智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

中文多技能对话评估

柳泽明, 程子豪, 刘晶晶, 杨晓, 郭园方, 王蕴红

引用本文:
柳泽明, 程子豪, 刘晶晶, 等. 中文多技能对话评估[J]. 智能系统学报, 2025, 20(5): 1281-1293.
LIU Zeming, CHENG Zihao, LIU Jingjing, et al. Evaluation of Chinese multiskill dialogues[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1281-1293.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202411001

DOI: 10.11992/tis.202411001
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250625.1316.002

中文多技能对话评估

柳泽明, 程子豪, 刘晶晶, 杨晓, 郭园方, 王蕴红
(北京航空航天大学 计算机学院, 北京 100191)

摘　要: 准确评估多技能对话系统的能力, 对满足用户多样化的需求, 例如社交闲聊、深入的知识对话、角色化聊天以及对话推荐至关重要。现有的基准仅针对特定对话技能的评估, 无法有效地同时评估多种对话技能。为解决这一问题, 本文构建了一个中文多技能评估基准 (multi-skill dialogue evaluation benchmark, MSDE), 它包含 1 781 个对话和 21 218 条话语, 覆盖 4 类常见的对话任务, 即闲聊、知识对话、画像聊天和对话推荐。然后, 本文基于 MSDE 做了大量实验, 并分析了自动评估指标和人工评估指标的相关性。实验结果表明: 1) 在 4 类对话任务中, 闲聊最难评估, 知识对话最容易评估。2) 不同指标在 MSDE 上的表现存在明显差异。3) 对于人工评估, 各指标在不同对话任务上的评估难度不同。部分数据发布在 https://github.com/IRIP-LLM/MSDE, 全部数据将在整理后发布。

关键词: 多技能对话; 对话评估; 闲聊; 开放域对话; 对话推荐; 画像聊天; 知识对话; 大语言模型
中图分类号: TP39　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)05−1281−13

中文引用格式: 柳泽明, 程子豪, 刘晶晶, 等. 中文多技能对话评估 [J]. 智能系统学报, 2025, 20(5): 1281–1293.
英文引用格式: LIU Zeming, CHENG Zihao, LIU Jingjing, et al. Evaluation of Chinese multiskill dialogues[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1281–1293.

Evaluation of Chinese multiskill dialogues

LIU Zeming, CHENG Zihao, LIU Jingjing, YANG Xiao, GUO Yuanfang, WANG Yunhong
(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: The accurate evaluation of the capabilities of a multiskilled dialogue system is important to satisfy the different demands of users, including social banter, profound knowledge-based discussions, role-playing conversations, and dialogue recommendations. Current benchmarks concentrate on assessing specific dialogue skills and cannot efficiently evaluate multiple dialogue skills concurrently. To facilitate the evaluation of multiskill dialogues, this study establishes a Chinese multiskill evaluation benchmark, which is the Multi-Skill Dialogue Evaluation Benchmark (MSDE). MSDE contains 1,781 dialogues and 21,218 utterances, which cover four common dialogue tasks: chit-chat, knowledge dialog, persona-based dialog, and dialog recommendations. We performed extensive experiments on MSDE and examined the correlation between automatic and human evaluation metrics. Results indicate that (1) among the four dialogue tasks, chit-chat is the most difficult to analyze, while knowledge dialogue is the easiest; (2) significant differences exist in the performance of various metrics on MSDE; (3) for human evaluation, the analysis complexity of each metric differs across varying dialogue tasks. Certain data will be made available on https://github.com/IRIP-LLM/MSDE, and all data will be released after sorting.

Keywords: multiskill dialogue; dialogue evaluation; chit-chat; open domain dialogue; conversational recommendation; persona-chat; knowledge-grounded dialogue; large language model

开发高质量的对话系统是人工智能的标志性任务之一。近年来, 随着深度学习和包括 Qwen[1]、Baichuan[2]、Llama[3] 和 ChatGLM[4] 在内的大模型技术的发展, 以及诸如 Meena[5]、BlenderBot[6] 和小冰[7-8] 等基于语音的聊天机器人的兴起[9], 不同对话任务的研究数量均有显著增加, 并取得了较大进展。

收稿日期: 2024−11−01.　**网络出版日期:** 2025−06−25.
基金项目: 国家重点研发计划项目 (2023YFF0725600); 国家自然科学基金项目 (62406015).
通信作者: 王蕴红. E-mail: yhwang@buaa.edu.cn.

但是，之前的研究大多分别研究不同的对话任务[10-11]，而在真实场景的人机对话中，用户以为对话系统是万能的，并期望对话系统能满足其多样化的需求，这需要对话系统具备多种技能[12]。例如，对话推荐技能、画像聊天技能、知识对话技能和闲聊技能等，因此，多技能对话具有很高的研究和产业应用价值。对于多技能对话研究，已有一些工作提出构建多技能对话基准数据集：dodecaDialogue[7] 是一个包含 12 个独立数据集的英文多技能基准数据集，其中每个独立数据集都对应一种对话技能；BlendedSkillTalk[13] 将多种技能融合到一个英文对话中，评估跨不同知识、情感和人物画像的多技能对话生成。此外，还有一些研究人员构建多技能对话模型，比如 DuRecDial[14] 是一种模块化框架，来应对多技能对话带来的挑战，它利用对话系统中的不同组件处理特定对话任务，具备更大的灵活性和适应性，还有一部分研究人员通过端到端的形式 训练模型处理多种对话类型或对话技能的能力[2,4]。

然而，这些研究更多关注多技能对话的建模方式和数据构建，缺乏对多技能对话评估方法的探索，这阻碍了多技能对话的发展。具体来说，现有的评估方法都是分别评估模型的不同技能，不同的自动指标在不同任务上的有效性会有很大差异，以这些指标来评估，不能有效地反映模型的对话能力[15-16]，这不仅影响了对多技能对话系统综合性能的准确评估，还阻碍了这些系统在多样化场景中的广泛应用。因此，亟须构建能够跨任务使用的多技能对话评估基准，以推动该领域的全面发展。

为促进多技能对话的发展，并评估各自动评估指标的有效性，本文首次提出了一个新的任务，即多技能对话评估。该评估旨在全面评估模型在多种常见对话任务上的生成能力，包括闲聊、知识对话、画像聊天和对话推荐等。为解决这个问题，本文构建了首个中文多技能对话评估基准 (multi-skill dialogue evaluation benchmark, MSDE)，其包含 1 781 个对话和 21 218 条话语，覆盖了 4 种对话类型，为多技能对话的评估奠定了基础。

基于 MSDE，本文使用多个模型做了大量实验，并分析了自动评估指标和人工评估指标在不同任务上的皮尔逊相关性 (Pearson correlation coefficient, PE) 和斯皮尔曼相关性 (Spearman's rank correlation coefficient, SP)[17]，具体来说，相关性较高证明该自动评估指标更加符合人类评估，相关性较低则证明该自动评估指标不能反映人类评估。本文发现在 4 类任务中，闲聊最难评估，知识对话最容易评估。同时在自动指标中 BERTscore 和 METEOR 与人类评估相关性较高，DIST1 和 DIST2 相关性较低。

综上所述，本文的主要贡献为：

1) 本文首次提出了多技能对话评估这一新的任务。

2) 为了解决这个问题，本文提出了首个中文多技能对话评估基准 MSDE。MSDE 包括 1 781 个对话和 21 218 条话语，覆盖 4 种常见的对话任务：闲聊、知识对话、画像聊天和对话推荐。

3) 基于 MSDE，本文做了大量实验。实验结果证明，METEOR 和 BERTscore 等指标在多技能对话评估中表现较好，而 DIST 指标在多技能对话评估中表现较差。另外在闲聊任务中，自动评估指标与人工评估指标的相关性普遍较低，难以通过自动评估指标进行精准的评估，而知识对话相对比较容易评估。

1　相关工作

为了促进对对话自动评估指标 (例如 BLEU[18]、METEOR[19] 等) 与人工评估相关性的研究，先前的一些研究已经创建了多个人工标注的数据集[20]，如表 1 所示。一般来说，这些数据集是通过以下步骤构建的：1) 选择一个现有的对话数据集 (例如 DailyDialog[21]、TopicalChat[22]、PersonaChat[23] 和 ConvAI2[24])；2) 在所选的对话数据集上训练一个回复生成模型 (例如 Transformer、Seq2seq 和 LSTM)[25]，并为所选数据集的验证集/测试集生成回复；3) 为生成的回复收集人工质量标注 (例如相关性、一致性、连贯性、吸引力和多样性)[26-27]。所有这些数据集都是通过静态评估收集的，在这种评估中，标注人员不直接与对话系统 (用于评估的模型) 交互，而是离线标注模型生成的回复。与它们相比，MSDE 中的每个对话都是通过人机交互的方式评估并收集的[28]。尽管和静态评估相比，交互式评估需要更多的时间和人力，但它能够让评估者通过交互更精确地了解对话系统的能力[29]。

表 1　MSDE 与其他对话评估数据集的对比

Table 1　Comparison of MSDE with other datasets for dialogue evaluation

数据集	是否支持多技能评估	是否通过人机交互评估	语言	对话条数	语句数
USR-TopicalChat[30]	否	否	英文	300	3 660

续表 1

数据集	是否支持多技能评估	是否通过人机交互评估	语言	对话条数	语句数
USR-PersonaChat[30]	否	否	英文	240	2 232
GRADE-ConvAI2[31]	否	否	英文	600	1 200
GRADE-DailyDialog[31]	否	否	英文	300	600
HolisticEval-DailyDialog[32]	否	否	英文	200	200
PredictiveEngage-DailyDialog[33]	否	否	英文	600	600
GRADE-EmpatheticDialogue[31]	否	否	英文	300	600
DSTC6[34]	否	否	英文	40 000	64 000
FED[35]	否	否	英文	500	5 900
DSTC9[36]	否	否	英文	2 200	59 840
MSDE(本文数据集)	是	是	中文	1 781	21 218

此外, MT-Bench 研究了大模型 (Claude-v1、GPT-3.5 和 GPT-4 等) 作为裁判进行评估和人工评估的相关性[37], 并没有分析传统自动指标 (例如 BLEU、METEOR 等) 和人工评估的相关性。而本文研究的是传统自动评估指标与人工评估的相关性。

2　数据集选取

本文从 Lic 2021[38] 和 CCF LUGE[39] 这 2 个比赛的数据中选取了开放域对话、知识对话、对话推荐以及画像聊天这 4 个对话任务的数据集构成 MSDE 数据集。之后, 对于每个比赛, 分别为其选取 14 个由不同模型组成的队伍, 使用这些队伍所代表的模型, 获得 2 个比赛数据上的模型生成回复。

2.1　开放域对话

LCCC[40] 是一个从新浪微博上抓取的大型中文短文本对话数据集, 该数据集中的每条数据包含一轮或是多轮的对话数据。这些对话数据来源于不同话题与背景下的新浪微博用户对话, 决定了 LCCC 数据集包含领域的开放性。

2.2　知识对话

DuConv[41] 是一个基于知识的中文真人对话数据集, 它利用结构化知识图谱和非结构化文本 (例如电影评论) 来构建知识图谱。对于给定的知识图谱, 将采样两个相关实体, 其中一个作为过渡主题, 另一个作为最终目标主题, 以构建一条知识路径来引导众包工作者收集对话。此外, 机器也可以使用这条知识路径来引导对话进行, 使其自然地走向最终目标话题。

2.3　对话推荐

DuRecDial[14] 是一个多轮中文真人对话推荐数据集, 包含 7 个应用领域及多种对话类型 (例如对话推荐、闲聊、任务对话和问答)。每个对话都附有一个知识图谱、一个搜索者介绍和一个任务模板。知识图谱使用结构化和非结构化知识构建。每个搜索者都有一个明确独特的介绍, 包含姓名、性别、年龄、居住城市、职业以及对领域和实体的偏好。任务模板包含: 1) 目标序列, 由对话类型和对话主题 2 个元素组成。2) 每个目标的详细描述。

2.4　画像聊天

CPC[42] 是一个使用角色画像的中文画像聊天数据集。在 CPC 中, 聊天机器人和用户在当前对话之前已经进行了一些对话, 机器人能够从这些历史对话中获取用户的个人信息, 因此机器人能够根据这些信息生成用户的个性画像。在每次聊天中, 机器人应该基于已知用户画像来控制对话有效进行且连贯深入。

3　人工质量标注

多技能的中文对话评估基准测量了自动指标与人类判断的相关性, 并用自动指标与人类判断的相关性来评估自动指标。本章描述了 MSDE 构建的 2 个关键步骤: 1) 众包工作者收集对话, 2) 众包工作者人工标注。

3.1　对话收集

本文开发了一个接口, 通过该接口, 大约 12 名众包员工与表 2 中的 20 种模型组合进行对话。具体来说, 基于对话附加信息 (例如, 知识对话中的知识, 对话推荐中的知识和推荐目标, 画像聊天中的角色信息), 标注者与每个模型进行 4 种类型的对话, 并收集对话大约 150 个。对于每个对话, 要求所有给定的信息都参与聊天, 或者达到 10 个回合 (20 句话)。然后, 3 名具有对话质量评估经验的数据专家将会对收集到的对话进行评估, 以进行质量验证。如果没有通过评估, 会要求众包工作人员按照上述方式重新收集对话, 直到符合标准 (如众包员工的话语应流利、自然、恰当、连贯和知识准确)。

表 2　Lic 2021 与 CCF LUGE 上使用的对话模型
Table 2　Models for Lic 2021 and CCF LUGE

序号	Lic 2021	CCF LUGE
1	RoBERT+MacBERT	BertSum+Seq2seq
2	PLATO+UniLM	RoBERTa+UniLM
3	RoBERTa	NEZHA+UniLM
4	BRoBERTa+UniLM	RoBERTa
5	ERNIE-GEN	BERT+GPT-2
6	BERT	DCial+PostKS
7	PLATO	GPT-2
8	NEZHA+UniLM	CDial+PostKS+KBRD
9	BERT+GPT-2	PLATO+PostKS
10	BERT-Chinese	BERT-Chinese
11	Baichuan2-7B-chat	
12	Qwen-7B-chat	
13	Llama3-8B-chinese-chat	
14	Chatglm4-9B	

3.2　质量标注

为了评估自动指标与人工判断的相关性，在对话收集后进行了人工质量标注。约 10 名具有对话经历的数据专家依次在丰富度 (0~2)、连贯性 (0~2)、知识准确率 (0~2) 和推荐成功率 (0~2) 等方面衡量收集到的对话。

3.2.1　丰富度

丰富度 (information-richness, Info.) 评估对话中提供的知识 (目标话题和话题属性) 的多少。

1) 分数 0(差): 没有提及任何知识。

2) 分数 1(一般): 只提及了一个知识点。

3) 分数 2(好): 提及了多个知识点。

3.2.2　连贯性

连贯性 (coherence, Coh.) 评估每个回复在给定当前目标和全局上下文时的流畅性、相关性和逻辑一致性。

1) 分数 0(差): 超过 2/3 的回复与给定的当前目标和全局上下文不相关或逻辑矛盾。

2) 分数 1(一般): 超过 1/3 的回复与给定的当前目标和全局上下文不相关或逻辑矛盾。

3) 分数 2(好): 少于 1/3 的回复与给定的当前目标和全局上下文不相关或逻辑矛盾。

3.2.3　知识准确性

知识准确性 (knowledge fidelity, Konw.) 评估回复中使用知识的正确性。

1) 分数 0(差): 所有使用的知识都是错误的，或没有使用任何知识。

2) 分数 1(一般): 部分使用的知识是正确的。

3) 分数 2(好): 所有使用的知识都是正确的。

3.2.4　推荐成功率

推荐成功率 (recommendation accuracy, Rec.) 评估用户是否最终接受了对话中的推荐。

1) 分数 0(差): 用户没有接受推荐。

2) 分数 1(一般): 用户部分接受推荐。

3) 分数 2(好): 用户完全接受推荐。

3.3　数据集统计

表 3 给出了 MSDE 的统计信息，标明了 MSDE 涉及的多对话任务类型以及丰富的人工标注信息。MSDE 包含了 4 种对话类型，每种对话类型都包含数百条对话，并且 MSDE 针对不同对话类型，分别选择了合适的人工评估标准进行标注。MSDE 的复杂性使其成为一个评估多技能对话 (多对话相关任务) 的优秀测试方法。

表 3　MSDE 统计
Table 3　Statistics of MSDE

数据集	对话类型	对话数量	话语数量	评估颗粒度
LCCC	闲聊	300	600	丰富度、连贯性
DuConv	知识对话	591	5 145	丰富度、知识准确率、连贯性
DuRecDial	对话推荐	590	10 593	丰富度、知识准确率、连贯性、推荐成功率
CPC	画像聊天	300	4 880	丰富度、知识准确率、连贯性

4　自动评估指标概述

本节介绍 4 类常见的自动评估指标，包括基于词汇重叠的评估指标，基于语义词汇重叠的评估指标和基于嵌入的评估指标和基于学习的评估指标。

4.1　基于词汇重叠的评估指标

F1 是一种广泛应用的评估指标，它通过计算准确率和召回率的调和平均值，综合衡量生成的文本质量。该指标在评估生成回复与目标回复之间的信息匹配程度时尤为有效，既强调生成回复中信息的覆盖面 (召回率)，又关注其与目标文本的一致性 (准确率)。

BLEU[18] 是一种常用的基于词汇重叠的评估指标，通常用于自然语言生成任务，如机器翻译和对话生成等。在对话评估中，它通过分析生成的回复和真实回复中的 n-gram 重叠率来评估生

成的质量。其中 n-gram 的核心思想是计算两个文本片段之间共享的连续 N 个词或字符占整个文本片段的比例。

ROUGE[43] (recall-oriented understudy for Gisting evaluation) 是一组用于评估文本生成的指标。本文使用 ROUGE-l, 它基于生成的回复和真实回复之间的最长公共子序列 (longest common subsequence, LCS) 计算 F1 值。LCS 指标是 2 个文本在相同词汇序列上的匹配程度, 与 n-gram 不同的是, 这些词汇不一定是连续出现的, 即在 LCS 中的词汇间可以夹杂其他词汇。

METEOR[19] 解决了 BLEU 一些局限性, 旨在改进对文本生成的评估。具体来说, METEOR 通过对生成文本和目标文本之间的词汇进行多层次匹配, 包括词形、词干、同义词和短语等, 并对匹配项进行对齐分析。其最终分数通过准确率和召回率的调和平均值计算得出, 提供了比单一 n-gram 重叠更为细致的评估。

DIST(distinct) [28] 是一组用于评估生成文本多样性的指标。它通过计算生成回复中非重复 n-gram 的比例来评估文本的多样性, 能够有效反映生成文本的多样性和避免冗余的能力, 特别适用于开放域对话生成任务。

CIDEr(consensus-based image description evaluation)[29] 是一种用于图像描述任务的自动评估指标。在评估对话生成的回复时, CIDEr 用于计算生成回复和真实回复间的相似度。

4.2　基于语义词汇重叠的评估指标

BERTscore(BS)[44] 是一种基于预训练 BERT 模型的上下文相关词嵌入的文本生成评估指标。具体来说, BERTscore 通过计算生成回复与真实回复中词嵌入的余弦相似度来评估生成的语义相关性。在对话评估中, BERTscore 通过匹配生成的回复和真实回复中的词嵌入, 计算 F1 值。

4.3　基于嵌入的评估指标

Skip-Thought(ST)[45] 模型是一种无监督学习模型, 它使用循环神经网络对给定句子进行编码, 将其转化为向量表示, 并基于这些向量来计算标准回复和预测回复之间的相关性指标。

Embedding Ave(EV)[46] 通过对句子中各个词的嵌入进行平均来计算句子的嵌入。它假设每个词对语义的贡献相等, 从而通过平均词嵌入来衡量生成文本的质量:

$$\overline{\boldsymbol{e}_R}=\frac{\sum_{w\in R}\boldsymbol{e}_w}{\left|\sum_{w'\in R}\boldsymbol{e}_{w'}\right|}$$

式中向量$\boldsymbol{e}_w$代表生成的回复R中单词w的词嵌入。

Vector Extrema(VE) 通过在组成回复的词嵌入中选择每个维度的最大值来计算句子的嵌入。这种方法从词嵌入中提取最具代表性的语义信息, 从而生成一个具有区分性的句子嵌入:

$$\boldsymbol{e}_{rd}=\begin{cases}\max\limits_{w\in R}\boldsymbol{e}_{wd}, & \boldsymbol{e}_{wd}>\left|\min\limits_{w'\in R}\boldsymbol{e}_{w'd}\right|\\ \min\limits_{w\in R}\boldsymbol{e}_{wd}, & \text{其他}\end{cases}$$

式中: d代表词嵌入向量的维度, r代表目标回复。

Greedy Matching(GM) 直接计算生成回复与目标回复之间的相似性分数。

$$S(R,r)=\frac{\sum_{w\in R}\max\limits_{\hat{w}\in r}\cos\_\mathrm{sim}(\boldsymbol{e}_w,\boldsymbol{e}_{\hat{w}})}{|R|}$$

$$G(R,r)=\frac{S(R,r)+S(r,R)}{2}$$

式中: $S(R,r)$表示以R为参考进行贪婪匹配, $G(R,r)$表示 GM 分数, $\hat{w}$表示目标回复中单词的词嵌入, cos_sim(·)表示计算余弦相似度, $|R|$表示生成回复R中单词的个数。

4.4　基于学习的评估指标

BLEURT[17] 用于评估一般自然语言生成任务, 它通过生成合成数据预训练 BERT 模型, 并使用均方误差 (mean squared error, MSE) 损失进行微调, 以预测生成文本的质量分数。

5　实验和结果

5.1　实验设置

评估的模型: 表 2 提供了所有用于评价的模型信息。

自动评估指标: 本文将 16 个常见自动评估指标进行了比较, 具体包括 F1、BLEU1、BLEU2、ROUGE、METEOR、DIST1、DIST2、CIDEr、Skip-Thought(ST)、Embedding Ave(EV)、Vector Extrema (VE)、Greedy Matching(GM)、BERTscore P(BS-P)、BERTscore R(BS-R)、BERTscore F1(BS-F1) 和 BLEURT。这些指标适用于未针对特殊任务或数据集进行训练的多任务评估。

5.2　自动指标与人工指标的相关性

本文分别采用自动指标和人工指标对模型回复进行评估, 并计算二者之间的相关性。评估流程如图 1 所示, 自动评估结果如表 4~9 所示, 人工评估结果如表 10 所示, 自动评估结果与人工评估结果的相关性如表 11~13 所示。

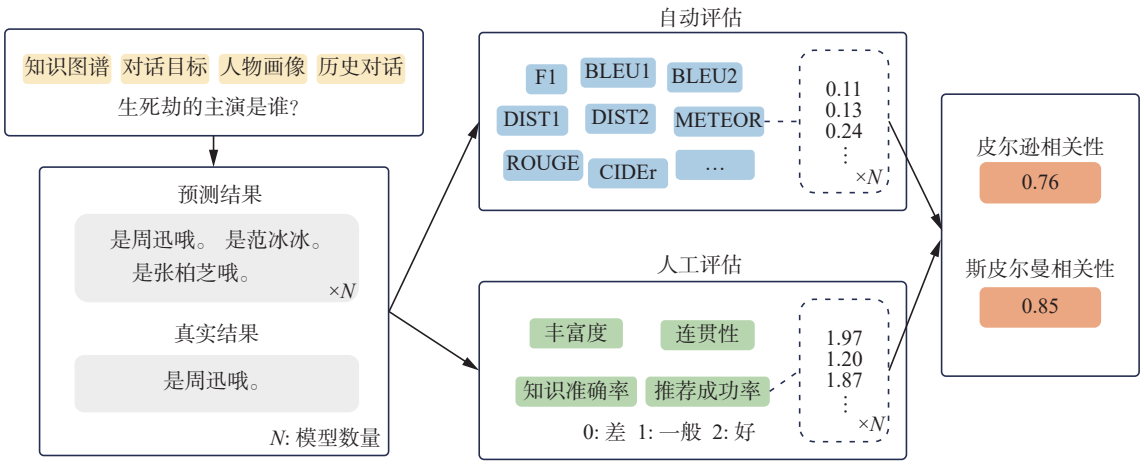

图 1　评估流程

Fig. 1　Evaluation process

表 4　Lic 知识对话任务各模型自动评估结果

Table 4　Automatic evaluation results of each model in the Lic knowledge-grounded dialogue task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	**0.49**	0.41	0.28	0.13	0.32	0.24	0.39	1.56	0.88	**0.98**	0.75	0.98	**0.75**	0.74	**0.74**	0.24
2	0.48	**0.44**	0.28	0.12	0.33	0.24	**0.40**	1.49	0.88	0.98	0.79	**0.99**	0.73	0.73	0.73	0.18
3	0.47	0.43	**0.29**	0.12	0.33	0.24	0.40	**1.63**	**0.88**	0.98	0.75	0.98	0.74	0.73	0.73	**0.24**
4	0.46	0.37	0.24	0.13	0.34	0.22	0.36	1.16	0.87	0.96	0.71	0.98	0.74	0.72	0.73	0.19
5	0.47	0.41	0.28	0.15	0.40	**0.24**	0.38	1.47	0.87	0.98	0.73	0.98	0.73	**0.74**	0.73	0.21
6	0.45	0.37	0.24	0.13	0.34	0.23	0.35	1.12	0.87	0.96	0.70	0.98	0.73	0.72	0.72	0.19
7	0.43	0.38	0.26	0.14	0.35	0.22	0.37	1.49	0.87	0.97	0.72	0.98	0.73	0.71	0.72	0.16
8	0.39	0.30	0.16	0.10	0.30	0.19	0.28	0.64	0.86	0.96	0.66	0.98	0.70	0.69	0.69	0.14
9	0.41	0.34	0.20	0.09	0.21	0.20	0.32	0.89	0.87	0.97	0.66	0.98	0.71	0.71	0.71	0.18
10	0.33	0.28	0.13	0.13	0.35	0.15	0.25	0.31	0.86	0.97	0.71	0.98	0.68	0.67	0.67	0.12
11	0.00	0.00	0.00	0.67	0.76	0.11	0.00	0.00	0.70	0.80	0.80	0.80	0.52	0.58	0.55	−0.25
12	0.12	0.10	0.04	0.26	0.61	0.13	0.02	0.00	0.68	0.81	0.81	0.81	0.57	0.61	0.59	−0.18
13	0.02	0.01	0.00	0.57	0.76	0.15	0.07	0.00	0.69	0.84	0.84	0.84	0.54	0.66	0.59	−0.60
14	0.00	0.00	0.00	**0.95**	**0.92**	0.18	0.05	0.01	0.75	0.85	**0.85**	0.85	0.57	0.67	0.61	−0.30

注: 加粗表示本列最优结果。

表 5　Lic 对话推荐任务各模型自动评估结果

Table 5　Automatic evaluation results of each model in the Lic conversational recommendation task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	**0.51**	0.45	0.35	0.06	0.16	**0.28**	**0.50**	2.26	0.81	0.91	0.62	0.96	0.78	0.77	0.77	0.32
2	0.49	0.46	0.34	0.06	0.24	0.23	0.48	1.99	0.81	0.91	0.62	0.96	**0.78**	0.75	0.76	0.30
3	0.49	**0.47**	**0.36**	0.08	0.22	0.25	0.49	2.33	0.81	0.92	0.63	**0.96**	0.78	0.76	**0.77**	**0.33**
4	0.49	0.44	0.34	0.06	0.19	0.27	0.49	2.13	0.81	**0.93**	0.66	0.96	0.77	0.76	0.77	0.32
5	0.43	0.42	0.32	0.06	0.20	0.22	0.49	**2.34**	**0.82**	0.92	0.63	0.96	0.77	**0.77**	0.76	0.33
6	0.50	0.44	0.34	0.06	0.17	0.26	0.47	1.87	0.80	0.90	0.60	0.95	0.77	0.75	0.76	0.28
7	0.43	0.36	0.28	0.06	0.20	0.19	0.46	2.02	0.81	0.91	0.60	0.95	0.77	0.74	0.75	0.28
8	0.50	0.44	0.34	0.06	0.17	0.26	0.48	2.01	0.81	0.91	0.60	0.95	0.78	0.76	0.77	0.29
9	0.37	0.34	0.25	0.06	0.19	0.17	0.41	1.66	0.80	0.90	0.56	0.95	0.74	0.72	0.73	0.23
10	0.34	0.24	0.14	0.09	0.31	0.14	0.33	0.82	0.77	0.87	0.53	0.94	0.73	0.70	0.71	0.21
11	0.01	0.00	0.00	0.53	0.63	0.15	0.01	0.00	0.64	0.77	0.77	0.77	0.54	0.62	0.57	−0.26
12	0.24	0.23	0.14	0.15	0.48	0.26	0.09	0.00	0.72	0.83	**0.83**	0.83	0.64	0.70	0.67	0.18
13	0.33	0.03	0.01	0.33	0.52	0.18	0.07	0.00	0.63	0.78	0.78	0.78	0.55	0.66	0.59	−0.32
14	0.01	0.00	0.00	**0.83**	**0.79**	0.08	0.00	0.00	0.72	0.69	0.69	0.69	0.61	0.73	0.66	−0.09

注: 加粗表示本列最优结果。

表 6　Lic 画像聊天任务各模型自动评估结果

Table 6　Automatic evaluation results of each model in the Lic persona-chat task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	0.36	0.31	0.19	0.07	0.30	0.16	0.33	0.82	0.87	0.98	0.80	0.99	**0.72**	0.70	**0.71**	0.30
2	**0.37**	**0.37**	**0.20**	0.05	0.27	0.16	**0.34**	0.78	**0.87**	**0.98**	0.82	0.99	0.70	0.69	0.70	0.26
3	0.35	0.32	0.19	0.06	0.27	0.16	0.33	**0.85**	0.87	0.98	0.81	**0.99**	0.71	**0.70**	0.70	**0.30**
4	0.35	0.31	0.18	0.06	0.26	0.16	0.32	0.65	0.87	0.98	0.81	0.99	0.71	0.69	0.70	0.29
5	0.35	0.31	0.18	0.07	0.27	0.16	0.32	0.74	0.87	0.98	0.80	0.99	0.71	0.70	0.70	0.29
6	0.31	0.28	0.14	0.05	0.22	0.14	0.28	0.42	0.87	0.97	0.79	0.99	0.68	0.67	0.68	0.23
7	0.34	0.29	0.18	0.06	0.26	0.14	0.32	0.83	0.86	0.97	0.77	0.99	0.71	0.69	0.70	0.28
8	0.27	0.23	0.11	0.04	0.22	0.11	0.25	0.34	0.86	0.97	0.75	0.98	0.68	0.65	0.67	0.26
9	0.30	0.27	0.14	0.05	0.19	0.13	0.28	0.45	0.87	0.98	0.76	0.99	0.68	0.67	0.68	0.27
10	0.29	0.28	0.14	0.06	0.30	0.13	0.27	0.38	0.86	0.97	0.79	0.99	0.68	0.67	0.67	0.22
11	0.01	0.00	0.00	0.73	**0.88**	0.15	0.00	0.00	0.64	0.81	0.81	0.81	0.55	0.64	0.59	−0.28
12	0.20	0.17	0.06	0.16	0.57	0.18	0.00	0.00	0.80	0.83	**0.83**	0.83	0.62	0.66	0.64	0.29
13	0.01	0.00	0.00	0.79	0.87	0.14	0.00	0.00	0.73	0.81	0.81	0.81	0.54	0.63	0.58	−0.38
14	0.01	0.00	0.00	**0.83**	0.83	**0.30**	0.08	0.00	0.80	0.82	0.82	0.82	0.58	0.67	0.62	−0.07

注: 加粗表示本列最优结果。

表 7　CCF LUGE 闲聊任务各模型自动评估结果

Table 7　Automatic evaluation results of each models for the CCF LUGE chit-chat task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	**0.32**	**0.30**	**0.28**	0.10	0.52	0.19	**0.32**	**2.58**	**0.74**	**0.92**	0.74	0.96	**0.69**	**0.69**	**0.69**	**0.28**
2	0.13	0.10	0.04	0.05	0.17	0.05	0.11	0.16	0.63	0.91	0.74	0.96	0.61	0.57	0.59	0.02
3	0.13	0.11	0.05	0.06	0.34	0.06	0.10	0.12	0.67	0.90	0.71	**0.96**	0.58	0.58	0.58	0.06
4	0.12	0.10	0.03	0.04	0.24	0.05	0.08	0.10	0.64	0.90	0.71	0.96	0.59	0.57	0.58	0.04
5	0.13	0.11	0.03	0.04	0.20	0.05	0.10	0.13	0.70	0.84	0.47	0.93	0.60	0.59	0.59	0.08
6	0.12	0.11	0.04	0.05	0.27	0.05	0.09	0.13	0.65	0.90	0.72	0.96	0.59	0.58	0.58	0.05
7	0.14	0.12	0.04	0.04	0.22	0.06	0.11	0.12	0.68	0.91	0.71	0.96	0.58	0.59	0.58	0.04
8	0.05	0.05	0.01	0.03	0.27	0.02	0.04	0.03	0.63	0.87	0.60	0.94	0.56	0.54	0.55	0.00
9	0.14	0.13	0.06	0.10	0.52	0.07	0.10	0.15	0.68	0.91	0.71	0.96	0.58	0.59	0.58	0.08
10	0.14	0.10	0.04	0.04	0.20	0.06	0.11	0.15	0.63	0.91	0.75	0.96	0.61	0.58	0.59	0.07
11	0.00	0.00	0.00	0.59	0.65	0.06	0.00	0.00	0.64	0.67	0.67	0.67	0.48	0.57	0.52	−0.53
12	0.07	0.06	0.02	0.13	0.52	0.07	0.00	0.00	0.72	0.71	0.71	0.71	0.54	0.59	0.56	−0.03
13	0.00	0.00	0.00	0.64	0.58	0.06	0.00	0.00	0.65	0.66	0.66	0.66	0.46	0.56	0.51	−0.56
14	0.02	0.00	0.00	**0.76**	**0.76**	**0.30**	0.08	0.00	0.68	0.82	**0.82**	0.82	0.50	0.60	0.54	−0.49

注: 加粗表示本列最优结果。

表 8　CCF LUGE 知识对话任务各模型自动评估结果

Table 8　Automatic evaluation results of each models for the CCF LUGE knowledge-grounded conversation task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	0.23	0.20	0.09	0.06	0.15	0.11	0.19	0.31	0.85	0.95	0.67	0.98	0.61	0.62	0.62	0.05
2	**0.46**	**0.37**	**0.24**	0.13	0.33	**0.23**	**0.35**	**1.15**	**0.87**	0.96	0.70	0.98	0.73	0.72	**0.73**	**0.20**
3	0.40	0.27	0.15	0.09	0.27	0.22	0.28	0.50	0.87	0.97	0.61	0.97	0.67	**0.72**	0.69	0.01
4	0.36	0.34	0.20	0.07	0.15	0.17	0.31	0.86	0.85	0.96	0.68	0.97	0.69	0.67	0.68	0.10
5	0.39	0.36	0.21	0.06	0.16	0.19	0.33	0.95	0.86	**0.98**	0.75	**0.99**	0.70	0.69	0.70	0.18
6	0.37	0.32	0.19	0.06	0.16	0.19	0.33	0.88	0.86	0.97	0.69	0.98	0.69	0.70	0.69	0.13

续表 8

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
7	0.42	0.32	0.19	0.11	0.27	0.19	0.32	0.91	0.86	0.96	0.69	0.98	**0.73**	0.70	0.71	0.18
8	0.37	0.35	0.20	0.06	0.13	0.18	0.32	0.92	0.85	0.97	0.72	0.98	0.70	0.68	0.69	0.16
9	0.37	0.34	0.18	0.07	0.18	0.18	0.31	0.81	0.85	0.96	0.69	0.97	0.68	0.69	0.68	0.12
10	0.38	0.31	0.17	0.09	0.24	0.18	0.30	0.68	0.86	0.98	0.72	0.98	0.70	0.69	0.69	0.16
11	0.00	0.00	0.00	0.67	0.76	0.11	0.00	0.00	0.70	0.80	0.80	0.80	0.52	0.58	0.55	−0.25
12	0.11	0.10	0.04	0.26	0.61	0.13	0.02	0.00	0.68	0.81	0.81	0.81	0.57	0.61	0.59	−0.18
13	0.02	0.01	0.00	0.57	0.76	0.15	0.07	0.00	0.69	0.84	0.84	0.84	0.54	0.66	0.59	−0.6
14	0.00	0.00	0.00	**0.95**	**0.92**	0.18	0.05	0.01	0.75	0.85	**0.85**	0.85	0.57	0.67	0.61	−0.3

注: 加粗表示本列最优结果。

表 9　CCF LUGE 对话推荐任务各模型自动评估结果

Table 9　Automatic evaluation results of each models for the CCF LUGE conversational recommendation task

序号	F1	BLEU1	BLEU2	DIST1	DIST2	METEOR	ROUGE	CIDEr	ST	EV	VE	GM	BS-P	BS-R	BS-F1	BLEURT
1	0.35	0.31	0.20	0.07	0.25	0.18	0.39	1.56	0.78	0.89	0.60	0.95	0.73	0.73	0.73	0.22
2	**0.50**	0.44	0.34	0.06	0.16	**0.27**	**0.48**	**2.06**	**0.81**	**0.92**	0.61	**0.96**	0.77	**0.77**	**0.77**	**0.31**
3	0.49	0.43	0.33	0.06	0.17	0.26	0.47	1.94	0.80	0.90	0.60	0.95	0.77	0.76	0.76	0.27
4	0.49	**0.46**	**0.35**	0.06	0.16	0.26	0.47	1.97	0.81	0.91	0.61	0.95	**0.78**	0.75	0.76	0.29
5	0.43	0.38	0.29	0.05	0.18	0.20	0.46	2.04	0.81	0.91	0.58	0.95	0.77	0.74	0.75	0.29
6	0.42	0.37	0.28	0.07	0.19	0.20	0.37	1.27	0.78	0.88	0.50	0.94	0.72	0.70	0.71	0.17
7	0.38	0.33	0.24	0.07	0.22	0.17	0.41	1.72	0.79	0.90	0.56	0.95	0.74	0.72	0.73	0.23
8	0.49	0.44	0.34	0.06	0.17	0.25	0.46	1.80	0.80	0.90	0.60	0.95	0.78	0.74	0.76	0.26
9	0.34	0.27	0.19	0.05	0.12	0.16	0.27	0.66	0.75	0.89	0.34	0.91	0.67	0.70	0.68	0.11
10	0.40	0.28	0.19	0.07	0.25	0.17	0.40	1.33	0.79	0.88	0.50	0.94	0.76	0.74	0.75	0.26
11	0.01	0.00	0.00	0.52	0.63	0.15	0.01	0.00	0.64	0.77	0.77	0.77	0.54	0.62	0.57	−0.25
12	0.22	0.21	0.13	0.16	0.51	0.25	0.08	0.00	0.71	0.83	**0.83**	0.83	0.64	0.70	0.67	0.15
13	0.01	0.00	0.00	0.56	0.71	0.19	0.06	0.00	0.66	0.78	0.78	0.78	0.54	0.65	0.59	−0.28
14	0.00	0.00	0.00	**0.99**	**0.98**	0.19	0.00	0.00	0.72	0.83	0.83	0.83	0.61	0.73	0.66	−0.16

注: 加粗表示本列最优结果。

表 10　CCF 各模型人工评估结果

Table 10　Human evaluation results of each models for CCF

序号	Lic 2021知识对话										CCF LUGE知识对话								
	知识对话			对话推荐				画像聊天			闲聊		知识对话			对话推荐			
	Info.	Coh.	Know.	Info.	Coh.	Know.	Rec.	Info.	Coh.	Know.	Info.	Coh.	Info.	Coh.	Know.	Info.	Coh.	Know.	Rec.
1	**1.20**	**1.65**	1.41	0.44	1.13	0.66	1.50	0.93	0.83	1.17	0.47	0.47	0.79	1.01	0.90	0.43	1.14	0.70	1.07
2	1.02	1.47	1.22	0.49	1.32	0.80	**2.00**	1.12	1.11	1.26	0.60	0.57	0.89	1.22	1.11	0.47	1.22	0.84	1.43
3	0.88	1.27	1.94	0.48	1.25	0.75	1.47	1.10	1.05	1.23	0.30	0.30	**1.06**	**1.29**	1.28	0.46	1.18	0.74	1.47
4	1.11	1.57	1.31	0.43	1.09	0.64	1.63	0.96	1.02	1.20	0.37	0.30	0.65	0.66	0.71	0.48	1.26	0.78	1.50
5	1.08	1.46	1.21	0.43	1.26	0.71	1.00	1.18	1.09	1.27	0.57	0.40	0.90	1.11	0.98	0.41	1.16	0.66	0.60
6	1.07	1.38	1.24	0.43	0.85	0.40	0.83	0.91	0.89	1.18	0.50	0.47	0.84	0.91	0.86	0.30	0.86	0.46	0.33
7	1.00	1.23	0.99	0.43	1.00	0.53	0.43	0.87	0.79	1.18	0.27	0.27	0.93	1.04	1.01	0.40	1.05	0.61	0.37
8	0.61	0.78	0.44	0.43	1.19	0.72	1.70	0.69	0.70	1.05	0.50	0.43	0.33	0.45	0.36	0.51	1.33	0.89	**1.73**
9	1.09	1.15	1.18	0.43	1.03	0.53	1.17	0.79	0.79	1.08	0.27	0.07	0.01	0.49	0.01	0.01	0.83	0.03	0.13
10	0.32	0.78	0.14	0.43	0.80	0.25	0.57	0.72	0.68	0.92	0.20	0.20	0.48	1.04	0.46	0.22	0.89	0.36	0.33
11	0.40	0.77	0.47	0.83	1.07	0.73	0.60	0.37	1.07	0.27	0.33	0.30	0.40	0.77	0.47	0.57	0.73	0.50	0.47
12	0.50	0.53	0.67	**1.27**	1.43	**1.40**	1.57	**1.30**	1.50	**1.37**	0.33	0.40	0.53	0.53	0.67	0.80	1.70	0.83	1.57

续表 10

序号	Lic 2021知识对话										CCF LUGE知识对话									
	知识对话			对话推荐				画像聊天			闲聊		知识对话			对话推荐				
	Info.	Coh.	Know.	Info.	Coh.	Know.	Rec.	Info.	Coh.	Know.	Info.	Coh.	Info.	Coh.	Know.	Info.	Coh.	Know.	Rec.
13	1.00	1.13	**1.50**	1.17	1.40	1.10	1.27	0.70	1.00	0.63	0.33	0.30	1.00	1.13	**1.50**	**0.97**	1.33	**0.97**	1.13
14	0.80	0.53	0.60	0.33	**1.50**	0.47	1.00	0.93	**1.70**	0.90	**1.67**	**1.80**	0.80	0.53	0.60	0.70	**1.47**	0.60	0.97

注: 加粗表示本列最优结果。

表 11　知识对话中自动指标与人工判断的相关性

Table 11　Correlation between automatic indicators and human judgment in knowledge-grounded conversation

自动评估指标	Lic 2021知识对话						CCF LUGE知识对话					
	Info.		Coh.		Know.		Info.		Coh.		Know.	
	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE
F1	0.70	0.53	0.91	0.75	0.60	0.43	0.42	0.04	0.52	0.31	**0.38**	0.00
BLEU1	0.58	0.53	0.84	0.75	0.55	0.45	−0.03	−0.08	0.08	0.18	0.02	−0.11
BLEU2	0.57	0.62	0.82	0.81	0.54	0.54	0.04	−0.02	0.12	0.21	0.07	−0.06
DIST1	−0.32	−0.28	−0.44	−0.56	−0.16	−0.28	0.13	0.07	0.08	−0.23	0.11	0.02
DIST2	−0.46	−0.36	−0.53	−0.59	−0.22	−0.28	0.26	0.11	0.20	−0.16	0.21	0.12
METEOR	0.70	**0.79**	0.86	0.80	**0.61**	**0.65**	**0.50**	0.30	0.45	0.35	0.36	0.22
ROUGE	0.63	0.62	0.86	0.80	0.60	0.51	0.20	0.04	0.22	0.28	0.18	−0.01
CIDEr	0.59	0.68	0.81	**0.84**	0.53	0.63	0.11	−0.02	0.14	0.19	0.08	−0.07
ST	0.65	0.48	0.85	0.66	0.53	0.31	0.41	0.08	**0.57**	0.31	0.34	−0.03
EV	0.49	0.50	0.71	0.68	0.37	0.35	0.12	0.07	0.24	0.31	−0.02	0.00
VE	−0.28	−0.16	−0.31	−0.32	0.14	0.02	−0.12	−0.07	−0.21	−0.34	−0.18	−0.05
GM	0.49	0.49	0.72	0.66	0.23	0.32	0.18	0.09	0.27	0.32	0.13	0.01
BS-P	0.71	0.56	0.88	0.73	0.59	0.42	0.11	0.04	0.21	0.23	0.07	−0.04
BS-R	**0.75**	0.78	0.89	0.80	0.58	0.61	0.44	**0.31**	0.47	**0.39**	0.33	**0.24**
BS-F1	0.74	0.65	**0.92**	0.78	0.58	0.49	0.34	0.14	0.41	0.31	0.27	0.06
BLEURT	0.67	0.33	0.83	0.57	0.55	0.22	0.02	−0.16	0.20	0.09	0.02	−0.27

注: 加粗表示本列最优结果。

表 12　对话推荐中自动指标与人工判断的相关性

Table 12　Correlation between automatic indicators and human judgment in conversational recommendation

自动评估指标	Lic 2021对话推荐								CCF LUGE对话推荐							
	Info.		Coh.		Know.		Rec.		Info.		Coh.		Know.		Rec.	
	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE
F1	−0.10	−0.54	−0.27	−0.40	−0.05	−0.25	0.44	0.35	−0.34	−0.58	0.00	−0.11	0.30	0.02	0.32	0.18
BLEU1	−0.01	−0.50	−0.12	−0.31	0.08	−0.17	0.51	0.42	−0.29	−0.52	0.05	−0.04	0.32	0.09	0.38	0.25
BLEU2	−0.04	−0.54	−0.15	−0.30	0.05	−0.19	0.48	0.41	−0.29	−0.49	0.05	−0.05	0.32	0.12	0.38	0.27
DIST1	0.30	0.17	0.39	0.45	0.20	0.01	−0.26	−0.27	0.57	0.59	0.24	0.24	0.02	0.09	−0.01	−0.01
DIST2	0.31	0.45	0.45	0.51	0.28	0.24	−0.25	−0.24	0.51	0.71	0.22	0.35	−0.02	0.18	−0.07	0.04
METEOR	0.20	0.05	0.02	−0.07	0.26	0.34	**0.62**	**0.59**	0.30	0.26	0.62	0.61	**0.67**	**0.65**	**0.75**	**0.85**
ROUGE	−0.09	−0.64	−0.14	−0.44	0.03	−0.36	0.43	0.26	−0.32	−0.57	0.03	−0.19	0.33	0.05	0.34	0.12
CIDEr	−0.24	−0.67	−0.17	−0.34	−0.06	−0.34	0.21	0.27	−0.35	−0.48	−0.02	−0.14	0.27	0.15	0.26	0.17
ST	−0.32	−0.74	−0.11	−0.33	−0.07	−0.42	0.29	0.25	−0.37	−0.56	0.06	−0.02	0.22	0.01	0.25	0.14
EV	−0.17	−0.40	−0.15	−0.46	0.04	−0.16	0.37	0.28	−0.37	−0.63	0.04	−0.06	0.25	−0.07	0.25	0.10
VE	**0.54**	**0.85**	**0.73**	**0.64**	**0.65**	**0.80**	0.24	0.14	**0.96**	**0.93**	**0.72**	**0.63**	0.58	0.63	0.64	0.48

续表 12

自动评估指标	Lic 2021对话推荐								CCF LUGE对话推荐							
	Info.		Coh.		Know.		Rec.		Info.		Coh.		Know.		Rec.	
	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE
GM	−0.06	−0.49	−0.12	−0.54	0.06	−0.28	0.43	0.22	−0.37	−0.65	0.03	−0.16	0.29	−0.08	0.28	0.04
BS-P	−0.13	−0.69	−0.09	−0.39	0.03	−0.39	0.50	0.27	−0.31	−0.54	0.09	−0.03	0.27	0.05	0.38	0.19
BS-R	−0.34	−0.67	0.04	−0.02	−0.09	−0.27	0.39	0.40	−0.18	−0.33	0.25	0.28	0.28	0.16	0.40	0.37
BS-F1	−0.23	−0.69	−0.15	−0.29	−0.06	−0.35	0.50	0.33	−0.32	−0.48	0.10	0.07	0.31	0.10	0.37	0.25
BLEURT	−0.20	−0.57	−0.07	−0.32	−0.01	−0.25	0.36	0.30	−0.34	−0.54	0.06	0.02	0.23	0.02	0.30	0.18

注: 加粗表示本列最优结果。

表 13　聊天中自动指标与人工判断的相关性

Table 13　Correlation between automatic indicators and human judgment in chatting

自动评估指标	Lic 2021画像聊天						CCF LUGE闲聊			
	Info.		Coh.		Know.		Info.		Coh.	
	SP	PE	SP	PE	SP	PE	SP	PE	SP	PE
F1	0.56	0.48	−0.03	−0.50	0.67	0.78	−0.33	−0.26	−0.24	−0.27
BLEU1	0.55	0.50	−0.04	−0.48	0.68	0.78	−0.26	−0.27	−0.21	−0.29
BLEU2	0.50	0.46	−0.10	−0.48	0.64	0.73	−0.33	−0.13	−0.20	−0.11
DIST1	0.08	−0.45	0.54	0.53	−0.27	−0.78	0.06	0.54	0.17	0.56
DIST2	−0.03	−0.36	0.49	0.59	−0.35	−0.73	0.00	0.43	0.08	0.45
METEOR	**0.78**	0.30	0.87	**0.87**	0.45	0.00	−0.28	**0.81**	−0.07	**0.83**
ROUGE	0.46	0.31	−0.07	−0.54	0.55	0.62	−0.11	0.04	−0.04	0.03
CIDEr	0.37	0.39	−0.20	−0.44	0.53	0.62	−0.04	−0.04	−0.01	−0.02
ST	0.46	0.57	−0.14	−0.32	0.59	0.86	−0.02	0.11	0.09	0.12
EV	0.48	0.27	−0.13	−0.64	0.60	0.63	−0.22	−0.02	−0.12	−0.06
VE	0.56	0.41	**0.88**	0.76	0.24	−0.03	0.05	0.29	**0.29**	0.39
GM	0.37	0.25	−0.35	−0.65	0.55	0.62	−0.22	−0.08	−0.15	−0.14
BS-P	0.44	0.46	−0.20	−0.47	0.58	0.77	**0.08**	−0.22	0.11	−0.24
BS-R	0.62	**0.63**	0.08	−0.05	0.61	0.72	−0.03	0.13	0.12	0.16
BS-F1	0.48	0.51	−0.14	−0.39	0.62	0.80	−0.01	−0.14	0.04	−0.14
BLEURT	0.64	0.61	0.03	−0.26	**0.75**	**0.89**	−0.20	−0.38	−0.19	−0.41

注: 加粗表示本列最优结果。

在所有自动评估指标中, BERTscore 和 METEOR 表现出相对较好的性能, 即这些指标与人工评估的正向相关性很高, 最高可以达到 0.92, 几乎达到线性正相关, 因此这些指标可以一定程度上代替人工评估模型; 而 DIST1 和 DIST2 在几乎所有任务上表现不佳, 在某些任务上甚至与人工评估呈现负相关性, 最差可以达到−0.78, 接近线性负相关, 这说明 DIST1 和 DIST2 在这些任务上评估模型能力是不可靠的。

此外, 在不同的对话任务下, 自动评估与人工评估的相关性也有很大差异。具体来说, 在知识对话中, 大部分自动评估指标都与人工评估指标正相关, 尤其是 METEOR 和 BERTscore, 最高相关度能达到 0.92, 证明在知识对话任务中, 模型的能力比较容易评估; 但 DIST1 和 DIST2 表现得仍然很差, 与很多指标呈现负相关。

在对话推荐任务中, 大部分自动评估指标和人工评估指标呈负相关, 说明自动评估对话推荐任务相对困难。而在这之中 METEOR 和 VE 表现得很好, DIST1 和 DIST2 表现依然不好。

画像聊天任务的相关性结果再一次证明了 BERTscore 和 METEOR 表现较好, 而 DIST1 和 DIST2 表现较差这一现象, 说明不同指标之间的可靠性存在明显的差异。同时, 画像聊天中连贯性与很多自动指标的相关性较低, 证明在该任务上自动评估连贯性很难。

在闲聊任务中, 几乎所有的自动评估指标与人工评估的相关性都很低, 大多数呈现负相关,

最差甚至可以达到−0.41。这在一定程度上表明闲聊任务是相对较难评估的任务。其可能的原因是, 闲聊的连贯性相对较差, 相同的上下文有很多种合理的回复, 导致无论是人工评估还是自动评估都很难准确地评估生成的回复质量, 这极大地增加了评估的难度。

此外, 根据所有指标的平均相关性, 发现在人工评估指标中, 信息丰富度的评估在知识对话和推荐对话中相对较难, 例如在推荐对话中最低的相关性可以达到−0.74, 在知识对话中最低的相关性可以达到−0.46。闲聊中几乎所有的指标都很难评估。

5.3　实验结果总结

实验结果证明了在不同对话技能中自动指标的有效性差异很大, 例如在画像聊天中信息丰富度和自动指标的相关性平均可以达到 0.40, 而闲聊任务中平均仅可以达到−0.08。此外, 知识对话任务较为容易评估, 大部分自动指标都和人工指标呈正相关; 而闲聊任务是最难评估的, 各指标的平均相关性为−0.01, 这样会造成利用自动评估方法很难真正评估一个模型的对话能力。

因此, 未来应着重于从 3 个方面改进自动评估指标的能力: 1) 自动评估指标要与人工评估指标呈现正相关, 最好达到线性正相关。2) 自动评估指标应在不同对话技能 (不同领域) 上表现出较好的鲁棒性。3) 自动评估指标要能具有一定的可解释性。

6　结束语

针对已有基准只在特定对话技能评估自动指标与人工指标的相关性的问题, 本文构建了首个中文多技能对话评估基准 MSDE, 覆盖 4 类常见对话任务, 以促进多技能对话评估的研究。基于 MSDE, 本文做了大量实验, 并分析了常用的自动评估指标与人工评估指标之间的相关性, 同时探讨了不同对话技能的评估复杂性。结果表明, 自动评估指标中, BERTscore、METEOR 表现较好, 而 DIST1 和 DIST2 较差。在不同对话任务中, 闲聊任务的评估最具挑战性, 而知识对话等任务则相对容易评估, 这为未来的对话系统研究提供了参考。此外, 本文发现几乎所有指标在不同技能中评估的有效性差异都很大。因此, 亟须研究新的自动评估指标, 以便在多种技能中都与人工评估表现出较好的相关性, 同时在不同技能中表现出较好的鲁棒性。

参考文献:

[1] BAI Jinze, BAI Shuai, CHU Yunfei, et al. Qwen technical report[EB/OL]. (2023−09−28)[2024−11−01]. https://arxiv.org/pdf/2309.16609.

[2] YANG Aiyuan, XIAO Bin, WANG Bingning, et al. Baichuan 2: open large-scale language models[EB/OL]. (2023−09−19)[2024−11−01]. https://arxiv.org/abs/2309.10305.

[3] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[EB/OL]. (2023−07−18)[2024−11−01]. https://arxiv.org/abs/2307.09288.

[4] ZENG Aohan, XU Bin, WANG Bowen, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB/OL]. (2024−07−30)[2024−11−01]. https://arxiv.org/abs/2406.12793v2.

[5] ADIWARDANA D, LUONG M T, SO D R, et al. Towards a human-like open-domain chatbot[EB/OL]. (2020−02−27)[2024−11−01]. https://arxiv.org/abs/2001.09977.

[6] ROLLER S, DINAN E, GOYAL N, et al. Recipes for building an open-domain chatbot[EB/OL]. (2020−04−30)[2024−11−01]. https://arxiv.org/abs/2004.13637v2.

[7] SHUSTER K, JU Da, ROLLER S, et al. The dialogue dodecathlon: open-domain knowledge and image grounded conversational agents[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 2453–2470.

[8] 马中红, 吴熙倡. 社交聊天机器人的性别偏见: 基于小冰系列的对话测试研究[J]. 国际新闻界, 2024, 46(4): 72−89.
MA Zhonghong, WU Xichang. Gender bias in social chatbots: a conversation test study based on xiaoice series of chatbots[J]. Chinese journal of journalism & communication, 2024, 46(4): 72−89.

[9] 赵妍妍, 陆鑫, 赵伟翔, 等. 情感对话技术综述[J]. 软件学报, 2024, 35(3): 1377−1402.
ZHAO Yanyan, LU Xin, ZHAO Weixiang, et al. Survey on emotional dialogue techniques[J]. Journal of software, 2024, 35(3): 1377−1402.

[10] 房小绵. 基于语音识别的英语智能对话机器人人机交互系统设计[J]. 自动化与仪器仪表, 2023(4): 225−228, 232.
FANG Xiaomian. Design of human-computer interaction system for English intelligent conversation robot based on speech recognition[J]. Automation & instrumentation, 2023(4): 225−228, 232.

[11] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学: 信息科学, 2023, 53(9): 1645−1687.
CHE Wanxiang, DOU Zhicheng, FENG Yansong, et al. Towards a comprehensive understanding of the impact of

large language models on natural language processing: challenges, opportunities and future directions[J]. Scientia sinica (informationis), 2023, 53(9): 1645−1687.

[12] 王曦, 曾广平, 乔柱. 面向心理健康的服务机器人设计与实现[J]. 制造业自动化, 2021, 43(6): 137−141.
WANG Xi, ZENG Guangping, QIAO Zhu. Design and implementation of mental health oriented service robot[J]. Manufacturing automation, 2021, 43(6): 137−141.

[13] SMITH E M, WILLIAMSON M, SHUSTER K, et al. Can you put it all together: evaluating conversational agents' ability to blend skills[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 2021−2030.

[14] LIU Zeming, WANG Haifeng, NIU Zhengyu, et al. Towards conversational recommendation over multi-type dialogs[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 1036−1049.

[15] 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[EB/OL]. (2016−03−25)[2024−11−01]. https://arxiv.org/abs/1603.08023.

[16] YEH Y T, ESKENAZI M, MEHRI S. A comprehensive assessment of dialog evaluation metrics[EB/OL]. (2021−07−07)[2024−11−01]. https://arxiv.org/abs/2106.03706v4.

[17] SELLAM T, DAS D, PARIKH A. BLEURT: learning robust metrics for text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 7881−7892.

[18] 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. Philadelphia: ACL, 2001: 311−318.

[19] 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. Ann Arbor: Association for Computational Linguistics, 2005: 65−72.

[20] 刘阳阳, 董涛. 基于对话模型的聊天机器人结构研究[J]. 信息技术与信息化, 2023(1): 13−16.
LIU Yangyang, DONG Tao. Research on the structure of chat robot based on dialogue model[J]. Information technology and informatization, 2023(1): 13−16.

[21] LI Yanran, SU Hui, SHEN Xiaoyu, et al. DailyDialog: a manually labelled multi-turn dialogue dataset[EB/OL]. (2017−10−11)[2024−11−01]. https://arxiv.org/abs/1710.03957v1.

[22] GOPALAKRISHNAN K, HEDAYATNIA B, CHEN Qinlang, et al. Topical-chat: towards knowledge-grounded open-domain conversations[C]//Interspeech 2019. Graz: ISCA, 2019: 1891−1895.

[23] ZHANG Saizheng, DINAN E, URBANEK J, et al. Personalizing dialogue agents: I have a dog, do you have pets too?[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018: 2204−2213.

[24] DINAN E, LOGACHEVA V, MALYKH V, et al. The second conversational intelligence challenge (ConvAI2)[C]//The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations. Cham: Springer International Publishing, 2020: 187−208.

[25] 魏泽林, 张帅, 王建超. 基于知识图谱问答系统的技术实现[J]. 软件工程, 2021, 24(2): 38−44.
WEI Zelin, ZHANG Shuai, WANG Jianchao. Implementation of question answering based on knowledge graph[J]. Software engineering, 2021, 24(2): 38−44.

[26] 叶健辉, 韩博文, 周帆, 等. 基于自然语言处理的人机对话调控机器人设计[J]. 中国科技信息, 2020(22): 63−65.
YE Jianhui, HAN Bowen, ZHOU Fan, et al. Design of man-machine dialogue control robot based on natural language processing[J]. China science and technology information, 2020(22): 63−65.

[27] 张雨璇, 沙立成, 王海霞, 等. 电网调度智能对话机器人的系统架构和关键技术研究[J]. 电子设计工程, 2022, 30(11): 45−49.
ZHANG Yuxuan, SHA Licheng, WANG Haixia, et al. Research on system architecture and key technologies of intelligent conversation robot for power grid dispatching[J]. Electronic design engineering, 2022, 30(11): 45−49.

[28] LI Jiwei, 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. San Diego: ACL, 2016: 110−119.

[29] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 4566−4575.

[30] MEHRI S, ESKENAZI M. USR: an unsupervised and reference free evaluation metric for dialog generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 681−707.

[31] HUANG Lishan, YE Zheng, QIN Jinghui, et al. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems[EB/OL]. (2020−10−08)[2024−11−01]. https://arxiv.org/abs/2010.0399

4v1.

[32] PANG Bo, NIJKAMP E, HAN Wenjuan, et al. Towards holistic and automatic evaluation of open-domain dialogue generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 3619–3629.

[33] GHAZARIAN S, WEISCHEDEL R, GALSTYAN A, et al. Predictive engagement: an efficient metric for automatic evaluation of open-domain dialogue systems[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 7789−7796.

[34] HORI C, HORI T. End-to-end conversation modeling track in DSTC6[EB/OL]. (2018−01−30)[2024−11−01]. https://arxiv.org/abs/1706.07440v2.

[35] MEHRI S, ESKENAZI M. USR: an unsupervised and reference free evaluation metric for dialog generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 681–707.

[36] GUNASEKARA C, KIM S, D'HARO L F, et al. Overview of the ninth dialog system technology challenge: DSTC9[J]. IEEE/ACM transactions on audio, speech, and language processing, 2024, 32: 4066−4076.

[37] ZHENG Lianmin, CHIANG W L, SHENG Ying, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023: 46595−46623.

[38] 中国计算机学会, 中国中文信息学会, 百度. 2021 语言与智能技术竞赛: 多技能对话任务[EB/OL]. (2021−05−16)[2024−11−01]. https://aistudio.baidu.com/aistudio/competition/detail/67.

[39] 中国计算机学会. 千言: 多技能对话[EB/OL]. (2021−01−24)[2024−11−01]. https://www.datafountain.cn/competitions/470.

[40] WANG Yida, KE Pei, ZHENG Yinhe, et al. A large-scale Chinese short-text conversation dataset[EB/OL]. (2022−04−26)[2024−11−01]. https://arxiv.org/abs/2008.03946v2.

[41] WU Wenquan, GUO Zhen, ZHOU Xiangyang, et al. Proactive human-machine conversation with explicit conversation goal[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: ACL, 2019: 3794–3804.

[42] XU Xinchao, GOU Zhibin, WU Wenquan, et al. Long time No see! open-domain conversation with long-term persona memory[C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin: ACL, 2022: 2639–2650.

[43] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74–81.

[44] ZHANG Tianyi, KISHORE V, WU F, et al. BERTscore: evaluating text generation with BERT[C]//Proceedings of the International Conference on Learning Representations. New Orleans: OpenReview.net, 2019: 1−43.

[45] KIROS R, ZHU Yukun, SALAKHUTDINOV R, et al. Skip-thought vectors[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2015: 3294−3302.

[46] FORGUES G, PINEAU J, LARCHEVÊQUE J M, et al. Bootstrapping dialog systems with word embeddings[C]//Proceedings of NIPS Modern Machine Learning and Natural Language Processing Workshop. Montreal: Curran Associates Inc., 2014: 1−5.

作者简介：

柳泽明，助理教授，博士，中国中文信息学会大模型与生成专业委员会委员，中国中文信息学会具身智能专业委员会 (筹) 副秘书长和创始委员。主要研究方向为自然语言处理、对话系统、大模型、具身智能。主持国家自然科学基金、国家重点研发计划青年科学家项目任务、CCF-百度松果基金、多个校企科研合作项目等。获北航卓越青年学者、中国国际大学生创新大赛北京赛区“优秀创新创业导师”等。获发明专利授权 10 项，发表学术论文 40 余篇，包括第一作者和通信作者论文 20 余篇。E-mail：zmliu@buaa.edu.cn。

程子豪，主要研究方向为自然语言处理和工具学习。E-mail：zihaocheng@buaa.edu.c。

王蕴红，教授，北京航空航天大学计算机学院院长，中国人工智能学会智能交互专委会主任、中国人工智能学会常务理事、中国图象图形学学会常务理事，国际电气与电子工程师学会会士、国际模式识别协会会士、中国计算机学会会士、中国人工智能学会会士。先后主持国家高技术研究发展计划项目、国家重点基础研究发展计划项目、国家自然科学基金项目等项目。曾获得国家技术发明二等奖、中国青年科技奖、北京市教学成果一等奖，曾被科技部授予 863 计划先进个人，入选教育部新世纪优秀人才计划。获得国际模式识别学会女性科学家 Maria Petrou 奖，是该奖设立以来第一位获得此奖项的华人。获发明专利授权 30 余项，发表学术论文 200 余篇。E-mail：yhwang@buaa.edu.cn。