DOI:10.11784/tdxbz202508010

# 基于语义对齐与模板可靠性的方面级情感四元组预测

雷子锟[1]，陈彦江[2]，张　凯[2]，刘　淇[1,2]
(1. 中国科学技术大学人工智能与数据科学学院，合肥 230027；
2. 中国科学技术大学计算机科学与技术学院，合肥 230027)

**摘　要**：方面级情感四元组预测(ASQP)任务旨在从文本中同时抽取方面项、方面类别、观点项和情感极性，是情感分析领域中的热门问题. 当下主流生成式方法通过多模板引导预训练语言模型生成四元组，然而在进行模板的选取时通常忽视了模板之间潜在的语义关联或各模板本身的性能差异. 为此，本文提出一种基于预训练语言模型 T5 的模板可靠性与语义对齐平衡(SMARTS)策略. 该策略在模板选择阶段，结合交叉熵和 Jensen-Shannon(JS)散度，构建多个兼顾生成性能与语义关联性的高质量模板分组. 首先，以交叉熵作为衡量指标，评估各候选模板在预训练语言模型下的生成性能，筛选出若干生成质量最优的模板；接着，考虑不同模板之间的结构和语义关联，模型进一步引入 JS 散度对模板间的分布差异进行建模，从而为每个性能最优模板匹配若干语义最相关的模板，形成多个高一致性的模板分组. 在推理阶段，为充分对齐 SMARTS 策略，本文提出了双层投票聚合策略，先在每组内选出最优结果，再在组间进行全局投票，提升预测稳定性和准确性. 在 2 个公开数据集 Rest15 和 Rest16 上实验结果表明了本文所提方法的有效性与优越性，且为了验证各个部分的有效性及更深入了解模型工作原理，文中进行了详细的超参数研究、消融实验以及案例研究.

**关键词**：方面级情感四元组预测；语义对齐；预训练语言模型；情感分析；深度学习
**中图分类号**：TP391.1　　**文献标志码**：A　　**文章编号**：0493-2137(2026)11-1187-10

# Aspect Sentiment Quad Prediction via Semantic Matching and Reliable Template Selection

Lei Zikun[1]，Chen Yanjiang[2]，Zhang Kai[2]，Liu Qi[1,2]
(1. School of Artificial Intelligence and Data Science，University of Science and Technology of China，Hefei 230027，China；
2. School of Computer Science and Technology，University of Science and Technology of China，Hefei 230027，China)

**Abstract**：Aspect sentiment quad prediction(ASQP) involves the simultaneous extraction of aspect terms，aspect categories，opinion terms，and sentiment polarities from text，presenting a significant challenge in sentiment analysis. Current mainstream approaches typically employ generative methods that utilize multiple templates to guide pre-trained language models in generating sentiment quads. However，these methods often neglect the semantic relationships between templates and the performance discrepancies among individual templates during the selection process. To overcome these limitations，this study introduces a strategy called semantic matching and reliable templates selection(SMARTS)，which is based on the pretrained language model T5 and aims to balance template reliability with semantic alignment. SMARTS first evaluates the generation performance of candidate templates using cross-entropy to filter out those with optimal generation quality. Subsequently，considering the structural and semantic correlations among different templates，it further introduces Jensen-Shannon(JS) divergence to model their distribution differences. This allows SMARTS to match each optimal template with several of the most semantically related ones，thereby forming multiple highly consistent template groups. To align with the SMARTS strategy during inference，a

**收稿日期**：2025-08-12；**修回日期**：2025-10-24.
**作者简介**：雷子锟(2003—　　)，男，硕士研究生，zklei0222@mail.ustc.edu.cn.
**通信作者**：刘　淇，qiliuql@ustc.edu.cn.
**基金项目**：国家自然科学基金资助项目(62337001).
**Supported by** the National Natural Science Foundation of China(No. 62337001).

two-stage voting aggregation mechanism is introduced. First, an intra-group voting process identifies the best candidate within each group. Then, a global voting stage aggregates results across groups to determine the final prediction, thereby enhancing both stability and accuracy. Experimental results on two widely-used benchmarks, Rest15 and Rest16, demonstrate the effectiveness and superiority of the proposed model. In addition, extensive experiments, including hyperparameter analysis, ablation studies, and case studies, are conducted to validate the contributions of individual components and to provide deeper insights into the model's behavior.

**Keywords**: aspect sentiment quad prediction; semantic matching; pre-trained language models; sentiment analysis; deep learning

情感分析作为自然语言处理领域中的一个重要分支,一直以来吸引着众多研究人员的关注,随着研究的发展,情感分析任务的处理方法和理论也逐渐完善,其应用也渐渐落地于不同领域,如在线论坛、电商平台、医疗等.

近年来,随着深度学习技术的发展,人们大多利用深度学习技术解决和探索各领域情感分析问题. 例如,利用深度学习技术对与微博评论进行情绪分类[1];又如将预训练大型语言模型(如 LLaMA2-7B)应用于金融新闻标题的情感分类任务[2];利用深度学习模型分析奥运会相关推文的情感,以理解全球观众的情绪等[3].

早期研究通常利用传统的机器学习方法进行情感分析,但是这种方式受到分类器和特征选择的影响,仍存在一定的局限性. 随着深度学习研究的发展,各种神经网络和变体不断被提出,相应地,在情感分析领域,研究者们也开始利用深度学习的理论方法解决情感分析问题. Wang 等[4]首次将注意力机制引入自然语言处理领域,提出了 ATAE-LSTM 模型,在方面级情感分析任务中取得良好成效. Tan 等[5]提出一种双注意力机制网络模型,更好地识别出文本中存在的冲突情感极性表现.

除了上述研究成果外,研究者们还利用图神经网络进行情感分析领域的探索. Zhang 等[6]提出了一种基于图卷积网络的方法,鲍小异等[7]设计了一种结合选择性注意力机制的模型.

近些年来,随着 Transformer[8]以及之后各种基于 Transformer 架构的预训练语言模型的诞生,利用诸如 BERT 等[9]预训练语言模型进行微调来解决 NLP 任务逐渐成为主流,这一趋势在情感分析领域同样显著. Li 等[10]利用 BERT 的上下文嵌入在端到端的方面级情感分析(aspect-based sentiment analysis, ABSA)任务上进行建模, Sun 等[11]对 BERT 模型进行了微调,从句子对分类的角度探讨了 ABSA 任务. 当下,人们仍在不断对 BERT 进行进一步的研究,探索其在情感分析[12]、问答系统[13]等任务中的潜力.

方面级情感四元组预测(aspect sentiment quad prediction, ASQP)是 ABSA 的一项重要任务,旨在一次性提取出文本中的 4 种情感元素:方面项、方面类别、观点项、情感极性. 早期 ABSA 的研究集中在单一 ABSA 任务上,而更近期的工作聚焦于研究复合 ABSA 任务,同时提取多个相关联的情感元素. Peng 等[14]提出了方面情感三元组提取任务, Wan 等[15]处理了目标方面情感检测任务.

在这些任务中最具挑战性的当属 ASQP,处理 ASQP 任务的方法主要分为流水线式和生成式两类.

流水线式方法将 ASQP 任务拆分为多个子任务依次完成. Cai 等[16]在最早提出情感四元组预测任务概念时,便提供了 4 个基于流水线方法的基准线. 这类方法虽然具有较高的可解释性,但易受误差累积影响,导致整体性能受限.

为解决流水线方法的误差传播问题,研究者们转而采用生成式方法. 这类方法将 ASQP 建模为序列到序列的生成问题,借助预训练语言模型的强大能力实现端到端处理. Zhang 等[17]提出了 Paraphrase 模型,利用预训练语言模型的生成能力将 ASQP 任务转换为释义生成问题. Hu 等[18]提出 DLO 和 ILO 模型,通过改变指定生成模板中情感元素的顺序,选取表现更佳的几个模板进行预测. Gou 等[19]提出了 MvP(multi-view prompting)方法,同样采用多模板提示学习并聚合结果.

上述基于多模板的生成式模型在 ASQP 任务上不断取得突破,但在模板选择上仍存在局限性:即在构建模板组合时,或侧重于评估各模板的生成质量而忽视了模板间的语义关联,或强调模板间的相关性却未能充分考虑模板本身的性能差异.

本文提出的语义对齐与模板可靠性(semantic matching and reliable template selection, SMARTS)策略正是为了同时解决这两个问题,通过综合考虑模板的自身性能(交叉熵)与模板间的语义一致性(Jensen-Shannon 散度, JS 散度)来构建高质量的模板分组,并在推理阶段采用双层投票机制进行聚合,取得了比

基准线更好的综合实验效果.

## 1　SMARTS模型结构及核心策略

### 1.1　SMARTS模型结构

本文提出的 SMARTS 模型整体框架基于预训练语言模型 T5 构建. 其核心在于通过创新的 SMARTS 策略与双层投票聚合策略，有效平衡模板的生成可靠性与语义对齐，以提升预测的准确性和稳定性. 模型结构如图 1 所示，其工作流程可分为 4 个主要阶段.

阶段 1：采取 SMARTS 策略选取参与微调的模板，得到模板的 $k$ 最大相关分组. ASQP 任务要求提取 4 种情感元素，作为生成式方法的生成模板中，所有的情感元素具有 24 种不同的排列. 对于所有 24 种顺序，先通过 T5 模型从中选取出性能最好的 $k$ 个模板，这 $k$ 个模板隶属于不同的组；接着对于 $k$ 个模板中的每一个主模板，通过 T5 模型找寻所有模板中和其计算 JS 值最小的 $m-1$ 个模板进行进一步分组.

阶段 2：利用上述所有模板构造生成目标对 T5 进行微调.

阶段 3：模型推理生成阶段，将待预测文本输入已经微调的 T5 模型，模型会按照 $k$ 最大相关分组多路生成 $k$ 组结果.

阶段 4：推理结果的聚合阶段，模型在 $k$ 组结果中的每一组中先进行一轮投票，总共选出 $k$ 个候选结果；接着在组间进行投票第 2 轮投票，选出最终结果.

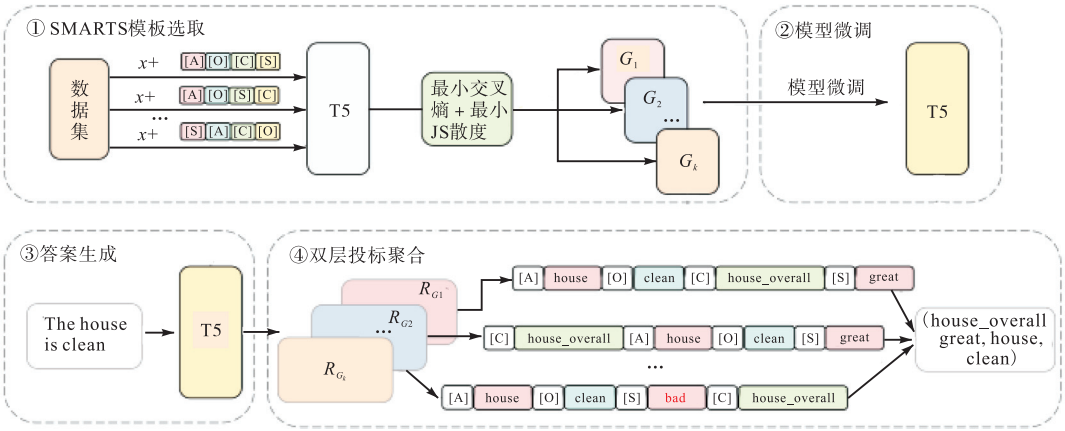



**图 1　模型结构示意**

**Fig.1　Schematic of the model structure**

### 1.2　交叉熵与高性能模板选择

不同情感元素顺序的模板对于生成式模型的引导具有不同性能，SMARTS 方法首先需要选取具有最好性能的 $k$ 个模板，其依据是平均交叉熵，在情感四元组预测任务中作为生成质量的评估指标. 依据此指标能够有效地指导模型选择在生成过程中的高质量模板. 对于模板 $O_i$，输入为 $x$，引导模型生成的输出为 $Y_i$，通过 T5 模型评分得到条件生成分数 $p=(Y_i\,|\,x)$，并在训练集 $D$ 上计算 $O_i$ 的平均分数作为此模板的最终性能指标，计算公式为

$$S_{O_i}=\frac{\sum_{D} p(Y_i\,|\,x)}{S_D} \tag{1}$$

式中：$S_D$ 为训练集样本数量；$D$ 为训练集.

对于具有不同情感元素顺序的模板 $O_i$，其分数 $S_{O_i}$ 能够直接反映模板在生成情感四元组时的准确度，$S_{O_i}$ 越低，表明模型根据引导模板抽取的结果精度越好. 因此，根据 $S_{O_i}$ 便可以对于 $O_i$ 的性能进行评估，按照分数从低到高进行排序，排序结果的前 $k$ 个就是在所有模板中表现性能最佳的 $k$ 个模板. 以平均交叉熵作为指标评估不同的模板性能的指标，模型可以有效筛除低质量模板，避免其在后续的推理过程中产生错误的情感推断，保障了后续推理的精度和质量.

### 1.3　JS 散度与语义相似性

在情感四元组预测中，模板中情感元素排列顺序的差异，可能导致生成结果在语义上不一致，这种不一致性会影响聚合阶段的准确性. 基于此种考量，SMARTS 方法利用 JS 散度度量模板之间的语义相似性，选择语义相近的模板进行分组，并通过后续聚合策略产生结果，以确保推理阶段生成的四元组在语义上更加一致，从而有效地避免生成过程中模板间产生的语义冲突，提高模型的生成精度和鲁棒性.

具体而言，对于输入 $x$，其对应模板 $O_i$ 和 $O_j$ 的生成结果是 $Y_i$ 和 $Y_j$，从 T5 解码器端得到对应生成序列的四元组部分 Logits 值为 $h_i$ 和 $h_j$. 那么对于 $O_i$ 和 $O_j$ 而言，它们的 JS 散度计算式为

$$D_{\mathrm{KL}}(p\,\|\,q)=\sum_{|V|}p\lg\left(\frac{p}{q}\right) \tag{2}$$

$$D_{\mathrm{JS}}(h_i\,\|\,h_j)=\frac{1}{2\,|\,h_i\,|}\sum_{|h_i|}D_{\mathrm{KL}}\left(h_i\,\|\,\frac{h_i+h_j}{2}\right)+$$

$$D_{\mathrm{KL}}\left(h_j\,\|\,\frac{h_i+h_j}{2}\right) \tag{3}$$

式中：$|V|$ 为词汇集大小；$|h_i|$ 为所选四元组对应 token 的数量；$D_{\mathrm{KL}}$ 为模板 $O_i$ 和 $O_j$ 的 KL 散度. 对于训练集 $D$，模板 $O_i$ 和 $O_j$ 的平均 JS 散度值计算为

$$S_{O_i,O_j}=\frac{1}{|D|}\sum_{D}D_{\mathrm{JS}}(h_i\,\|\,h_j) \tag{4}$$

式中 $S_{O_i,O_j}$ 为 $O_i$ 和 $O_j$ 之间的相关程度，值越小则表示两模板的语义相关性越高.

对于所有 24 个模板，SMARTS 方法首先两两计算它们之间的 $S$ 值，再对于第 1.2 中所选取 $k$ 个高质量模板中的每一个模板，找到与其 $S$ 值最小的 $m-1$ 个模板，聚合形成一组，从而得到 $k$ 组最大相关模板分组.

## 2 模型工作流程

### 2.1 基于元素顺序的提示学习

模型工作的第一阶段，模型首先需要对于训练集的输入和标签依据不同元素顺序的模板进行目标构造. 具体而言，模型遵循 DLO[18]的构造方法，本文使用元素标记来表示情感元素的信息结构，对应方面项、方面类别、观点项、情感极性这 4 元素的标签分别为[A]、[C]、[O]、[S]. 对于训练集中的数据 $(x,\ y)$，$x$ 为输入，对应待分析的文本；$y$ 为情感四元组，可能存在多个. 将模板对应的元素标记按序排列在输入末尾，形成新的输入 $X$；将模板对应元素标记嵌入目标头部，连接成句，多句用[SSEP]进行分割，形成新目标 $Y$. 例如，对于模板[A] [C] [O] [S]，对于以下例子，原输入是 $x$："The house is clean"，原目标是 $y$：(house_overall, great, house, clean)，即

$X$: The house is clean [A] [C] [O] [S]

$Y$: [A] house [C] house_overall [O] clean [S] great

此种对于训练集输入和目标设计可以有效地促进 T5 模型对于情感元组的生成和提取能力，为后续生成任务铺垫.

当输入文本经过元素顺序模板拼接后，形成新的序列化输入形式 $X$，其中包括原始句子与代表情感元素的标记(如 [A] [C] [O] [S])作为提示信息. 该输入被送入 T5 分词器进行子词级编码，生成 token 序列，并通过共享词嵌入矩阵映射为 embedding 表示

$$\boldsymbol{X}_{\mathrm{emb}}\in\mathbf{R}^{L_x\times d_{\mathrm{model}}} \tag{5}$$

式中：$L_x$ 为输入序列长度(最大为 512)；$d_{\mathrm{model}}=768$ 为本文使用 T5-base 模型的特征维度. 每个 embedding 表示中包含词嵌入(token embedding)和位置编码(position embedding)，用于建模序列信息.

接下来，嵌入表示 $X_{\mathrm{emb}}$ 被输入至 T5 的 Encoder 模块，编码器由 12 层 Transformer 堆叠构成，每层包括多头自注意力机制与前馈神经网络，并通过残差连接与 LayerNorm 实现稳定训练. 编码器输出的上下文表示 $\boldsymbol{H}_{\mathrm{enc}}$ 为

$$\boldsymbol{H}_{\mathrm{enc}}\in\mathbf{R}^{L_Y\times d_{\mathrm{model}}} \tag{6}$$

式中 $L_Y$ 为目标序列长度.

在生成阶段，T5 的 Decoder 采用自回归生成策略. 在训练过程中，使用 teacher forcing 机制，将目标序列 $Y$ 作为输入送入解码器的下一个位置. 每一个时间步中，解码器利用之前已生成的 token 与编码器输出进行交互，通过交叉注意力机制得到当前步的隐藏状态. Decoder 输出为

$$\boldsymbol{H}_{\mathrm{dec}}\in\mathbf{R}^{L_y\times d_{\mathrm{model}}} \tag{7}$$

最后，Decoder 输出经过一个共享的词嵌入矩阵 $\boldsymbol{W}\in\mathbf{R}^{V\times d_{\mathrm{model}}}$ (其中 $V=32\,128$ 为词表大小)进行线性投影，并通过 softmax 得到每个时间步对应的 token 预测概率.

整个训练过程以最小化预测序列与真实目标之间的 token-level 交叉熵损失为目标函数. 通过"模板构造→编码→解码→输出"的处理流程，T5 能够将模板中注入的结构信息编码入生成语义中，从而有效提升情感四元组的结构完整性与生成精度.

### 2.2 高性能模板组构造

SMARTS 方法利用预训练语言 T5 计算每一个模板 $O_i$ 上训练样本的平均交叉熵损失，进行 24 种情感元素组合顺序模板的性能评估，得到 $O_i$ 的表现分数 $S_{O_i}$. 选取 $k$ 个具有最小交叉熵(最佳性能)模板后，利用 JS 散度以这些模板为主模板进行聚合、分组，每个模板聚合与其语义相似度最大的 $m-1$ 个其余模板，得到语义相关、性能最佳的 $k$ 个模板组.

### 2.3 训练阶段

当选择出 $k$ 组最大相关模板后，其中的 $k\times m$ 个模板中可能存在重复，先要对重复模板进行去重，得到最终的 $n$ 个模板作为候选模板. 对于这 $n$ 个模板，按照第 2.1 所述的方式进行输入和目标的构造，再对 T5 模型进行微调，优化负对数似然损失函数为

$$L(X,Y_i)=-\sum_{t=1}^{L_Y}\lg p_\theta\left(Y_{i,t}\mid X,Y_{i'<t}\right) \tag{8}$$

### 2.4　推理阶段

在推理阶段，针对 $k$ 最大相关分组的特点，本文提出一种双层投票机制，保证最终结果的精度.

#### 2.4.1　组内选取

对于每个最大相关分组 $G_i$，模型首先将输入对应 $G_i$ 中的每一个模板 $T_i$ 进行输入构造，经过已经微调的 T5 模型推理生成 $m$ 个结果. 接着，模型提取 $m$ 个结果序列对应的情感四元组，然后从这 $m$ 个四元组候选项中选举出一个代表结果

$$R_i=\left\{r\mid r\in\bigcup_{i=1}^{m}T_i,\left(\sum_{i=1}^{m}\mathbb{I}\left[T_i(r)\geqslant\tau\right]\right)\right\} \tag{9}$$

式中：$R_i$ 为每组的结果；$\mathbb{I}[\cdot]$ 表示指示函数，输入为 True 时 $\mathbb{I}=1$，输入为 false 时 $\mathbb{I}=0$；$\tau$ 为选取的阈值，统计出现次数需要超过 $\tau$ 才会被选举，这里的 $\tau$ 设置为大于等于 $\frac{m}{2}$，这是因为要求得票数过半可以提升决策的置信度，避免因票数分散而产生预测歧义，从而提高精确率，避免预测误差对结果产生干扰.

#### 2.4.2　组间选取

在 $k$ 组最大相关分组中的每一组选取出了候选项 $R_i$ 后，模型最后推理最终结果 $P$. 和组内投票策略类似，$k$ 组候选项代表了 $k$ 个性能最佳模板生成的四元组结果，取其中超过半数的结果作为最终结果. 组间投票公式为

$$P=\left\{p\mid p\in\bigcup_{i=1}^{k}R_i,\left(\sum_{i=1}^{k}\mathbb{I}\left[R_i(p)\geqslant\frac{k}{2}\right]\right)\right\} \tag{10}$$

式中 $P$ 代表最终结果四元组.

## 3　实验结果及分析

### 3.1　数据集

本文实验在 2 个公开数据集上进行：Rest15 和 Rest16. 这 2 个数据集最早源自 SemEval 任务，后来逐渐被研究者们标注完善. 在 Zhang 等[17]比对和补充后，两个数据集中的每一个实例都包含一个评论句，每句话都包含 1 个或多个情感四元组. 这 2 个数据集的统计数据如表 1 所示.

### 3.2　评价指标

ASQP 任务要求预测出的四元组的四元素和样本标签完全一致才视作预测正确，否则不计入结果. 为评估模型效果且与基准线评估方式保持一致，本文使用精确率 $r_{\mathrm{p}}$、召回率 $r_{\mathrm{R}}$ 和 F1 分数 3 种指标对于模型进行评估，其中 F1 分数由精确率和召回率计算得到，是最为关键的指标，更能够综合综合评估模型性能，其计算公式为

$$\mathrm{F1}=\frac{2r_{\mathrm{p}}r_{\mathrm{R}}}{r_{\mathrm{p}}+r_{\mathrm{R}}} \tag{11}$$

**表 1　Rest15 和 Rest16 数据集数据统计**

**Tab.1　Statistics for the Rest15 and Rest16 datasets**

| 数据集 | Rest15 | | | | Rest16 | | | |
|---|---|---|---|---|---|---|---|---|
| | 样本数量 | 积极四元组数量 | 中性四元组数量 | 消极四元组数量 | 样本数量 | 积极四元组数量 | 中性四元组数量 | 消极四元组数量 |
| 训练集 | 834 | 1 005 | 34 | 315 | 1 264 | 1 369 | 62 | 558 |
| 验证集 | 209 | 252 | 14 | 81 | 316 | 341 | 23 | 143 |
| 测试集 | 537 | 453 | 37 | 315 | 544 | 583 | 40 | 176 |

### 3.3　实验细节及参数设置

本文实验使用 Python 作为开发语言以及 Pytorch 框架对模型进行编写，采用 Huggingface 中的 T5-BASE 模型作为预训练模型，在解码阶段使用贪婪解码进行搜索.

训练主要涉及的超参数有：分组的组数 $k$、每组的大小 $m$ 以及组内投票的阈值 $\tau$.

首先对 $k$ 和 $m$ 进行设置. 由于选择参与微调的模板数量越多，训练开销越大，带来的噪声也越多；同时考虑到覆盖率问题，假设 $k$ 最大相关组中无重复模板，$k\times m$ 的值就是参与微调模板的数量，此数值应当控制在一定的区间，即 24 种不同组合数量一半的附近或以内. 实验主要采用了 3 种规格的 $k$ 和 $m$ 组合，分别为：①$k=1$，$m=7$；②$k=3$，$m=3$；③$k=3$，$m=5$.

其次是有关 $\tau$ 的数值，即组内投票阈值，对于 $k=1$，$m=7$、$k=3$，$m=3$ 在两个数据集上分别设置为 6.0、3.0，对于 $k=3$，$m=5$ 在 Rest15 上设置为 2.5，在 Rest16 上设置为 4.0. 实验训练的学习率设置为 $1\times10^{-4}$，使用 AdamW 优化器作为优化算法，迭代次数为 20 次，训练批量大小设为 16，验证批量大小设为 64.

### 3.4　对比基准方法

本实验选取近年来处理 ASQP 任务中的先进基

准方法(模型)作为对比,包括:

(1) TAS-BERT[15]:TAS 用于提取情感三元组,经扩展后改变标注模式,可以用于提取四元组;

(2) Extract-Classify[16]:在原始文本中提取出方面项和观点项,并通过分类的方法得到方面类别和情感极性;

(3) GAS[20]:直接将情感四元组序列作为目标序列进行目标生成;

(4) Paraphrase[17]:同样作为生成式方法,Paraphrase 通过将生成目标转义为文本进行训练,能够更好利用预训练语言模型的知识;

(5) DLO/ILO[18]:基于不同元素顺序的生成模板作为数据增强扩充 ASQP 数据集;

(6) MvP[19]:基于不同元素顺序模板提示学习方法,多视角聚合结果进行预测.

### 3.5 实验结果

本实验方法(SMARTS)和基准线的数据对比如表 2 所示.

**表 2 实验结果对比**

**Tab.2 Comparison of experimental results**

| 模型(方法) | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|
| | $r_P$/% | $r_R$/% | F1 | $r_P$/% | $r_R$/% | F1 |
| TAS-BERT | 44.24 | 28.66 | 34.78 | 48.65 | 39.68 | 43.71 |
| Extract-Classify | 35.64 | 37.25 | 36.42 | 38.40 | 50.93 | 43.77 |
| GAS | 45.31 | 46.70 | 45.98 | 54.54 | 57.62 | 56.04 |
| Paraphrase | 46.16 | 47.72 | 46.93 | 56.63 | 59.30 | 57.93 |
| DLO | 47.08 | 49.33 | 48.18 | 57.92 | 61.80 | 59.79 |
| ILO | 47.78 | 50.38 | 49.05 | 57.58 | 61.17 | 59.32 |
| MvP(top-3) | 49.59 | 48.93 | 49.32 | 60.27 | 58.94 | 59.62 |
| MvP(top-11) | 49.26 | 50.19 | 49.72 | 59.56 | 61.58 | 60.55 |
| SMARTS(1 × 5) | 50.89 | 50.44 | 50.66 | 61.22 | 60.08 | 60.64 |
| SMARTS(1 × 7) | 52.13 | 49.31 | 50.68 | 60.81 | 58.07 | 59.41 |
| SMARTS(3 × 3) | 50.45 | 49.56 | 50.00 | 58.65 | 60.70 | 59.66 |
| SMARTS(3 × 5) | 49.69 | 50.94 | 50.31 | 61.74 | 61.20 | 61.47 |
| SMARTS(5 × 3) | 49.20 | 50.06 | 49.63 | 58.87 | 61.08 | 59.95 |

实验中,SMARTS 方法括号后的 $k \times m$ 表示选取模板设置为 $k$ 组,每组聚合 $m$ 个最大相关模板.

如表 2 所示,SMARTS 方法的 1 × 7 模型在 Rest15 上取得最佳表现,相较于表现最好的基准线 MvP(top-11)在 F1 分数上有着 0.96%的提升;3 × 5 模型在 Rest16 数据集上表现最佳,相较 MvP(top-11)有 0.92%的提升. 值得注意的是,对于 1 × 5 模型而言,即使相较于 3 × 5 模型在规模上有所减小,但是仍然能够取得比基准线更好的表现. 实验结果表明了 $k$ 最大相关分组选取策略的有效性,从模型的性能和稳定性两方面考虑,比基准线方法更加全面,从而取得了更好的效果.

### 3.6 超参数研究

本文模型涉及到的关键超参数主要有 3 个:模板分组组数 $k$、每组成员个数 $m$ 以及组内投票阈值 $\tau$. 为了进一步探究这些超参数对于模型性能的影响,本文通过设置这些参数不同组合进行实验.

#### 3.6.1 组数 $k$ 和组成员个数 $m$

模型参数 $k$ 表示预训练时依据最小交叉熵选取的表现性能最好模板数量;$m$ 表示根据每组的组代表模板和依据最大相关标准选取后每一组成员的数量. 如果不考虑选取的重复,那么最终进行微调的模板数量是 $k \times m$,而模板数量越多,模型计算开销越大,语义不相关模板带来的噪声也会越多,因此要将 $k \times m$ 控制在一个合理的范围内,本文所选取 $k$ 和 $m$ 的组合不超过 15,在所有模板总数 24 的半数 12 附近.

针对 $k \times m$ 的具体设置,本文选择了 3 × 3、3 × 5 和 5 × 3 3 种模型在 Rest15、Rest16 两个数据集上的 F1 分数进行研究和分析. 实验结果如图 2 所示,分别展示了 3 种参数设置下模型在 Rest15 和 Rest16 上 F1 分数.

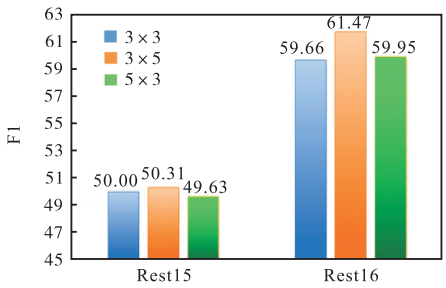



**图 2 3种参数设置下模型 F1分数对比**

**Fig.2 F1 score of the model under the three parameter settings**

（1）固定 $k=3$，增加 $m$. 在代表模板数 $k=3$ 固定的条件下，将每组内模板数 $m$ 从 3 扩展至 5，模型的 F1 分数在 Rest15 和 Rest16 数据集上分别出现 0.31%与 2.08%的提升. 这表明，适度扩大每组的高相关模板规模，可从多个语义视角增强预训练语言模型对任务结构的理解能力. 同时，在推理阶段，更多组内成员参与投票也增强了输出结果的容错性和稳定性. 因此，增加 $m$ 有助于提升生成质量与决策鲁棒性.

（2）固定 $m=3$，增加 $k$. 当组内成员数 $m=3$ 不变，将代表模板数 $k$ 从 3 扩展至 5，模型在 Rest16 上略有提升（0.29%），但在 Rest15 上出现性能下降（–0.37%）. 由于 $k$ 控制的是选取交叉熵最小的主模板，其间往往缺乏语义关联，过多不相关的主模板在预训练阶段难以形成协同学习信号，甚至可能干扰模型学习统一的生成模式. 在数据资源较为有限的场景下，这种“信息分散”带来的负面影响尤为显著. 因此，仅单方面增加 $k$ 对性能提升的效果有限.

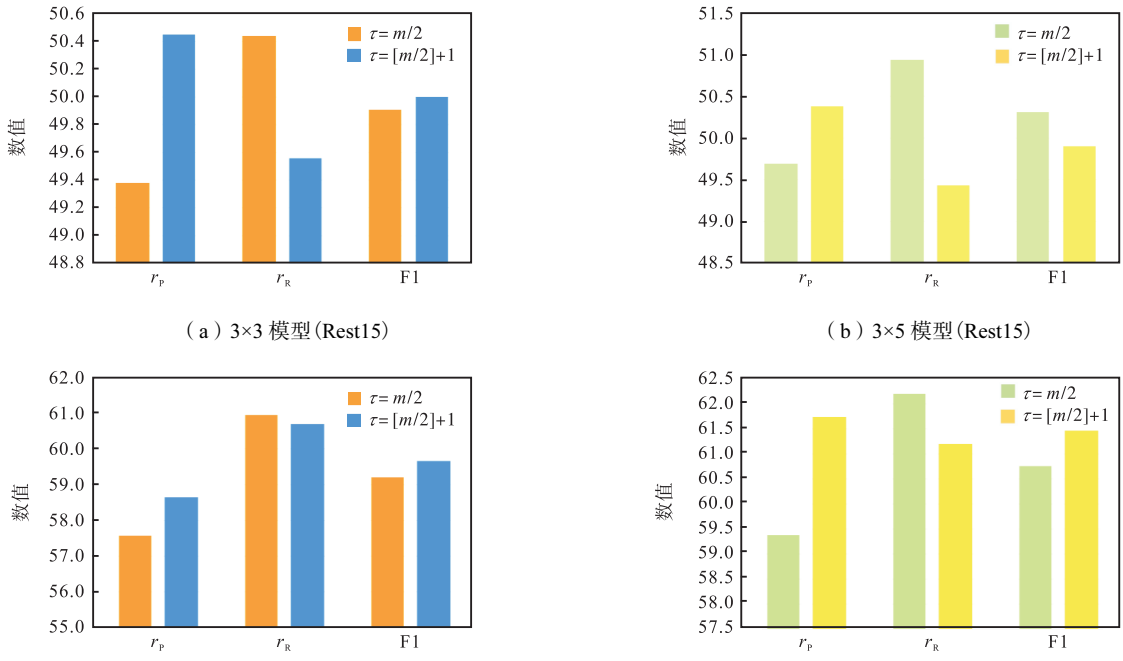
（3）固定总模板数，对比 $3\times5$ 与 $5\times3$. 在保持参与预训练的总模板数量（即 $k\times m=15$）一致的前提下，对比不同 $k$ 与 $m$ 组合可以更清楚地揭示性能与相关性的权衡情况. 实验结果显示，采用 $3\times5$ 结构在 Rest15 和 Rest16 上的 F1 分数相较于 $5\times3$ 分别提升了 0.68%和 1.79%. 这说明，在相同训练资源约束下，优先选择更少但性能强的主模板，并结合更多语义相近的辅助模板，不仅能提高信息一致性，还有助于模型更充分地学习到任务结构，从而增强模型的泛化能力.

综上，合适的 $k$ 值和 $m$ 值对于模型表现有着较为明显的影响，不能通过盲目增加 $k$ 或者增加 $m$ 来实现更好的效果，而是需要选取适当的值来达到 T5 模型预训练的视角数、模板质量等因素的平衡. 实际应用中，需要根据具体数据集和具体的需求观察并动态调整 $k$ 和 $m$ 的大小.

### 3.6.2　组内投票阈值 $\tau$

组内投票阈值 $\tau$ 表示模型的推理阶段中每组代表被选举出来的置信水平. 较高的 $\tau$ 值表示由每组预测的四元组更频繁地被多个模板匹配，从而增强每组代表预测正确性的置信度. 图 3 展示了 $3\times3$ 和 $3\times5$ 两种 $k$ 和 $m$ 参数设置下不同 $\tau$ 的设置下模型的表现结果.

由图 3 观察可知，在同种 $k$ 和 $m$ 的参数设置、同一数据集上，提升阈值 $\tau$ 会引起精确率和召回率的动态变化，进而影响 F1 分数的变化. 具体而言，当阈值 $\tau$ 提升时，会使得模型答案置信度变高，从而使得精确率变高，召回率降低. 当这种动态变化维持在一定范围中时，可以使得模型综合性能（F1 分数）得到提升. 除了在 $3\times5$ 规格、Rest15 数据集中，其余 3 种情况提升阈值 $\tau$ 能够使得 F1 分数提升，其中在 $3\times5$ 设置下 Rest16 上能够将 F1 分数提高 0.71%.




（a）3×3 模型（Rest15）　（b）3×5 模型（Rest15）

（c）3×3 模型（Rest16）　（d）3×5 模型（Rest16）

**图 3　两种参数设置下模型不同 $\tau$ 值实验结果**

**Fig.3　Experimental results of different $\tau$ values of the model under two parameter settings**

另外,本文在 3 × 5 规格模型中固定 $\tau$ 值为 $m/2$,对组间投票阈值 $\tau_k$ 进行调整,得到数据如表 3 所示. 表 3 展示了 3 × 5 规格模型不同 $\tau$ 和 $\tau_k$ 值在 Rest15 和 Rest16 上精确率、召回率和 F1 分数. 由表 3 可知,尽管提升 $\tau_k$ 也能带来精确率和召回率的动态调整,但是都不能取得最好的效果. 相较于调整 $\tau$ 值,调整 $\tau_k$ 在精确率和召回率上变化幅度会减小. 由于组间投票是推理聚合的第 2 阶段,在第 1 阶段已经有相当的门槛使得组代表入选,此时再调整阈值对于答案输出的变化会不如在第 1 阶段调整那般敏感.

**表 3 模型取不同 $\tau$ 和 $\tau_k$ 值时实验结果对比**

**Tab.3 Comparison of experimental results under different values of $\tau$ and $\tau_k$**

| 模型参数 | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|
| | $r_P$/% | $r_R$/% | F1 | $r_P$/% | $r_R$/% | F1 |
| $\tau=\frac{m}{2}$, $\tau_k=\frac{k}{2}$ | 49.69 | 50.94 | 50.31 | 59.38 | 62.20 | 60.76 |
| $\tau=\left[\frac{m}{2}\right]+1$, $\tau_k=\frac{k}{2}$ | 50.38 | 49.43 | 49.90 | 61.74 | 61.20 | 61.47 |
| $\tau=\frac{m}{2}$, $\tau_k=\left[\frac{k}{2}\right]+1$ | 50.32 | 49.94 | 50.13 | 60.56 | 62.08 | 61.31 |

综上,在实际情况中,应当设置适当的 $\tau$ 值平衡精确率和召回率的值,必要时,还可以进一步调整 $\tau_k$. 通过这种平衡,可以使模型综合性能得到进一步提升.

### 3.7 消融实验

为了探究本文方法的有效性,本文在 Rest15 和 Rest16 上针对模板选取阶段的 $k$ 最大相关分组策略和推理阶段的聚合策略开展消融实验,设置模型参数为 $k=3$, $m=5$, $\tau=4$.

#### 3.7.1 模板选取策略

针对模板选取使用的最小交叉熵策略和最小 JS 散度策略,为验证其有效性,在保证最终聚合策略均为双层投票机制的情况下,本文设计并开展如下实验,实验结果如表 4 所示. 表 4 展示了不同模板选取方法在 Rest15 和 Rest16 上精确率、召回率和 F1 分数表现.

由表 4 可以观察到,SMARTS 策略的 F1 分数在两个数据集上均优于其他策略,尤其是在 Rest16 数据集上,优势更明显,在 F1 分数上平均提升了 1.15%,表明更多量的数据可以更加有效发挥 $k$ 最大相关分组策略的潜力. 这证明了 $k$ 最大相关分组模板选取方式能够有效促进 T5 预训练语言模型的学习,提升其生成能力,从而更好解决 ASQP 任务.

**表 4 不同模板选取方法实验结果对比**

**Tab.4 Comparison of experimental results with different template selection methods**

| 模版选取依据 | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|
| | $r_P$/% | $r_R$/% | F1 | $r_P$/% | $r_R$/% | F1 |
| 最大交叉熵 + 最小 JS 散度 | 50.46 | 48.81 | 49.62 | 60.45 | 60.45 | 60.45 |
| 最小交叉熵 + 最大 JS 散度 | 50.98 | 49.18 | 50.06 | 60.20 | 60.20 | 60.20 |
| 随机选取 | 48.87 | 51.21 | 50.02 | 59.64 | 61.01 | 60.32 |
| 最小交叉熵 + 最小 JS 散度(SMARTS) | 49.69 | 50.94 | 50.31 | 61.74 | 61.20 | 61.47 |

#### 3.7.2 聚合策略

针对模型的生成推理阶段采用的双层投票机制,为验证其有效性,在保证模板选取策略均为 SMARTS 策略的前提下,本文设计与单层投票和随机选取对比,并开展如下实验,实验结果如表 5 所示,表 5 展示了不同聚合策略在 Rest15 和 Rest16 上精确率、召回率和 F1 分数表现. 其中单层投票机制与双层投票机制不同在于最终推理阶段,单层投票不考虑 $k$ 最大相关分组,直接选取出现次数最多的答案作为结果.

从表 5 结果可以看出,相较于其他两种聚合策略,双层投票策略有着更高的 F1 分数,相较于单层投票机制,在 Rest15 和 Rest16 数据集上分别有 1.70%和 1.60%的提升. 并且,双层投票机制有着更为平衡的精确率和召回率. 这是由于和单层投票机制将所有候选项“混为一谈”相比,双层投票机制针对不同组选择组代表,再进行结果的选取,充分对齐 SMARTS 策略,是更为合理的聚合策略,故而表现出更好的性能.

**表 5 不同聚合策略实验结果对比**

**Tab.5 Comparison of experimental results with different aggregation strategies**

| 聚合策略 | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|
| | $r_P$/% | $r_R$/% | F1 | $r_P$/% | $r_R$/% | F1 |
| 单层投票 | 44.31 | 53.84 | 48.61 | 55.69 | 64.96 | 59.97 |
| 随机选取 | 48.32 | 50.69 | 49.48 | 59.67 | 63.33 | 61.45 |
| 双层投票 | 49.69 | 50.94 | 50.31 | 61.74 | 61.20 | 61.47 |

### 3.8 案例研究

为了更加全面了解本文提出模型的优势和不足,本文开展了案例研究,使用模型参数为 $k=3$, $m=5$, $\tau=4$. 其中一个案例是模型预测成功的例子,另一个是模型预测失败的例子,具体案例如表 6 和表 7 所示.

**表 6　案例 1：模型成功预测案例**

**Tab.6　Case 1：model successful prediction case**

| 项　目 | 结　果 |
|---|---|
| 输入案例 1 | As part of a small party of four, our food was dropped off without comment. |
| 标签 | (service general, null, bad, dropped off without comment) |
| 组 1 预测 | (service general, null, bad, dropped off without comment) × 4(选中)<br>(service general, food, bad, dropped off without comment) × 1(舍弃) |
| 组 2 预测 | (service general, null, bad, dropped off without comment) × 5(选中) |
| 组 3 预测 | (service general, null, bad, dropped off without comment) × 3(舍弃)<br>(service general, food, bad, dropped off without comment) × 2(舍弃) |
| 组代表集合 | (service general, null, bad, dropped off without comment) × 2(选中) |
| 最终预测结果 | (service general, null, bad, dropped off without comment) |

**表 7　案例 2：模型失败预测案例**

**Tab.7　Case 2：model failure prediction case**

| 项　目 | 结　果 |
|---|---|
| 输入案例 2 | Oh, I forgot to mention that they don't have bread service. |
| 标签 | (service general, null, bad, don't have bread service) |
| 组 1 预测 | (service general, null, bad, don't have bread service) × 2(舍弃)<br>(service general, bread service, bad, don't have bread service) × 3(舍弃) |
| 组 2 预测 | (service general, null, bad, don't have bread service) × 1(舍弃)<br>(service general, bread service, bad, don't have bread service) × 4(选中) |
| 组 3 预测 | (service general, null, bad, don't have bread service) × 2(舍弃)<br>(service general, bread service, bad, don't have bread service) × 3(舍弃) |
| 组代表集合 | (service general, bread service, bad, don't have bread service) × 1(舍弃) |
| 最终预测结果 | 无 |

在表 6 中，模型成功预测出了正确答案. 组 1 和组 3 在预测中存在成员模板预测错误. 其中组 1 中存在一个错误，4 个正确答案大于等于阈值 $\tau$，被成功选举作为组 1 代表，同时错误答案被模型成功过滤. 组 3 中存在 2 个错误答案和 3 个正确答案，根据聚合策略，没有答案等于或超过阈值 $\tau$，不输出任何值. 最终产生 2 个组代表，均为正确答案，大于 $k$ 值的一半，成功预测出结果. 由此可知，虽然在预测过程中存在成员的错误预测结果，但是通过双层投票机制能够有效防止偶然错误的干扰，从而提升模型性能.

在表 7 中，模型未能正确预测答案. 其中组 1 和组 3 没有代表产生，而组 2 错误答案超过阈值 $\tau$，成为唯一一个候选项，没能超过阈值，因此无答案产生. 观察组内正确答案和错误结果之间的区别仅在于方面项，这说明模型在生成能力上仍存在一定的局限性，经过训练后的 T5 模型未能稳定且正确生成包含正确方面项的答案，因而本文模型仍存在一定的潜力.

## 4　结　语

情感分析在当下发挥着越来越重要的作用，在舆情分析、电商销售、医疗健康等众多领域正得到广泛应用. 作为方面级情感分析最具挑战性的任务，ASQP 在近年来备受关注. 为解决现有生成式方法在模板选择上或忽视模板间语义关联、或忽视模板本身性能差异的局限性，本文提出的模型基于预训练语言模型 T5 的语义匹配及可靠模板分组选取策略(SMARTS)和双层投票机制聚合策略，在 2 个公开数据集 Rest15 和 Rest16 上相较于基准线取得了更好的成效. 尽管本文方法相较于基准线性能有所提升，但仍存在一定的局限性. 例如，模型表现受限于 T5 的生成能力以及固定的 24 种模板. 未来的工作可将此策略迁移至更先进的预训练语言模型，探索更丰富、灵活的模板设计(如大模型自动生成模板)，并尝试将该方法扩展至多模态情感分析任务，以处理更复杂的数据形式.

**参考文献：**

[1] 郭贤伟，赖　华，余正涛，等. 融合情绪知识的案件微博评论情绪分类[J]. 计算机学报，2021，44(3)：564-578.
Guo Xianwei，Lai Hua，Yu Zhengtao，et al. Emotion classification of case-related microblog comments integrating emotional knowledge[J]. Chinese Journal of Computers，2021，44(3)：564-578(in Chinese).

[2] Luo W，Gong D H. Pre-trained large language models for financial sentiment analysis[EB/OL]. https：//arxiv.org/abs/2401.05215，2024-01-25.

[3] Bandyopadhyay I, Karmakar R. Deep learning-based sentiment analysis of Olympics tweets[EB/OL]. https://arxiv. org/abs/2407. 12376, 2024-07-17.

[4] Wang Y Q, Huang M L, Zhu X Y, et al. Attention-based LSTM for aspect-level sentiment classification [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 606-615.

[5] Tan X W, Cai Y, Zhu C X. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong, China, 2019: 3426-3431.

[6] Zhang C, Li Q C, Song D W. Aspect-based sentiment classification with aspect-specific graph convolutional networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong, China, 2019: 4568-4578.

[7] 鲍小异, 姜晓彤, 王中卿, 等. 基于跨语言图神经网络模型的属性级情感分类[J]. 软件学报, 2023, 34(2): 676-689.
Bao Xiaoyi, Jiang Xiaotong, Wang Zhongqing, et al. Cross-lingual aspect-level sentiment classification with graph neural network[J]. Journal of Software, 2023, 34(2): 676-689(in Chinese).

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems(NeurIPS 2017). Long Beach, USA, 2017: 6000-6010.

[9] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186.

[10] Li X, Bing L D, Zhang W X, et al. Exploiting BERT for end-to-end aspect-based sentiment analysis[C]// Proceedings of the 5th Workshop on Noisy User-Generated Text(W-NUT 2019). Hong Kong, China, 2019: 34-41.

[11] Sun C, Huang L Y, Qiu X P. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 380-385.

[12] 林 煌, 李弼程. 基于 BERT 模型和图注意力网络的方面级情感分析[J]. 计算机科学, 2024, 51(增 2): 240400018-7.
Lin Huang, Li Bicheng. Aspect-based sentiment analysis based on BERT model and graph attention network[J]. Computer Science, 2024, 51(Suppl 2): 240400018-7(in Chinese).

[13] Sengupta P S, Sarkar S, Das D. HIJLI_JU at SemEval-2024 Task 7: Enhancing quantitative question answering using fine-tuned BERT models[C]//Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). Mexico City, Mexico, 2024: 279-284.

[14] Peng H Y, Xu L, Bing L D, et al. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 8600-8607.

[15] Wan H, Yang Y F, Du J F, et al. Target-aspect-sentiment joint detection for aspect-based sentiment analysis[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 9122-9129.

[16] Cai H J, Xia R, Yu J F. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 340-350.

[17] Zhang W X, Deng Y, Li X, et al. Aspect sentiment quad prediction as paraphrase generation[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 9209-9219.

[18] Hu M T, Wu Y K, Gao H, et al. Improving aspect sentiment quad prediction via template-order data augmentation[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 7889-7900.

[19] Gou Z B, Guo Q Y, Yang Y J. MvP: Multi-view prompting improves aspect sentiment tuple prediction [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 4380-4397.

[20] Zhang W X, Li X, Deng Y, et al. Towards generative aspect-based sentiment analysis[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 504-510.

(责任编辑: 王晓燕)