愿者对数据集中给出的不同问题的不同回复进行评分，并且对人工评价的分数做了分析，根据人工评价得到的分数特点结合现有的上文与回复，再将通过预训练方法生成的大量数据加入到实验数据集中。通过计算自动评价与人工评价之间的 Spearman 和 Pearson 的相关性，表明 ADEM 的评价结果要好于 BLEU 和 ROUGE[10]。

任务型对话系统评价技术

随着任务型对话系统的诞生，与其对应的评价方法也逐渐成为一个活跃的研究方向。1997 年沃克 (Walker) 提出了一个将对话持续时间及其他许多特征融入线性方程的系统 PARADISE，用于推测用户的满意度 [11]。该系统主要采用一个已标注用户满意度的对话数据集和一个客观评价的数据集，通过线性回归方法对已标注的数据集求出一个可以表示用户满意度的权重指标。指标的决定因素是对话成功率和对话成本消耗（如对话时长、系统给出确认性质回复的次数等），再通过强化学习将这个指标变成一个损失函数作为网络的奖励 (Reward)。由于该方法很好地考虑了对话系统的多个不同因素，后来被用于对话策略学习 [12]。实际操作中发现，对话系统的成功率和对话的长度是最重要的两个指标，因此，后续的对话系统评测也往往将最大化成功率与最小化对话长度作为任务型对话系统评测的指标。

然而，当系统真正地与人进行交互时，任务完成的程度是很难界定的，不仅如此，生成模型理论上的有效性等一系列问题使得这种评价系统的效果不尽如人意 [13]。因此，基于标注语料的数据驱动型对话评价模型成为广泛讨论的方向。2012 年，有研究者提出用协同过滤的方法实现对用户反馈的表示 [14]，利用重塑反馈函数也可以达到加速对话策略学习的目的 [15]。于尔泰斯 (Ultes) 与明克 (Minker) 等人的研究发现，专家满意度对对话系统的回复成功率影响很大 [16]。所有方法和尝试都表明，优质的训练数据对于对话系统的生成结果至关重要。但是得到优质的标注数据是非常困难的，耗费大量的专家资源来对数据进行系统完整的标注是非常不明智的，所以有研究者提出用机器模拟人类标注数据的过程，这样既能减少人工消耗，也可以产生更多的可用数据。基于这个想法，有研究者提出了主动学习 (Active Learning) 的方法，为了减少标注误差而采用多种方式相结合的办法来对数据进行自动标注 [17]。

史蒂夫·杨 (Steve Young) 等人在 2012 年总结了任务型对话系统评价的基本情况 [18]，即数据驱动的自然语言处理任务评价有很多方法，但由于多轮交互性，对对话系统的评价难度更大，尽管目前已经有很多针对不同指标的评价矩阵（即将客观指标用特征矩阵的方式表示，从而达到可运算的目的），但如何将他们很好地结合起来用于评价整个对话系统仍是个难题。事实上，对话系统评价的最终目标是测评用户的满意度，但总有许多因素使我们无法将评价结果与用户的体验感受完全吻合，即使通过给出的人工制定的评价指标来对一个系统进行测试，也会造成不同程度的偏差，而且也很难将所有的特征罗列出来并加以对比达到更好的评价效果，因此现有的评价过程大都无法准确地满足用户的要求。针对这些问题，文章提出了两个任务型对话系统评价的对策：(1) 通过构造某种特定形式的用户模拟系统进行评价；(2) 人工评价。

用户模拟评价

用户模拟是最有效最简单的评价策略，并且是最有可能覆盖最大对话空间的方法，因为通过模拟不同情境下的对话，可以在大范围内进行有效的测试和评价 [19~21]。然而，这种方法的缺点也很明显，就是真实用户的反应与模拟器的反应之间潜在的矛盾，这个矛盾的影响大小某种程度上取决于用户模拟器的好坏。即使这个矛盾无法解决，用户模拟仍然是任务型对话系统评价中最常用的评价方法，曾被用于评价多种不同的基于部分可观察马尔可夫决策过程 (Partially Observable Markov Decision Process, POMDP) 的对话策略 [22, 23]。

人工评价

人工评价是指通过雇佣测试人员对对话系统生成的结果进行评价，这样做的好处是能够产生更多真实的评价数据。目前，这种评价方法更多出现在实验室等研究资源雄厚的环境中，测试人员在预定任务领域内对系统进行评测，通过一些预设的询问方式与系统进行对话，根据对话结果对系统的表现

进行评分。

通过 POMDP 进行对话策略学习的对话系统已经采用了这种评价方法，并将结果应用于人工编码策略与马尔可夫决策过程 (Markov Decision Process, MDP) 策略的结果提升任务中 [23~25]。不论评价者是否能够全方位地代表用户，这种方法最大的问题在于如何雇佣足够多的测评人员，很明显这需要大量的开销，后期出现的外包模式以及借助网络媒介延迟较小的特点在网络上进行实时评价等方法都可能解决该问题。例如使用 AMT 服务 [26]，给出预定义的任务以及基础的培训指令，雇佣评测员对指定的任务进行评测，评测员可以通过免费电话对对话系统进行评价，每次对话后给出反馈信息。该方法可以有效地产生大量对话系统与人的真实对话数据，从而产生大量的数据统计结果 [27]。

除开销巨大外，这种方法还存在外包选择的评测人员是否真的能够代表所有用户的问题，事实上，如果没有很好地监控实验集合，人工评价动机和目的将成为最后评价结果的重要影响因素，也有事实证明人工评价并没有非常准确地表现出对话的准确程度 [28]。

总结及展望

本文介绍了目前对话领域的两种主流评价思路。虽然每种方法都有优势，但也有不足，对话系统评价研究还有很大的空间需要探索。

通过对客观指标进行评分，可以很好地根据每个指标对对话系统模型进行修改和提高，并根据模拟人工评分的结果，对评价本身有一个整体的宏观认知，这些为对话系统评价提供了研究思路。对于任务型对话系统的评价仍然需要继续提高与人工评价结果的拟合程度；对于开放域对话系统，如何利用客观指标和模拟评分二者不同的特点来提高评价的效率和准确率，将是未来研究的重点；如何减轻模型比较和选择的负担，增强评价系统的可扩展性，实现一种可迁移到不同数据集中进行评价任务的评价方法，也是未来的重要研究方向。 ■

张伟男

CCF专业会员。哈尔滨工业大学讲师、硕士生导师。主要研究方向为人机对话与自然语言处理。
wnzhang@ir.hit.edu.cn

车万翔

CCF高级会员。哈尔滨工业大学副教授、博士生导师。主要研究方向为自然语言处理。
car@ir.hit.edu.cn

参考文献

[1] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//*In Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL)*, 2002: 311-318.

[2] Banerjee S, Lavie A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments[C]//*In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005: 65-72.

[3] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//*In Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004.

[4] Rus V, Lintean M. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics[C]//*In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012: 157-162.

[5] Wieting J, Bansal M, Gimpel K, et al. Towards universal paraphrastic sentence embeddings[C]//*CoRR*, abs/1511.08198, 2015.

[6] Forgues G, Pineau J, Larcheveque J M, et al. Bootstrapping dialog systems with word embeddings[J]. 2014.

[7] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. *In Advances in neural information processing systems*, 2013: 3111-3119.

[8] 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[OL].(2017). arXiv preprint arXiv: 1603.08023.

更多参考文献：www.ccf.org.cn/sztsg/cbw/zgjsjxhtx/