DOI: 10.16725/j.1673-808X.202381
引用本文: 王涵, 李龙. 基于多任务学习的对话伦理判别方法[J]. 桂林电子科技大学学报, 2026, 46(3): 284-294.
Wang Han, Li Long. A multi-task learning method for ethical dialogue detection[J]. Journal of Guilin University of Electronic Technology, 2026, 46(3): 284-294.

基于多任务学习的对话伦理判别方法

王　涵，李　龙

（桂林电子科技大学 计算机与信息安全学院, 广西 桂林　541004）

摘　要： 随着人工智能的发展, 相关的伦理问题和潜在风险愈发引起人们的关注。其中, 对话作为人们日常交流方式之一, 其与人工智能结合带来的伦理问题更加值得考虑。要应对这一问题, 首要任务是实现对话的伦理判别。基于此, 通过影视字幕、多轮对话数据集等多种来源方式, 构建了与伦理相关的中文多轮对话数据集 EthicDialog, 并对数据集的特点进行了分析。此外, 借助多任务学习实现了对话的伦理判别, 将对话中话语的伦理属性偏移检测任务和对话伦理判别任务结合, 从伦理标签变化和语义 2 个角度建模, 从多种角度捕捉上下文的伦理相关信息; 通过对 Transformer 层的掩码设计, 关注了对话的全局、局部及说话者相关信息, 从而更全面地考虑语义上下文、说话者信息等。实验结果表明了构造数据集的作用, 显示出所提出模型较基准具有着一定的效果提升。

关键词： 对话伦理判别; 多轮对话; 伦理对话数据集; 多任务学习; Transformer
中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1673-808X(2026)03-0284-11

A multi-task learning method for ethical dialogue detection

WANG Han，LI Long

（School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China）

Abstract: Ethical issues and potential risks associated with artificial intelligence have attracted increasing attention in recent years. As one of the most common forms of human communication, dialogue introduces important ethical concerns when combined with artificial intelligence systems. To address this issue, dialogue ethics detection has become an important research task. In this study, a Chinese multi-turn dialogue dataset, EthicDialog, is constructed from multiple sources, including film subtitles and existing dialogue corpora, and its statistical characteristics are systematically analyzed. Furthermore, a multi-task learning framework for dialogue ethics detection is proposed, which jointly performs ethical attribute shift detection and dialogue-level ethics classification. By introducing dedicated masking strategies into Transformer layers, the model effectively captures global contextual information, local semantic information, and speaker-related information, thereby enabling a more comprehensive understanding of semantic context. Experimental results demonstrate the effectiveness of the constructed Ethicdialog dataset and verify that the proposed model consistently outperforms baseline methods.

Key words: ethical dialogue detection; multi-turn dialogue; ethical dialogue dataset; multi-task learning; Transformer

目前, 人们越来越关注人工智能技术发展带来的伦理问题。国内外均有许多伦理道德规范准则被提出, 如《机器人伦理报告初步草案》《新一代人工智能伦理规范》等[1-2]。这些都体现出在使用人工智能技术的同时, 也需要考虑处理潜在伦理风险的必要性。

对话的主观性质导致聊天机器人的使用存在潜

收稿日期: 2023-06-29
基金项目: 广东省基础与应用基础研究基金(2023A1515012846); 广州市科技计划(202201011128); 认知无线电与信息处理省部共建教育部重点实验室基金(CRKL220208); 广西精密导航技术与应用重点实验室基金(DH202230)
通信作者: 李龙(1989—), 男, 教授, 博士, 研究方向为可信机器学习、人工智能安全、大数据隐私保护。E-mail: lilong@guet.edu.cn

在危险。微软的 TwitterBot Tay 于 2016 年发布, 但其发布种族主义、攻击性言论引发舆论强烈反弹后, 被官方紧急下架关停。到现在为止, 对话安全仍然是生成性对话模式的致命弱点[3-4]。毋庸置疑, 对话中伦理问题的存在不容忽视。要有效减少对话生成中出现的伦理问题, 首要任务是实现对话的伦理判别。

目前对话伦理方面的研究主要是对句子[5-8]、故事[9-10]等文本的伦理判定或生成。而短文本和长文本的伦理判别不同于对话。对话有着上下文话语, 且具有 2 个及以上说话者, 这使得对话的伦理判别不仅需要关注上下文, 而且不同的上下文之间也存在不同的关联影响。因此, 对话的伦理判别问题更复杂。Sun 等[4]对对话中的伦理问题做了考虑, 基于对话生成中存在的问题, 提出了对话安全的概念, 将回复分为对话安全和 6 个对话不安全类别; 并基于设定的分类法, 构建了相应的单轮对话数据集, 但其只针对英文单轮对话进行判别, 而多轮对话的伦理判别比单轮对话更复杂。

多轮对话的伦理判别与话语的上下文和说话者信息存在一定联系。在一组对话中, 上下文的伦理合规与否会影响后续话语的伦理属性, 并导致后续话语的伦理状态发生变化。此外, 说话者信息对话语的伦理属性也存在影响。一方面, 说话者本身的伦理属性存在一定的稳定性; 另一方面, 说话人之间的话语伦理属性也存在一定影响。目前, 深度学习技术广泛应用于图像[11-12]、文本[13-14]等领域, 大量多轮对话理解的相关研究均基于深度学习技术[15-17]。Shen 等[18]基于预训练模型实现的方式改进 XLNet, 提出了 DialogXL 模型, 针对当前对话与上下文间的不同关系, 通过改进 XLNet 的 Transformer 层, 实现了对多轮对话中整体、局部上下文、说话者本身和说话者之间的上下文语义信息的捕捉关注。在一组对话中, 话语间的伦理属性具有一定的相关性。一个人持有的伦理观念具有一定的稳定性, 因而同一说话者的话语具有类似的伦理属性; 此外, 在 2 个说话者交互对话过程中, 可能出现说话者的伦理态度冲突并相互影响的情况。

综上所述, 目前研究缺乏对多轮对话伦理判别的关注, 且对话理解中更多关注的是话语间的语义关系, 忽视了话语间的伦理属性变化影响。当前研究大都忽视从伦理标签变化角度建模, 或者只对伦理和语义进行建模, 存在对这两部分考虑不全面的问题。在中文对话伦理判别领域, 现有的研究忽视了话语的伦理属性偏移, 而在实际对话过程中, 话语的伦理属性会随语境变化而变化。

多任务学习旨在联合学习多个相关辅助任务, 使任务中包含的知识能被其他任务利用, 对相关任务的信息进行补充学习, 从而提高所有任务的泛化性能[19-20]。伴随着深度学习发展和海量数据的使用, 相比单任务学习模型, 深度多任务学习模型表现出更好的性能。目前许多研究通过多任务学习实现对话相关任务。Chen 等[21]通过 3 个文本理解的子任务, 实现对对话中的话语意图、领域等信息提取, 从而更好地实现对话理解。Zhang 等[22]通过一个基于胶囊神经网络的多任务学习对对话语义层次结构进行了研究。

基于以上背景, 构建了一个中文多轮对话的伦理数据集 EthicDialog, 依托该数据集引入多任务学习机制, 从伦理的偏移和语义 2 个方面对说话者和上下文进行考虑, 提出了一个分类模型 ESD-DialogXL, 实现了结合上下文、说话者信息的对话伦理判别。该模型基于 DialogXL 改进, 通过多任务学习, 将对话伦理判别作为主任务, 伦理属性的偏移检测作为辅助任务, 在原来 DialogXL 关注了语义方面的全局、局部上下文和说话者信息的情况下, 通过辅助任务中对同一说话者的话语伦理属性偏移情况的关注, 对伦理偏移方面的说话者信息进行编码。最终通过对 2 个任务的训练交互, 提升了对话伦理判别的效果。

1　EthicDialog 数据集

通过影视字幕网站的字幕文件、网络新闻等公开的聊天记录、公开对话数据集、人为构造对话等方式, 构造了一个中文多轮对话的伦理数据集 EthicDialog, 其中包含符合伦理道德(eth)、中性(neu)、不符合伦理道德(une)三个类别。

1.1　数据采集和标注

通过多种来源收集伦理相关的对话数据。主要通过以下 4 种方式收集数据: 利用影视字幕文件筛选出符合条件的对话数据; 公开数据集; 人工编写符合条件的对话; 社交平台等其他来源的公开聊天记录。

1.1.1　影视字幕文件

参考其他对话数据集的构造来源, 考虑伦理对话数据集的构造。目前的对话数据集, 一般通过社交平台、影视、模型生成等方式进行获取。然而, 社交平台内容繁杂杂乱, 有效伦理对话样本获取难度大, 难以筛选高质量标注数据。而模型生成无法保证获取的聊天数据与伦理相关。为了使对话的来源更加广泛和自然, 考虑在影视字幕中选取对话并构造。因此, 通过射手字幕网、OpenSubtitles 等字幕网站获取公开的字幕文件。为更容易找到体现伦理特点的对话,

选择犯罪题材和一些体现伦理道德题材的电影字幕，从中筛选满足条件的对话片段。

1.1.2 公开数据集

电影字幕中的对话存在口语化、体现隐晦的问题。虽然有些对话有与伦理相关的场景，但内容并不直接，很多话语是隐晦地表现出当时的情况。此外，收集到的数据主要为一组四五句的短对话，缺乏长对话。为了解决此问题，从现有的对话数据集 M3ED[23] 中筛选伦理相关的长对话数据。此外，通过检索“杀”“死”“犯罪”“救”“帮”等字词，加快对数据的筛选。

1.1.3 人工构造对话

从影视字幕文件、现有对话数据集中筛选合适数据，需耗费大量时间与人力，且所得对话往往不够自然。因此，进一步通过人工编写对话的方式获取数据。给出已经构造的典型对话数据示例，并基于帮助、道德、犯罪等主题要求，由多人编写双人对话，再将编写结果进行筛选，将符合要求的对话加入数据集。其中，对话需体现说话者存在帮助、守规、伤害等行为或意图，或者对相关行为有悔改、批判、认同等态度。

1.1.4 社交平台等其他来源的对话

为获取更多、更广泛、更自然的对话数据，从社交网络平台、剧本等搜集可用对话数据。通过在微博检索“偏见”“骂”“怼”等相关字眼，并关注一些热点争议新闻，查看是否有满足对话形式的评论交互以及公开的聊天记录。获取到相关聊天记录后，去除其中涉及个人隐私的部分，只提取逻辑连贯的有效文字片段。此外，还通过检索具有教育意义的寓言、剧本，添加少量符合伦理的对话交互。

从多个数据源获得满足条件的对话数据后，对数据进行清洗，再对格式做统一处理。数据的标签标注由 5 人进行，每组对话数据分给两人标注，当两人标注结果不一致时，由第 3 个人对数据再打标签，若存在较大争议，则舍弃该组对话数据。将对话中的每条话语标签分为 eth(符合伦理道德)、neu(中性，与伦理与否无关)、une(不符合伦理道德)3 类。所有数据通过 json 格式展示。每条话语包含话语、说话者、对应话语标签，如表 1 所示。

基于功利主义的伦理理论对数据进行了标注。功利主义关注整体的幸福感最大化[24]，其行为的效用值M计算式为[25]

$$M=\sum_{i=1}^{N}W_iP_i。$$

其中：N为利益相关者的数量；W_i为每个利益相关者的权重；P_i用于权衡每个利益相关者的幸福度。

表 1 数据集中 1 条对话话语的信息

内容类型	内容
text	只有把我们孩子卖掉了
speaker	A
label	une

1.2 数据统计分析

收集到的数据有 341 组对话，2 761 条话语，每组对话的平均长度为 7.82。所构建对话话语数据中，eth∶neu∶une 为 350∶1 666∶655。现实情况中，一组对话中的中性话语居多，数据集中的中性话语数据亦会更多。

为体现 3 种类别的话语在对话中的伦理属性变化情况，对话上下文的伦理变化情况进行统计，结果如图 1 所示，其中伦理属性突变为 eth、neu、une 的概率分别为 0.701、0.218、0.760。从图 1 可看出，符合伦理和不符合伦理的部分更易突变，这种突变主要源于中性话语。这主要由于选取的话语主要针对话语伦理属性会在上下文发生变化的情况，而这种变化更多是日常生活中本无任何伦理属性(中性)的话语所引起的。此外，还可看出，当前一句话为 eth 或 neu 时，其后续的话语更大概率仍保持这种属性。而当前一句话语为 une 时，其后续话语更倾向于某种伦理属性，而非中性。这是由于人们在对话过程中，对他人不符合伦理的话语做出的反应，更易表现出其当下反应是否符合伦理。

此外，对同一说话者的前后话语标签变化进行了统计，如图 2 所示，可看出，同一说话者的伦理属性突变为 eth、neu、une 的概率分别为 0.618、0.172、

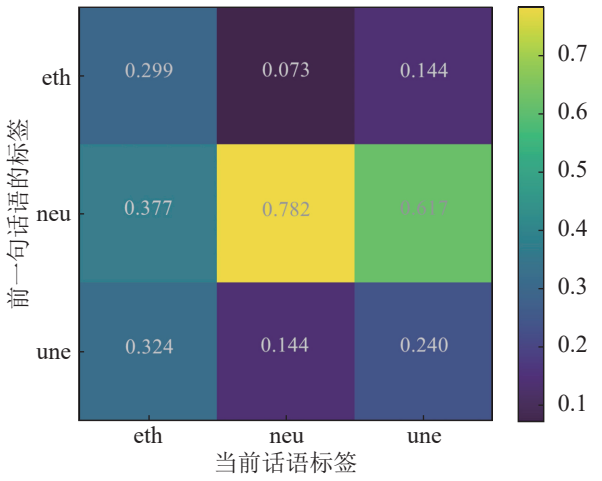


图 1 对话上下文变化概率矩阵

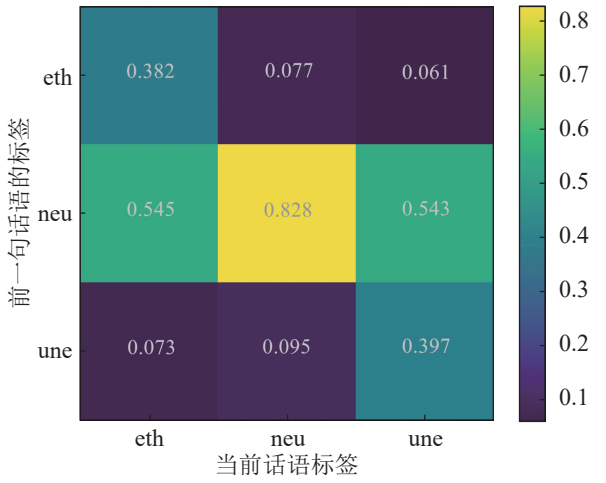


图 2　同一说话者的上下文变化概率矩阵

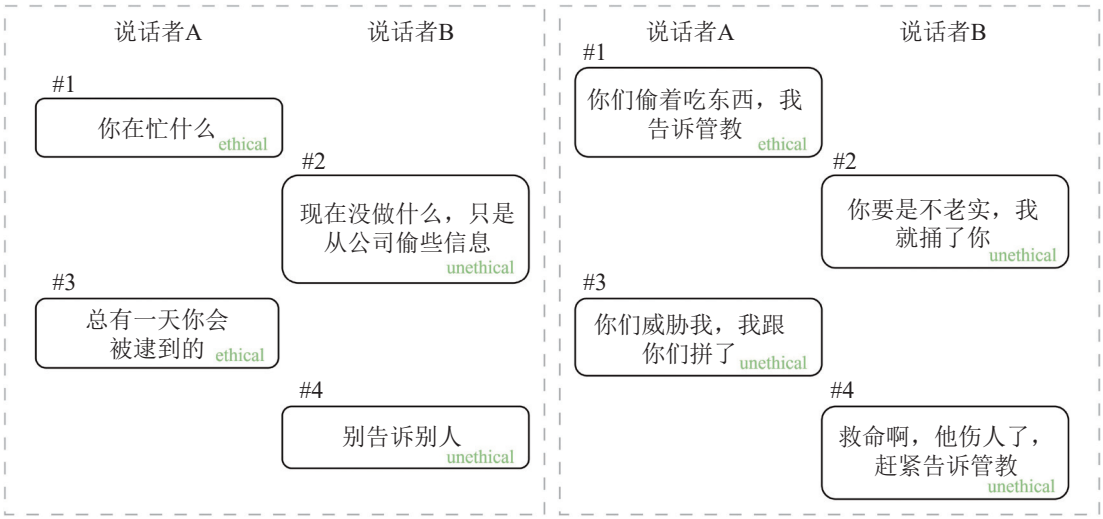
0.603, 与整体对话上下文的变化概率类似, 但均产生了一定幅度的减小。当同一说话者的前一句标签为 eth、neu、une 时, 其后一句话语标签保持不变的概率分别为 0.382、0.828 和 0.397; 相较于转变为其他类型的标签, 与当前话语说话者相同的后一句话语的标签, 更大概率保持原有的伦理属性, 而同一说话者的话语标签突变为 eth 和 une 的概率在分布上存在一些相似之处。这是因为说话者本身的伦理属性具有一定的稳定性和突变的规律。根据矩阵的标签变化概率分布可看出, 同一说话者的前一句为 eth、neu、une 时, 其后一句话语最大概率会保持原有的伦理属性。这体现出在 EthicDialog 数据集中, 说话者本身的伦理属性具有一定的稳定性。

2　ESD-DialogXL 模型

在一组对话中, 上下文的伦理与否, 会影响到后续话语, 并使得后续话语的伦理与否发生变化。图 3 为两组伦理相关对话示例。其中, 第一个对话示例中, 单看句子, 话语#4 与伦理并不相关, 是中性的; 但结合上下文, 可看出说话者 B 正在窃取公司信息, 并警告说话者 A, 表现出试图掩盖罪行的行为, 因此这句话语不符合伦理。

上下文对原本话语的伦理属性存在影响, 此外, 说话者信息对对话话语的伦理属性也存在影响。一方面, 说话者本身的伦理属性存在一定的稳定性, 如图 3(a)中, 说话者 B 的话语从始至终都不符合伦理; 而另一方面, 说话人之间的话语伦理属性也存在一定影响。第二组对话中, 说话者 A 第一句话语符合伦理, 但被说话者 B 不符合伦理的话语影响, 其回应也变得不符合伦理。

图 3 表明, 多轮对话的伦理判别与话语的上下文、说话者信息存在一定联系。因此, 为解决对话伦理判别的问题, 需关注话语上下文和说话者信息。



(a) 第一组对话示例　　(b) 第二组对话示例

图 3　伦理相关的对话示例

2.1　任务定义

2.1.1　对话伦理判别

根据 DialogXL 等对情感分析任务的定义, 可对对话伦理判别任务做如下定义:

给定一组对话$U=\{u_1,u_2,\cdots,u_N\}$和对应的说话者$\{s_1,s_2,\cdots,s_N\}$, 其中: N为一组对话的话语总数; s_i为话语u_i对应的说话者。对话伦理判别任务是, 给定一个话语u_t和它对应的说话者s_t、上下文$\{u_1,u_2,\cdots,u_{t-1}\}$, 输出这个话语对应的伦理标签$y_t$。

2.1.2　基于多任务学习下的对话伦理判别

参考情感偏移作为辅助任务的对话情感分析任

务[26]，可以定义伦理属性偏移作为辅助任务的对话伦理判别任务，定义如下：

给定一组对话d，该组对话可用一个四元组$d=(U_l,S_l,Y_l^{\mathrm{ET}},Y_l^{\mathrm{ES}})$表示。其中，话语序列$U_l=\{u_1,u_2,\cdots,u_N\}$，说话者序列$S_l=\{s_1,s_2,\cdots,s_N\}$。$Y_l^{\mathrm{ET}}=\{y_1^{\mathrm{ET}},y_2^{\mathrm{ET}},\cdots,y_N^{\mathrm{ET}}\}$为伦理判别的标签序列，$y_i^{\mathrm{ET}}\in S^{\mathrm{ET}}$，$S^{\mathrm{ET}}$为伦理标签的集合。$Y_l^{\mathrm{ES}}=\{y_1^{\mathrm{ES}},y_2^{\mathrm{ES}},\cdots,y_N^{\mathrm{ES}}\}$为伦理属性偏移的标签序列，$y_i^{\mathrm{ES}}\in S^{\mathrm{ES}}$，其中$S^{\mathrm{ES}}=\{0,1,[\mathrm{start}]\}$，为伦理相关标签偏移的标签集合，集合中元素分别表示无偏移、偏移、初始状态。初始状态表示说话人第一次说话时的话语伦理属性。最初状态由一个零向量表示。对于一个给定话语序列U_l和说话者序列S_l的对话d，主任务为输出伦理判别标签Y_l^{ET}，辅助任务为输出伦理属性偏移的标签序列Y_l^{ES}。

2.2 模型介绍

ESD-DialogXL 模型整体结构如图 4 所示。该模型基于 DialogXL 改进，通过多任务学习将语义上下文建模(对话伦理判别)作为主任务，伦理属性的偏移检测作为辅助任务，在原 DialogXL 关注了语义方面的全局、局部上下文和说话者信息的情况下，通过辅助任务中对同一说话者的话语伦理属性偏移情况的关注，对伦理偏移方面的说话者信息进行编码。

考虑到抛开语义上的伦理影响，话语在上下文中的伦理属性之间也可能存在一定关联。基于将伦理偏移检测作为辅助任务的多任务学习，提出了 ESD-DialogXL 模型。其结构如图 5 所示，首先将当前话语和与其同说话者的前一句话语送入 XLNet 预训练模型中，得到对应话语的语义表征。

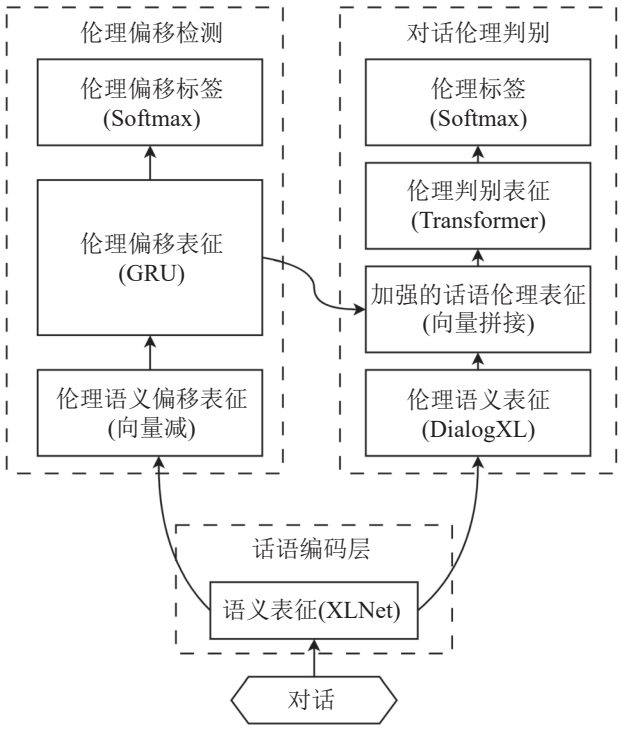


图 4 ESD-DialogXL 模型整体结构

对于辅助任务，将得到的 2 个表征向量相减，得到伦理语义偏移表征；将其作为输入，送入双向 GRU 编码器，得到伦理偏移表征，再通过 MLP 和 Softmax 函数，得到最终的伦理偏移标签序列。

而对于主任务，则将得到的话语语义表征先放入 DialogXL 的多层 Transformer 中，得到包含伦理全局、局部和说话者关系的上下文信息的伦理语义表征，再将伦理语义表征和伦理偏移表征拼接，得到加强的话语伦理表征，再放入 Transformer 中，对两部分信息进行训练学习，得到最终的伦理判别表征。通过 MLP 和 Softmax 函数，得到最终的伦理判别标签序列。

具体而言，可将模型结构分为伦理偏移检测任务部分和伦理判别任务 2 部分。其中，伦理偏移检测任务部分的模型结构分为文本编码输入层、伦理偏移编码层和伦理偏移输出层；而伦理判别任务部分的模型结构分为文本编码输入层、伦理语义编码层、融合伦理偏移表征信息的伦理判别编码层和伦理判别输出层。

2.2.1 文本编码输入层

文本编码输入层在伦理偏移检测任务和伦理判别任务中存在共享变量的情况，且部分处理流程相似。其中，伦理偏移检测任务的输入话语和伦理判别任务中的输入话语的关系为：前者的说话者和后者的说话者相同，而前者为后者的前一句话。

2 个任务中，本层的通用处理流程为：首先，对于每个时间步t，在输入的话语前加入“[CLS]”，得到的输入话语u_t、u_t^{ES}分别为：

$$u_t=\{[\mathrm{CLS}],w_{t1},w_{t2},\cdots,w_{tn_t}\},$$

$$u_t^{\mathrm{ES}}=\{[\mathrm{CLS}],w_{t1}^{\mathrm{ES}},w_{t1}^{\mathrm{ES}},\cdots,w_{tn_t^{\mathrm{ES}}}^{\mathrm{ES}}\}。$$

其中：u_t和u_t^{ES}分别为伦理判别任务和伦理偏移检测任务的输入话语；n_t和n_t^{ES}分别表示u_t和u_t^{ES}中包含的 token 数。

将话语u_t和u_t^{ES}分别传入嵌入层，得到各个话语对应的词向量级表示。使用 XLNet 预训练模型，得到话语的句子表示。得到的语义表征表达式为：

$$\boldsymbol{h}_t=f_{\mathrm{e}}(u_t),$$

$$\boldsymbol{h}_t^{\mathrm{ES}}=f_{\mathrm{e}}(u_t^{\mathrm{ES}})。$$

其中：$\boldsymbol{h}_t$为伦理语义表征，作为伦理判别任务中的输入，后续被送入 DialogXL 的多层 Transformer 中；$\boldsymbol{h}_t$、$\boldsymbol{h}_t^{\mathrm{ES}}\in\mathbf{R}^{d_{\mathrm{input}}}$；$d_{\mathrm{input}}$为 XLNet 模型的维度；$f_{\mathrm{e}}$为获取词嵌入的函

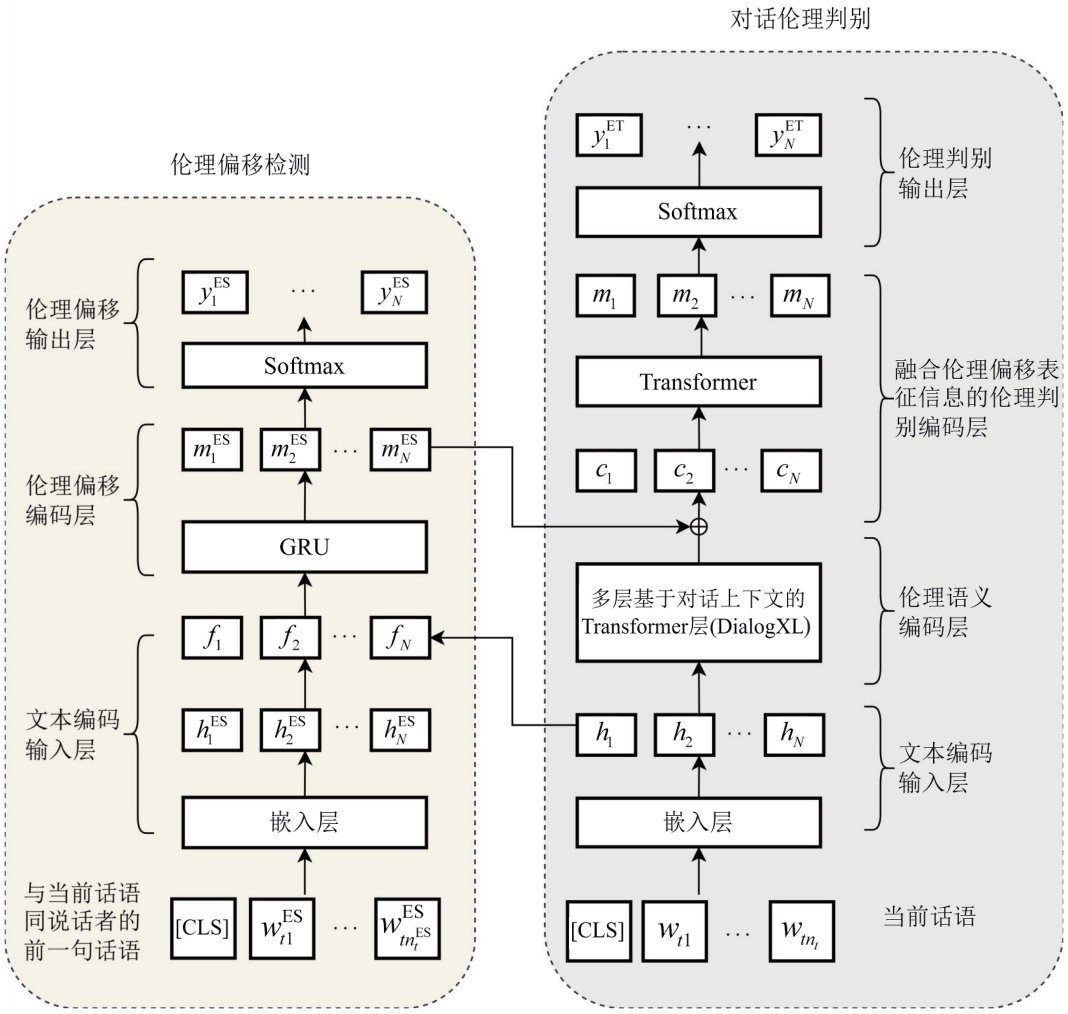


图 5　ESD-DialogXL 模型结构

数, 为 XLNet 模型。

由于部分u_t为说话者s_t所说的第一句话无对应的前一句话, u_t^{ES}为空值, 通过一个零向量$\boldsymbol{h}_{\mathrm{start}}$来表示$u_t^{\mathrm{ES}}$对应的$\boldsymbol{h}_t^{\mathrm{ES}}$。

要得到伦理语义偏移表征, 还需将前面得到的句子表示进行计算处理。根据句子相似度分类任务[27-29]中的 Sentence-BERT 等模型思想, 采用向量减的方式获得每条话语的伦理语义偏移表征。具体地, 通过同一说话者前后话语的语义表征相减, 得到最终的语义偏移表征。

伦理语义偏移表征的计算式为

$$\boldsymbol{f}_t = \boldsymbol{h}_t^{\mathrm{ES}} - \boldsymbol{h}_t, s_t = s_t^{\mathrm{ES}}。$$

其中: 话语u_t和话语u_t^{ES}分别为同一说话者的当前话语和前一句话; $\boldsymbol{h}_t$和$\boldsymbol{h}_t^{\mathrm{ES}}$分别为话语$u_t$和$u_t^{\mathrm{ES}}$的句子语义表示; s_t和s_t^{ES}分别为话语u_t和u_t^{ES}的说话者。

2.2.2　伦理偏移检测任务

1)伦理偏移编码层。

对话文本存在连贯的上下文, 为模拟对话上下文的伦理偏移过程, 将得到的伦理语义偏移表征$\boldsymbol{f}_t$放入 BiGRU 编码器中, 得到每条话语的伦理偏移表征。BiGRU 是由 2 个单向的、方向相反的 GRU 组成的模型, 其利用 2 个方向相反的 GRU 来处理正向和反应序列, 获得 2 个方向的特征信息。在每个时刻, 输入会同时提供 2 个方向相反的 GRU, 而输出则是 2 个 GRU 输出的组合, 具体流程为:

$$\overrightarrow{\boldsymbol{m}_t^{\mathrm{ES}}} = \overrightarrow{\mathrm{GRU}}\left(\overrightarrow{\boldsymbol{m}_{t-1}^{\mathrm{ES}}}, \boldsymbol{f}_t\right),$$

$$\overleftarrow{\boldsymbol{m}_t^{\mathrm{ES}}} = \overleftarrow{\mathrm{GRU}}\left(\overleftarrow{\boldsymbol{m}_{t+1}^{\mathrm{ES}}}, \boldsymbol{f}_t\right),$$

$$\boldsymbol{m}_t^{\mathrm{ES}} = \left[\overrightarrow{\boldsymbol{m}_t^{\mathrm{ES}}} \oplus \overleftarrow{\boldsymbol{m}_t^{\mathrm{ES}}}\right]。$$

其中: $\overrightarrow{\boldsymbol{m}_t^{\mathrm{ES}}}$和$\overleftarrow{\boldsymbol{m}_t^{\mathrm{ES}}}$分别表示$t$时刻输入的话语的前向和后向的 GRU 特征, $\oplus$为拼接操作, $\boldsymbol{m}_t^{\mathrm{ES}} \in \mathbf{R}^{2d_{\mathrm{GRU}}}$; d_{GRU}为 GRU 的隐藏层状态总数。

2)伦理偏移输出层。

伦理偏移输出层包含一个多层感知机 MLP 和 Softmax 函数。首先将$\boldsymbol{m}_t^{\mathrm{ES}}$送入 MLP 中得到每个类别

的概率表示，再通过 Softmax 对表征进行归一化，得到最终话语的伦理属性偏移概率序列$\hat{p}_t^{\mathrm{ES}}$。$\hat{p}_t^{\mathrm{ES}}$的每维代表话语u_t属于伦理偏移的每种标签的对应概率，

$$\hat{p}_t^{\mathrm{ES}} = \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{m}_t^{\mathrm{ES}}))。$$

2.2.3 伦理判别任务

1）伦理语义编码层。

伦理语义编码层使用 DialogXL 作为伦理的上下文语义编码器，其由多层改进的 Transformer 层组成，该 Transformer 层设计四类专属自注意力机制，实现了对全局、局部、说话者自身和说话者之间的上下文语义信息的关注，并具有可存储话语历史的缓存机制。将话语语义表征$\boldsymbol{h}_t$作为本层的输入，最终得到带有全局、局部、说话者相关的上下文信息的伦理语义表征$\boldsymbol{o}_t$。

$$\boldsymbol{o}_t = \mathrm{DialogXL}(\boldsymbol{h}_t)，$$

其中，$\boldsymbol{o}_t \in \mathbf{R}^{d_{\mathrm{input}}}$。对各部分的注意力掩码设计如下。

全局自注意力：关注所有历史上下文信息，与普通的自注意力并无区别，关注整体上下文。其掩码值为$s_{ij}^{\mathrm{g}} = 0$。

局部自注意力：只关注w个接近的上下文话语，其中w为超参数。由于当前话语的伦理表达主要被更近的话语交互所影响，对于局部自注意力，需将较远的历史话语的注意力遮盖。对应掩码值表示为

$$s_{ij}^{\mathrm{lo}} = \begin{cases} +\infty, j \notin I(\{u_{t-w}, u_{t-w+1}, \cdots, u_t\}), \\ 0, \text{其他情况}。 \end{cases}$$

其中，$I(\cdot)$为将话语中的 token 映射到对应的注意力中$\boldsymbol{K}$矩阵的函数。

说话者自身的自注意力：由于说话者自身的稳定性，其具有相对统一的伦理属性，其所说话语的伦理性质存在一定的稳定性，因此需要关注说话者自身的上下文话语。这部分只考虑当前说话者所说话语的历史上下文，不关注非当前说话者所说的话语。

$$s_{ij}^{\mathrm{sp}} = \begin{cases} +\infty, j \in I(\{k | s_k \neq s_t\}), \\ 0, \text{其他情况}。 \end{cases}$$

说话者之间的自注意力：这部分自注意力只考虑其他说话者的话语。由于说话者会被其他说话者影响，而产生不同的与伦理相关的回应，故需要对说话者之间的话语进行关注。对应的注意力掩码为

$$s_{\mathrm{i}j}^{Ii} = \begin{cases} +\infty, j \in I(\{k | s_k = s_t\}), \\ 0, \text{其他情况}。 \end{cases}$$

2）融合伦理偏移表征信息的伦理判别编码层。

为了更好地实现对话伦理判别，从语义和伦理趋势变化 2 个角度实现对话的伦理判别，这部分通过将伦理偏移检测任务得到的伦理偏移表征$\boldsymbol{m}_t^{\mathrm{ES}}$和前面伦理语义编码层得到的伦理语义表征$\boldsymbol{o}_t$进行拼接，得到拼接伦理表征$\boldsymbol{c}_t$，再送入 Transformer 编码器中，得到最终的伦理判别表征$\boldsymbol{m}_t$。此过程可表示为

$$\boldsymbol{c}_t = [\boldsymbol{m}_t^{\mathrm{ES}} \oplus \boldsymbol{o}_t]，$$

$$\boldsymbol{m}_t = \mathrm{TransformerEncoder}(\boldsymbol{c}_t)。$$

其中，$\boldsymbol{c}_t$、$\boldsymbol{m}_t \in \mathbf{R}^{2d_{\mathrm{GRU}}+d_{\mathrm{input}}}$。

3）伦理判别输出层。

这部分与伦理偏移任务的伦理偏移输出层类似，包含一个多层感知机 MLP 和 Softmax 函数。首先将$\boldsymbol{m}_t$送入 MLP 中，得到每个类别的概率表示，再通过 Softmax 对表征进行归一化，得到最终话语的伦理属性偏移概率序列$\hat{p}_t^{\mathrm{ET}}$，$\hat{p}_t^{\mathrm{ET}} = \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{m}_t))$。

2.2.4 损失函数

用交叉熵损失作为主任务和辅助任务的损失函数。对于对话伦理判别任务，损失函数为

$$L^{\mathrm{ET}} = -\frac{1}{\sum_{1}^{N} T_l} \sum_{l=1}^{N} \sum_{i=1}^{\mathrm{T}_l} [y_i^{\mathrm{ET}}] \log(\hat{p}_i^{ET})，$$

其中：N为数据集中对话的组数；T_l为第l组对话中话语总数。

$$L^{\mathrm{ES}} = -\frac{1}{\sum_{1}^{N} T_l} \sum_{l=1}^{N} \sum_{i=1}^{T_l} [y_i^{\mathrm{ES}}] \log(\hat{p}_i^{\mathrm{ES}})。$$

最终 2 个任务的多任务学习的总损失为

$$L = L^{\mathrm{ET}} + L^{\mathrm{ES}}。$$

3 实验设计与结果分析

实验分为两部分，一部分是将 EthicDialog 应用于其他模型中，查看其结果，以验证数据集的效果；另一部分是将改进的模型应用于多个不同数据集中，查看结果，以体现本模型的改进效果。

3.1 度量标准

EthicDialog 的分类任务属于多分类，为计算模型在不同类别上的准确率和 F1 值，采用准确率A、宏 F1 值$F_{1,\mathrm{m}}$和加权 F1 值$F_{1,\mathrm{w}}$作为模型预测的评估指标，

$$A = \frac{T}{N_{\mathrm{a}}}，$$

$$P_{\mathrm{m}} = \frac{1}{K} \sum_{i=1}^{K} P_i，$$

$$R_{\mathrm{m}} = \frac{1}{K} \sum_{i=1}^{K} R_i，$$

$$F_{1,\mathrm{m}}=\frac{1}{K}\sum_{i=1}^{K}\frac{2P_{\mathrm{m}}R_{\mathrm{m}}}{P_{\mathrm{m}}+R_{\mathrm{m}}},$$

$$P_{\mathrm{w}}=\frac{1}{K}\sum_{i=1}^{K}w_iP_i,$$

$$R_{\mathrm{w}}=\frac{1}{K}\sum_{i=1}^{K}w_iR_i,$$

$$F_{1,\mathrm{w}}=\frac{1}{K}\sum_{i=1}^{K}\frac{2P_{\mathrm{w}}R_{\mathrm{w}}}{P_{\mathrm{w}}+R_{\mathrm{w}}}。$$

其中: N_{a}为所有类别样本的总数; T为所有样本中预测正确的总数; K为类别数; P_i为精确率; R_i为召回率; w_i为第i类标签在总样本类别中的占比。

3.2　实验设计

3.2.1　同一模型下，不同数据集的表现情况

为对比表现 EthicDialog 数据集的质量效果, 用 DialogXL 模型对 4 个常用对话数据集进行实验, 以验证其效果。

DailyDialog[30]: 该数据集为一个多轮情感对话数据集, 由人工编写, 包含 13 000 多轮对话, 共 102 000 条话语, 并由人工标注了 7 种情绪和 4 种对话行为。由于其无说话者信息, 这部分参考 DialogXL 中的处理。

MELD[31]: 该数据集是一个多模态的情感对话数据集, 基于美剧《老友记》构建而成, 包含 1 433 组多轮对话, 共 13 708 条话语, 其标签分为 7 类情绪。

M^{3}ED: 该数据集是一个中文的多模态情感对话数据集。其通过 56 个国内影视剧获取情感对话数据, 包含 990 组对话, 24 449 条话语, 其标签分为 7 类情绪。

CPED[32]: 该数据集是一个大型中文多模态对话数据集。其通过 40 多个影视节目构建数据, 对数据集中每句话的说话者性别、大五人格、情绪、对话行为、场景进行了标注, 共有 12 000 组对话, 133 000 条话语。其对情感部分的分类分为 3 类情感、13 类情绪, 本研究选择了 3 类情感作为标签。DailyDialog 和 MELD 为英文数据集, CPED、M^{3}ED 为中文数据集。仅选择各数据集的文本部分进行输入和训练。

关于各数据集的划分方式, DailyDialog 和 MELD 数据集参考 Shen 等[14] 的划分情况进行划分。而 CPED 使用其原作者的划分方式, 按照训练集：验证集：测试集为 7：1：2 的比例进行划分。由于 M^{3}ED 并未划分为不同数据集, 将数据集打乱后, 根据对话组数划分, 按照训练集：验证集：测试集为 6：2：2 的比例进行划分。这些数据集的训练集、验证集和测试集的数据分布情况如表 2 所示。

表 2　不同数据集的划分情况

数据集	语言	类别	对话数（训练/验证/测试）	话语数（训练/验证/测试）
DailyDialog	英文	7 类	11 118/1 000/1 000	87 170/8 069/7 740
MELD	英文	7 类	1 038/114/280	9 989/1 109/2 610
M^{3}ED	中文	7 类	592/199/199	14 748/4 808/4 893
CPED	中文	3 类	8 086/934/2 815	94 187/11 137/27 438
EthicDialog	中文	3 类	203/69/69	1 652/503/516

3.2.2　伦理对话数据集下，不同模型的表现情况

为提供一个有效的基准, 选取了两类基准模型: 仅将当前一句话作为输入的对话分类模型和将当前话语和对话上下文均作为输入的对话分类模型。

1)仅将当前一句话作为输入的分类模型, 包括 TextCNN[33]、TextRNN[13]、BERT[34]。其使用句子级的语言模型 LM 来获取话语的表征后, 再通过 MLP 来预测当前话语的伦理属性, 表达式为

$$\boldsymbol{h}=F_{\mathrm{l}}(u_t),$$

$$y_t=\mathrm{MLP}(\boldsymbol{h})。$$

其中 F_{l} 为不同的语言模型。将 EthicDialog 中的每组的一轮轮话语进行拆分, 作为模型的输入。

2) 将当前话语和对话上下文均作为输入的分类模型, 包括 bcLSTM[35]、DialogueRNN[36]、DialogueGCN[37]、DialogXL[18]。

bcLSTM 是一个双向上下文 LSTM 模型, 由 2 个单向且反向的 LSTM 堆叠而成。

DialogueRNN 是用 3 个 GRU 分别对对话的全局信息、说话者信息和伦理信息进行建模。

DialogueGCN 由顺序上下文编码器、说话者级编码器和伦理分类器 3 部分组成, 其中的说话者编码器通过图神经网络对不同位置的说话者内容进行编码。

DialogXL 是对 XLNet 预训练模型做出改动, 一方面加入了基于对话的自注意力, 另一方面对其中的缓存机制做改进, 节省了内存。

3.3　实验结果分析

3.3.1　同一模型下，不同数据集的表现情况

不同数据集在 DialogXL 下的分类结果如表 3 所示。

通过对比不同数据集在 DialogXL 下的各指标结果, 发现 EthicDialog 数据集在各个评估指标上均较高。原因在于: 一方面, 除 CPED 外, 其他数据集中的类别较多, 使得训练的难度变大; 另一方面, EthicDialog 中存在一些特定的与伦理相关的场景, 如悔改、对不符

表 3 不同数据集在 DialogXL 模型下的分类结果 %

数据集	准确率	宏 F1 值	加权 F1 值
DailyDialog	—	27.03	51.60
MELD	66.28	46.27	62.41
M^3ED	49.87	47.54	35.50
CPED	51.50	48.02	—
EthicDialog	79.26	67.67	76.82

合伦理行为的劝解等,这使得与伦理相关的发言前后也更倾向于与伦理相关,这也如图 1、2 对数据集上下文标签变化所示情况;此外,一组对话中也存在一定的上下文内容为与伦理无关的中性交流,这会使对中性标签的分类效果较好,且这部分也是对话交流场景下的多数话语,从而对整体结果产生一定影响。

3.3.2 伦理对话数据集下,不同模型的表现情况

将学习率设为1×10^{-6},将 dropout 设为 0.7,batch_size 设为 8,全局、局部、说话者上下文信息的注意力权重一致,以此为参数设定,得出 ESD-DialogXL 的结果,其结果和其他各个对比模型结果如表 4 所示。

表 4 EthicDialog 数据集在不同模型下的结果 %

模型	准确率	宏 F1 值	加权 F1 值
TextCNN	61.52	25.39	46.87
TextRNN	65.18	44.61	61.52
BERT	72.87	61.13	71.37
bcLSTM	69.38	51.17	66.67
DialogueRNN	68.99	46.59	64.48
DialogueGCN	69.38	46.78	65.51
DialogXL	79.26	67.67	76.82
ESD-DialogXL	80.72	68.12	77.88

在仅将当前一句话作为输入的 ERC 模型中,BERT 在所有指标中表现最好,甚至高过一些结合了上下文和说话者信息的对话。而 TextCNN、TextRNN 的宏 F1 值相较加权 F1 值过低。通过对分类结果进行分析,发现其原因为:在 eth 和 une 标签的学习效果不佳,这使得将各标签分类结果平均后的结果不够好。而其效果不佳的原因为:对话中存在上下文、说话者信息,这都会对话语的符合伦理与否产生影响;其中对话中存在一定量的中性话语上下文,这使得这部分标签的学习效果相对略好。此外,EthicDialog 中的中性话语较多,这种样本不均衡的问题也影响了各个模型的结果。

而对于将当前话语和对话上下文均作为输入的 ERC 模型中,DialogXL 模型的结果最好,一方面是其基于 XLNet 预训练模型改进,相较其他几个模型使用 word2vec 作为词嵌入的部分,其对词的表征具有更多的信息;另一方面,DialogXL 在语义方面,对全局、局部和说话者相关的信息均进行了关注,其关注更多层面的信息,也使其训练效果更为突出。

此外,对比 EthicDialog 数据集在各个模型下的结果可看出,ESD-DialogXL 在准确率、macro-F1 和 weighted-F1 下均获得了最佳结果。将其与未加入多任务学习的 DialogXL 比较,可发现其结果在 3 个指标下均有了一定的提升。这既体现出在话语的语义和伦理变化 3 个角度进行建模是具有一定效果的,又证明了本模型在对话伦理判别任务下的有效性。

4 结束语

对多轮对话的伦理判别进行研究,并根据多轮对话中上下文和说话者之前的影响和伦理属性自身在对话中的变化情况,对相关部分加以关注建模,提出了基于中文多轮对话的对话伦理判别方法。一方面,通过影视字幕等多种方式构建了关于伦理的日常对话数据集,对其分析可以发现说话者上下文等方面具有一定的规律,这值得探索学习;另一方面,通过多任务学习,将对话中话语的伦理属性偏移检测和对话伦理判别任务结合,从伦理自身变化和语义上下文变化的 2 个角度对对话的伦理判别进行建模,更好地实现了对话伦理判别;而根据对话的特点,在语义部分对上下文的全局、不同局部上下文各个部分进行关注,也使得其获取了不同的信息,促成了效果的提升。

然而本研究也存在一些不足之处:1)由于现有资源难以获取,且中文多轮对话领域的发展还有很大空间,目前数据集的规模有待扩大;2)存在数据不均衡的问题,这是因为在很多伦理相关的对话场景中,存在一些局部上下文的话语与伦理无关,但却对整体对话上下文存在影响;3)数据集的构建部分来源广泛,日常对话中的用语灵活,也使得数据集中数据不够规范,存在一定的噪声数据。在后续的研究工作中,将补充更多数据,根据数据集在说话者上下文部分的规律特点,更好地对其进行建模。

参考文献:

[1] 腾讯研究院. 人工智能各国战略解读: 联合国人工智能政策报告[J]. 电信网技术, 2017(2): 26-28.

[2] 张兆翔, 张吉豫, 谭铁牛. 人工智能伦理问题的现状分析与对策[J]. 中国科学院院刊, 2021, 36(11): 1270-1277.

[3] Henderson P, Sinha K, Angelard-Gontier N, et al. Ethi-

cal challenges in data-driven dialogue systems[C]//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM Press, 2018: 123-129.

[4] Sun Hao, Xu Guangxuan, Deng Jiawen, et al. On the safety of conversational models: taxonomy, dataset, and benchmark[C]//Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2022: 3906-3923.

[5] Hendrycks D, Burns C, Basart S, et al. Aligning AI with shared human values[C]//International Conference on Learning Representations. Ithaca, NY: OpenReview. net, 2021: 1-29.

[6] Forbes M, Hwang J D, Shwartz V, et al. Social chemistry 101: learning to reason about social and moral norms[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2020: 653-670.

[7] 古天龙, 冯旋, 李龙等. 基于社会新闻数据集的伦理行为判别方法[J]. 计算机研究与发展, 2021, 58(2): 253-263.

[8] Feng Xuan, Gu Tianlong, Bao Xuguang, et al. Behavior-based ethicalunderstanding in Chinese social news[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 2349-2360.

[9] Nahian M S A, Frazier S, Riedl M, et al. Learning norms from stories: a prior for value aligned agents[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM Press, 2020: 124-130.

[10] Emelin D, Le Bras R, Hwang J D, et al. Moral stories: situated reasoning about norms, intents, actions, and their consequences[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2021: 698-718.

[11] 欧阳宁, 任天宇, 林乐平. 基于梯度指导的双通道深度压缩感知图像重建方法[J]. 桂林电子科技大学学报, 2023, 43(1): 1-6.

[12] 黄新, 沈英超. 基于深度学习的疲劳驾驶检测方法[J]. 桂林电子科技大学学报, 2020, 40(3): 201-206.

[13] Liu Pengfei, Qiu Xipeng, Huang Xuanjing. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 2873-2879.

[14] 文益民, 杨鹏, 文博奚, 等. 基于深度学习的中文网络招聘文本中的技能词抽取方法[J]. 桂林电子科技大学学报, 2020, 40(4): 338-348.

[15] Chen Hongshen, Liu Xiaorui, Yin Dawei, et al. A survey on dialogue systems: recent advances and new frontiers[J]. ACM SigKDD Explorations Newsletter, 2017, 19(2): 25-35.

[16] Ghosal D, Majumder N, Mihalcea R, et al. Utterance-level dialogue understanding: an empirical study[PP/OL]. arXiv (2020-09-29)[2023-03-22]. https://arxiv.org/abs/2009.13902v5.

[17] Ni Jinjie, Young T, Pandelea V, et al. Recent advances in deep learning based dialogue systems: a systematic survey[J]. Artificial Intelligence Review, 2023, 56(4): 3055-3155.

[18] Shen Weizhou, Chen Junqing, Quan Xiaojun, et al. DialogXL: all-in-one XLNet for multi-party conversation emotion recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13789-13797.

[19] Zhang Yu, Yang Qiang. A survey on multi-task learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(12): 5586-5609.

[20] Vandenhende S, Georgoulis S, Van GANSBEKE W, et al. Multi-task learning for dense prediction tasks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1.

[21] Chen Y N, Hakkani-Tür D, Tur G, et al. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding[C]//The 17th Annual Meeting of the International Speech Communication Association. San Francisco, CA: ISCA, 2016: 3245-3249.

[22] Zhang Chenwei, Li Yaliang, Du Nan, et al. Joint slot filling and intent detection via capsule neural networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2019: 5259-5267.

[23] Zhao Jingming, Zhang Tengan, Hu Jingwen, et al. M3ED: multi-modal multi-scene multi-label emotional dialogue database[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2022: 5699-5710.

[24] Tolmeijer S, Kneer M, Sarasua C, et al. Implementations in machine ethics: a survey[J]. ACM Computing Surveys, 2021, 53(6): 1-38.

[25] Gips J. Toward the ethical robot [M]. Cambridge, MA: MIT Press, 1995: 243-252.

[26] Gao Qingqing, Cao Biwei, Guan Xin, et al. Emotion recognition in conversations with emotion shift detection based on multi-task learning[J]. Knowledge-Based Systems, 2022, 248: 108861.

[27] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-Networks[C]//Proceed-

ings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL Press, 2017: 670-680.

[28] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2017: 670-680.

[29] Cer D, Yang Yinfei, Kong Shengyi, et al. Universal sentence encoder for English[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, PA: ACL Press, 2018: 169-174.

[30] Li yanran, Su Hui, Shen Xiaoyu, et al. DailyDialog: a manually labelled multi-turn dialogue dataset[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei: Asian Federation of Natural Language Processing, 2017: 986-995.

[31] Poria S, Hazarika D, Majumder N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2019: 527-536.

[32] Chen Yirong, Fan Weiquan, Xing Xiaofen, et al. CPED: a large-scale chinese personalized and emotional dialogue dataset for conversational AI[PP/OL]. arXiv(2022-05-29)[2023-05-29]. https://arxiv.org/abs/2205.14727v1.

[33] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Stroudsburg, PA: ACL Press, 2014: 1746-1751.

[34] Kenton J D M W C, Toutanova L K. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. Stroudsburg: ACL Press, 2019: 4171-4186.

[35] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2017: 873-883.

[36] Majumder N, Poria S, Hazarika D, et al. DialogueRNN: an attentive rnn for emotion detection in conversations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6818-6825.

[37] Ghosal D, Majumder N, Poria S, et al. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL Press, 2019: 154-164.

编辑：张所滨