标签先验知识增强的方面类别情感分析方法研究

吴任伟[1], 李琳[1], 何铮[2], 袁景凌[1]
[1]武汉理工大学 计算机与人工智能学院,武汉 430070
[2]德勤有限公司,上海 510623
{wrw, cathylilin}@whut.edu.cn, zhhe@deloitte.com.cn, yuanjingling@126.com

摘要

当前，基于方面类别的情感分析研究旨在将方面类别检测和面向类别的情感分类两个任务协同进行。然而，现有研究未能有效关注情感数据集中存在的噪声标签，影响了情感分析的质量。基于此，本文提出一种标签先验知识增强的方面类别情感分析方法（AP-LPK）。首先本文为面向类别的情感分类构建了自回归提示训练方式，可以激发预训练语言模型的潜力。同时该方式通过自回归生成标签词，以期获得比非自回归更好的语义一致性。其次，每个类别的标签分布作为标签先验知识引入，并通过伯努利分布对其进行进一步精炼，以用于减轻噪声标签的干扰。然后，AP-LPK将上述两个步骤分别得到的情感类别分布进行融合，以获得最终的情感类别预测概率。最后，本文提出的AP-LPK方法在五个数据集上进行评估，包括SemEval 2015和2016的四个基准数据集和AI Challenger 2018的餐厅领域大规模数据集。实验结果表明，本文提出的方法在F1指标上优于现有方法。

关键词： 基于方面类别的情感分析 ； 提示学习 ； 标签先验知识

Aspect-Category based Sentiment Analysis Enhanced by Label Prior Knowledge

Renwei Wu[1], Lin Li[1], Zheng He[2], Jingling Yuan[1]
[1]School of Wuhan University of Technology, Wuhan 430070, China
[2]Department of Deloitte Limited, Shanghai 510623, China
{wrw, cathylilin}@whut.edu.cn, zhhe@deloitte.com.cn, yjl@whut.edu.cn

Abstract

Current aspect-category based sentiment analysis researches aim at performing joint aspect category detection and category-oriented sentiment classification. However, most of existing studies have not paid much attention on the noisy labels that often occur in sentiment datasets. To cope with this problem, we propose an aspect-category based sentiment analysis approach with Label Prior Knowledge(AP-LPK). Specifically, we firstly construct an Autoregressive Prompting training that can stimulate the potential of pre-trained language models. And then label words are generated through autoregression for better semantic consistency than non-autoregression. Secondly, we introduce the label distribution of each category as label prior knowledge that is refined through Bernoulli distribution to mitigate the interference of noisy labels. And then, the outputted labels from the autoregressive prompting and label prior knowledge refined work together based on the distributions of sentiment polarities to obtain final predictions. Finally, our AP-LPK approach is evaluated on five datasets that include the four benchmark datasets from SemEval 2015 and 2016 and the Restaurant-domain dataset from AI Challenger 2018. Experimental results demonstrate that our approach outperforms existing ones in terms of F1.

Keywords: Aspect-Category based Sentiment Analysis , Prompt learning , Label prior knowledge

第二十一届中国计算语言学大会论文集，第545页-第556页，南昌，中国，2022年10月14日至16日。

1 引言

随着互联网的发展，网络生活中衣食住行等服务或产品已经逐渐融入人们的日常生活，人们可以在网络生活中浏览、购买和使用这些产品，并分享自己对产品的看法或评论。以文本为主的评论涉及产品的不同方面，用户从产品的不同角度进行描述，包含有非常丰富的和有价值的多方面信息。因此文本评论具有广泛的研究场景，比如基于方面类别的消费评论（外卖、电商等）的情感分析，而对评论细粒度的情感分析有助于相关用户从中高效快捷地获取各方面的信息，为后续决策提供支持。

基于方面类别的情感分析(Aspect-Category based Sentiment Analysis(ACSA))在实际应用中逐渐变得越来越流行 (Schmitt et al., 2018; Dai et al., 2019; Guo et al., 2020; Cai et al., 2020; Fu et al., 2021; Hu et al., 2019; Liang et al., 2021)。ACSA旨在识别评论句子中的多个方面类别，并共同预测每个已识别类别的情感极性。

已有的研究重点在于如何建立类别-情感联合的神经网络模型，通过在情感标签空间中添加一个维度来指示每个类别的出现，例如Schmitt等人 (2018)，Dai等人 (2019)和Guo等人 (2020)。近年来，ACSA 转向预训练和微调范式。Cai等人 (2020)介绍了几种基于微调的方法，即Cartesian-BERT、Pipeline-BERT和AddOneDim-BERT。此外，他们将ACSA重新形式化为类别-情感层次预测问题，他们的方法可以对多个类别之间的内在关系以及类别与情感标签之间的相互关系进行建模，即Hier-BERT、Hier-Transformer-BERT和Hier-GCN-BERT。不过，尽管这些基于微调范式的模型稳定地优于基于传统神经网络的工作，但微调范式仍然存在一个主要障碍，即无法充分激发预训练语言模型的潜力，甚至会导致灾难性的遗忘问题 (Liu et al., 2021)。

在情感数据集中不可避免地会出现噪声标签，这是由于数据标注者的能力以及他们对标注标准的不同理解 (Zhou, 2018; Li et al., 2021)。以图 1中的两条评论为例。 虽然两条评论关于类

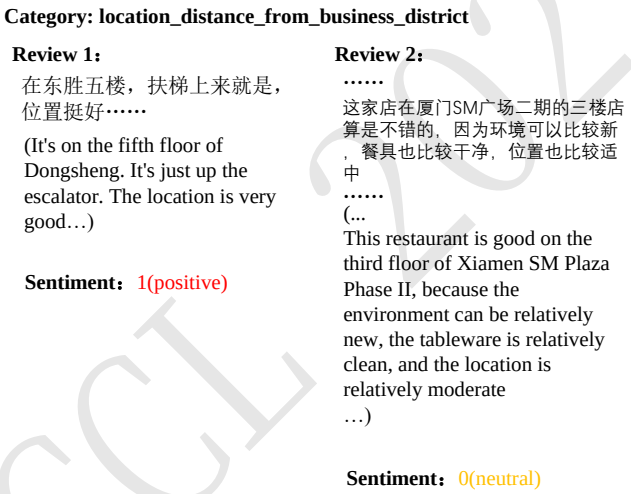


Figure 1: 两条评论样例

别*location_distance_from_business_district*的描述基本相同，但标注的情感极性不同。事实上，两者的情感极性都应该是积极的。更重要的是，这种噪声标签的情况在每个类别中并不少见。由于这些噪声标签会干扰模型性能 (Pan et al., 2022)，并且提示训练在各种NLP任务中表现出有效的性能，因此在ACSA任务中同时考虑它们是一个挑战。

针对该问题，本文提出了一种基于方面类别的带有标签先验知识的情感分析方法，即带有标签先验知识的自回归提示(Autoregressive Prompting with Label Prior Knowledge(AP-LPK))。在本文的自回归提示中，完形填空式的提示模板是一个包含类别的句子，含有多个连续的汉字掩码。由此，通过自回归语言模型 (Yang et al., 2019)，答案生成是自回归的。为了解决噪声标签的问题，本文通过结合伯努利分布和标签先验知识来校准自回归提示训练的输出标签，以生成最终预测。

本文的实验是在五个公开的数据集上进行，其中包括四个基准数据集和一个具有超过100k条带标签样本的中文大规模数据集。实验结果验证了提示训练在ACSA任务中的有效性。在此基础上，本文提出的AP-LPK方法可以通过处理标签噪声始终能促进预测质量的提升。

2 相关工作

近年来，基于方面类别的情感分析已经取得了一定的进展。现有研究按近代自然语言技术发展的范式 (Liu et al., 2021)大致分为两类，传统的神经网络模型和基于预训练和微调范式的模型

2.1 基于传统神经网络的模型

基于传统神经网络的工作采用的是基于神经网络的完全监督学习范式 (Liu et al., 2021)。在该范式下，基于方面类别的情感分析研究经历了两个阶段。

研究初期，学者们对于ACSA任务多采用Pipeline方法/多任务方法，即将ACSA任务分为方面类别提取(aspect category detection(ACD))任务和方面级情感分类(aspect level sentiment classification(ALSC))任务。Ruder等人 (2016)提出了分层神经网络模型来进行方面级情感分类，其中ACD假定基于Pipeline框架中的一些其他系统（如，SVM），而ALSC是模型的主要任务。

类似的，Hu等人 (2019)使用多任务的方式研究。在辅助任务ACD和主任务ALSC的设定下，他们提出了一种注意力网络CAN，以约束注意力权重分配，帮助学习更好的具体方面的句子表示。

上述的两项工作在公开数据集上都具有不错的表现。但是，Pipeline方法/多任务方法的任务分离会带来误差累积/传播问题 (Guo et al., 2020)，在一定程度上影响着模型效果。

因此，为了解决这个问题，近年来也有一些其他研究在探索联合学习ACD任务和ALSC任务。Schmitt等人 (2018)提出了一种联合模型，被Cai等人 (2020)称为AddOneDim-LSTM。该模型通过标签扩维，即在情感标签空间中添加一个维度来指示每个类别的出现，从而将ACD任务融入ALSC任务中，达到联合建模的目的。这代表了当时最先进的工作。

之后，一些其他的学者也以标签扩维为基础对ACSA进行研究。

Dai等人 (2019)为了捕获文本中的多共享特征以及特定类别的特征，提出了模型MMAM。该模型采用一个多头文档注意力机制作为记忆单元以编码共享的文档特征，并且采用一个多任务注意力机制来提取特定类别的特征。在两个真实数据集上的实验结果表明，MMAM具有良好的预测效果。

Guo等人 (2020)提出了一个改良的多路匹配深度神经网络模型以进行细粒度的情感分析，模型通过直接在多轮校准结构中捕获过去的注意力来改善现有的注意力，以预防误差传播和注意力缺失。

另外，Fu等人 (2021)基于双BiLSTM的多角度注意力，并在模型中引入特定方面类别的信息，提出了MPADB和MPADB_joint结构。MPADB通过丰富的上下文表示和多角度注意力机制避免方面和相应情感极性的错误匹配，而MPADB_joint是为了解决传统Pipeline方法的误差累积问题以及联合模型的注意力权重重叠问题，并提升模型的可解释性。在两个真实数据集上的实验结果表明了这项工作在准确性和可解释性方面的有效性。

然而，从2017年到2019年，自然语言处理模型的学习发生了巨大的变化，这种完全监督的范式现在发挥的作用越来越小 (Liu et al., 2021)。

2.2 基于预训练和微调范式的模型

近年来，随着BERT等一系列大规模预训练语言模型在NLP领域大放异彩，预训练和微调范式成为学术界和工业界广泛关注的重点。因此有学者开始应用该范式来研究基于方面类别的情感分析。Cai等人 (2020)基于预训练BERT和微调，提出了一系列方法，同时结合图卷积网络（GCN）来进行ACSA的研究。

- Cartesian-BERT：以BERT为句子编码器的笛卡尔法。
- Pipeline-BERT：ACD和ALSC都以BERT为编码器进行建模。
- AddOneDim-BERT：将AddOneDim-LSTM (Schmitt et al., 2018)中的LSTM替换为BERT。
- Hier-BERT：以BERT为句子编码器的类别-情感层次预测方法。
- Hier-Transformer-BERT：在Hier-BERT的基础上，Transformer被用来建模类别间的内在关系和类别与情感之间的相互关系。
- Hier-GCN-BERT：基于Hier-BERT，以层次图卷积网络（Hier-GCN）进行关系学习。

在四个公开数据集上的大量实验表明了Cai等人 (2020)工作的有效性，并且上述基于预训练BERT和微调的方法均在F1值方面优于Schmitt等人 (2018)的工作。

此外，Liang等人 (2021)通过探索基于外部知识的Beta分布引导的方面感知图构建，从一个新颖的角度研究ACSA任务。他们提出了AAGCN-BERT、AAGCN-BERT-c等模型，并且在六个基准数据集上的实验表明，他们的方法显著优于最先进的基线方法。

综上所述，基于方面类别的情感分析领域上的研究有一定的进展，具有非常好的理论指导作用。但几乎所有的ACSA工作都对数据集中经常出现的噪声标签缺乏关注，同时针对现有工作对提示学习的研究还不够丰富，因此我们的研究主要基于自回归提示学习，采用标签先验知识修正噪声标签问题，提出了一种方面类别情感分析方法AP-LPK，对用户评论文本进行预测。

3 任务定义

本文的任务是对于给定的待测评论文本和m个预定义的方面类别，能够提取所有提及的方面类别，并且识别每个被检测到的类别的情感。

给定一段具有n个字符的评论文本$\boldsymbol{r} = [w_1, ..., w_n]$，令$\boldsymbol{C} = \{c_1, ..., c_m\}$为$m$个预定义方面类别的集合，并且$s = \{negative, neutral, positive, not\ mentioned\}$为情感极性的标签集。因此，对于每一个输入$\boldsymbol{r}$，本项工作中ACSA的目标是生成情感极性集$y = \{..., \hat{y}_i, ..., \hat{y}_m\}$，其中$\hat{y}_i$表示评论文本$\boldsymbol{r}$中第i个方面类别对应的情感极性。

根据上述的定义，我们所提出的AP-LPK方法，其映射关系表示如下：

$$[\boldsymbol{r}^1, ..., \boldsymbol{r}^I, ..., \boldsymbol{r}^N] \xrightarrow{f_i(\cdot)} [\hat{y}_i^1, ..., \hat{y}_i^I, ..., \hat{y}_i^N] \tag{1}$$

其中，N表示数据集的大小，则$\boldsymbol{r}^N$表示第N条评论文本；$f_i(\cdot)$表示第i个类别对应的模型；$\hat{y}_i^I$表示第I条评论文本中第i个方面类别的情感极性。

4 带有标签先验知识的自回归提示

针对上述研究现状和任务定义，本文提出了AP-LPK方法，该方法结合提示学习、自回归模型与标签先验知识对输入的用户评论文本进行模型学习。AP-LPK能够构建ACSA任务的提示学习训练，得到面向特定方面类别的自回归语言模型，并且将提示学习的答案工程与标签先验知识相结合来实现噪声标签干扰的减轻。

4.1 AP-LPK框架

本文们提出方法的整体框架如图 2所示，该方法由两部分组成，自回归提示学习和标签先验知识的引入。 在我们的提示训练中，提示模板工程和答案工程是手动设计的。自回归语言模型（例

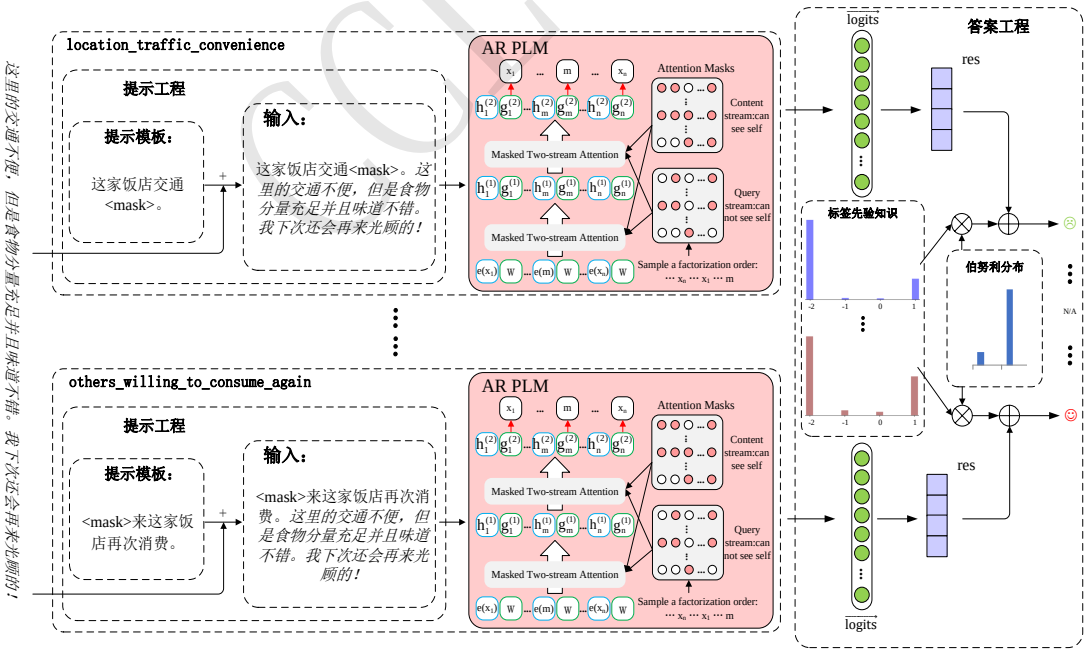


Figure 2: AP-LPK框架图

如，XLNet (Yang et al., 2019)）被引入来训练每个类别，以构建多类别提示训练（第 4.2节）。

此外，标签先验知识和伯努利分布的结合协助精炼提示训练的输出，以进一步获得最终预测（第 4.3节）。

4.2 自回归提示

自回归提示学习部分主要如图 2左侧所示，主要分为：提示模板工程与答案工程的设计和自回归预训练语言模型的引入。

4.2.1 提示模板工程与答案工程的设计

现今已有一些关于设计提示模板工程与答案工程的研究，其中设计方式主要包含手动设计和自动化设计 (Liu et al., 2021; Chen et al., 2022; Schick et al., 2020)。我们采用手动方式来设计提示学习中的提示模板工程和答案工程，这种方式策略直观而且已经被证明在提示任务上可以实现稳定的性能 (Schick and Schütze, 2021; Hu et al., 2021; Gao et al., 2021)。

对于每个方面类别，基于完形填空式的提示学习，手动设计的提示模板需要贴合评论文本的上下文表达，以进一步保证模型输入的语义质量。另外，提示模板还必须对应待测的特定方面类别信息，如图 2中所示的**提示模板**。同时，提示模板应尽量简洁，避免引入其他干扰信息。

由此，答案空间以及到输出空间的映射可根据提示模板和评论文本的上下文语境手动设计。在每个方面类别下，答案空间中的标签词是与情感极性对应的具有正确语义的中文字词，从而也构建起从答案空间到输出空间的映射。当这些字词应用到模板中的掩码位置（即，标记<**mask**>）时，它们可以和对应的提示模板构成具有合理语义的评论文本。

提示模板拼接在每个类别的每段评论文本的开头，以生成模型输入，如图 2中所示的**输入**，其中标记<**mask**>是模型的预测对象。

4.2.2 自回归预训练语言模型的引入

基于完形填空式的提示学习一般采用Masked LM作为预训练模型，比如BERT、RoBERTa (Schick and Schütze, 2021; Hu et al., 2021)。BERT、RoBERTa作为自编码语言模型，对于上述**输入**，在训练过程中可以根据上下文，预测标记<**mask**>处被掩盖的字词。BERT等自编码语言模型虽然性能不错，但是对于被掩盖的字符的数目多于1且连续时，它的预测效果便无法保证。特别的，对于我们实验中基于中文字词的待测标签词（一般是由多个连续的中文字符组成，比如中性情感词“一般”等），这些模型还可能生成不合语义的答案词（比如“不般”等），从而干扰后续的标签映射。这是因为BERT等自编码语言模型对于被掩盖的多个字符是相互独立预测的，即非自回归生成。

因此，预训练中文XLNet (Cui et al., 2020)被选择作为我们提示学习的预训练语言模型，它将自回归语言模型和自编码语言模型的优点进行了巧妙的结合。具体来说，对于我们的提示训练，每个方面类别将上述的模型**输入**馈送给XLNetLMHeadModel[0]，并使用答案空间中对应的标签词作为训练标签。该模型在训练过程中使用的损失函数是交叉熵。

如前所述，自回归提示的提出是为了激发预训练语言模型的潜力，并生成具有更好语义一致性的标签词。例如，在图 1中，提示学习会使得下游任务去适应预训练语言模型。在这种情况下，预训练语言模型可以充分利用预训练时的数据信息，其中可能包含类似于图 1的描述；而微调需要预训练语言模型来适应下游任务，可能会导致这些信息的遗忘。此外，生成的标签词也为第4.3节奠定基础。

4.3 标签先验知识

动机：如前所述，本文提出使用每个类别的标签分布作为标签先验知识，来帮助减轻噪声标签的干扰。对于数据标注来说，尽管由于标注不当而产生了少量噪声标签，但数据集的整体标注准确率和习惯是趋于稳定。同样以图 1为例。数据集中具有相似描述的评论基本上被标记为类别*location_distance_from_business_district*上的积极情感极性，并且该知识可以反映在标签分布中，用以校准噪声标签的问题。

具体来说，对于每个类别，本文将训练集中四种情感极性（即消极、中性、积极、未提及）的频率统计为标签先验知识$\boldsymbol{K} \in R^4$。然而，该先验知识$\boldsymbol{K}$不一定是完全有益的。受BERT的启发，BERT在进行基于掩盖的语言模型任务时，使用伯努利分布来决定随机对输入序列中的某些位置进行遮罩 (Devlin et al., 2019)，因此这里伯努利分布被采用来决定$\boldsymbol{K}$可用的概率，该概率在BERT中为0.15，而在这里是可学习的。

[0]https://huggingface.co/docs/transformers/v4.17.0/en/model_doc/xlnet

基于自回归模型对标记<**mask**>的预测$\overrightarrow{logits}$，我们提取对应于四个情感极性的标签词的logits，以生成原始输出$\boldsymbol{res}$。因此，进行噪声标签干扰的缓解如下：

$$\boldsymbol{F} = bernoulli(\boldsymbol{K}) \bigoplus \boldsymbol{res} \tag{2}$$

其中，$\bigoplus$表示基于四种情感极性的对应相加。

最后，可以得到最终输出如下：

$$\hat{y}_i^I = argmax(softmax(\boldsymbol{F})) \tag{3}$$

4.4 算法描述

综上所述，AP-LPK的主要算法描述如算法 1所示。

在算法中，每个方面类别下的模型是相互独立的，因此本文对所有预定义方面类别进行循环（第2行）。在每个轮次的循环（第5行）中，算法首先通过训练数据集进行XLNet训练（第6行），并将本轮次训练所得的XLNet用于验证集并得到验证结果（第7行）。因为实验主要以Macro-F1值进行评估，因此判断验证结果中的F1值是否在本轮次中有所提升（第8，13行）。若验证结果中的F1值相比上一轮次有提升，则保存本轮次训练所得XLNet并利用该模型进行测试集和标签先验知识上的评估（第11，12行）；否则，若验证集上的F1值在3个epoch内没有进一步提升，当前方面类别的训练过程才结束（第14，15，16行）。

算法 1 AP-LPK算法

输入： 带有提示模板的训练集评论文本$train_input$，对应$train_input$的标签词$train_label_word$;
带有提示模板的验证集评论文本$valid_input$，对应$valid_input$的标签词$valid_label_word$;
带有提示模板的测试集评论文本$test_input$，对应$test_input$的标签词$test_label_word$;
标签先验知识lpk

输出： 评价指标($test\_result['macro\_f1']$)，每个方面类别的最终预测模型model

```
1: function AP-LPK
2:     for each category do
3:         max_macro_f1 ← −1.0
4:         early_stop ← 0
5:         for each epoch do
6:             training: XLNet(train_input, train_label_word)
7:             validating: valid_result['macro_f1'] ← XLNet(valid_input, valid_label_word)
8:             if valid_result['macro_f1'] > max_macro_f1 then
9:                 early_stop ← 0
10:                max_macro_f1 ← valid_result['macro_f1']
11:                saving model
12:                testing: test_result['macro_f1'] ← XLNet(test_input, test_label_word, lpk)
13:            else
14:                early_stop ← early_stop + 1
15:                if early_stop >= 3 then
16:                    break
17:                end if
18:            end if
19:        end for
20:    end for
21: end function
```

5 实验

5.1 数据集

本文的方法在5个数据集上进行评估，数据集相关的统计信息如表 1所示。 其中，4个基准数据集来自SemEval 2015和2016 (Pontiki et al., 2015; Pontiki et al., 2016)。REST-AI Challenger 2018为

	Restaurant-15	Laptop-15	Restaurant-16	Laptop-16	REST-AI Challenger 2018
训练集样本数	1102	1397	1680	2037	105k
测试集样本数	572	644	580	572	15k
预定义类别数	30	198	30	198	20

Table 1: 数据统计

中文数据集，即在线用户评论数据集的细粒度情感分析2018 (AI Challenger 2018)[1]。原REST-AI Challenger 2018虽然包含测试集A和测试集B，但由于原测试集无法下载且未被标注，因此本文将AI Challenger竞赛中的验证集作为本次实验的测试集。

REST-AI Challenger 2018在数据量上远大于4个基准数据集，因此本文选择它作为实验的主要数据集。REST-AI Challenger 2018中的评价对象按照粒度不同划分为两个层次，第一层为粗粒度的评价对象，例如评论文本中涉及的环境、价格等要素；第二层为细粒度的情感对象，例如“环境”属性中的“装修情况”、“嘈杂情况”等要素。具体情况如表 2所示。

粗粒度层面	细粒度层面
位置(location)	交通是否便利(traffic convenience)
	距离商圈远近(distance from business district)
	是否容易寻找(easy to find)
服务(service)	排队等候时间(wait time)
	服务人员态度(waiter's attitude)
	是否容易停车(parking convenience)
	点菜/上菜速度(serving speed)
价格(price)	价格水平(price level)
	性价比(cost-effective)
	折扣力度(discount)
环境(environment)	装修情况(decoration)
	嘈杂情况(noise)
	就餐空间(space)
	卫生情况(cleaness)
菜品(dish)	分量(portion)
	口感(taste)
	外观(look)
	推荐程度(recommendation)
其他(others)	本次消费感受(overall experience)
	再次消费的意愿(willing to consume again)

Table 2: 评价对象的具体划分

每个细粒度要素有4种情感倾向，如表 3所示。

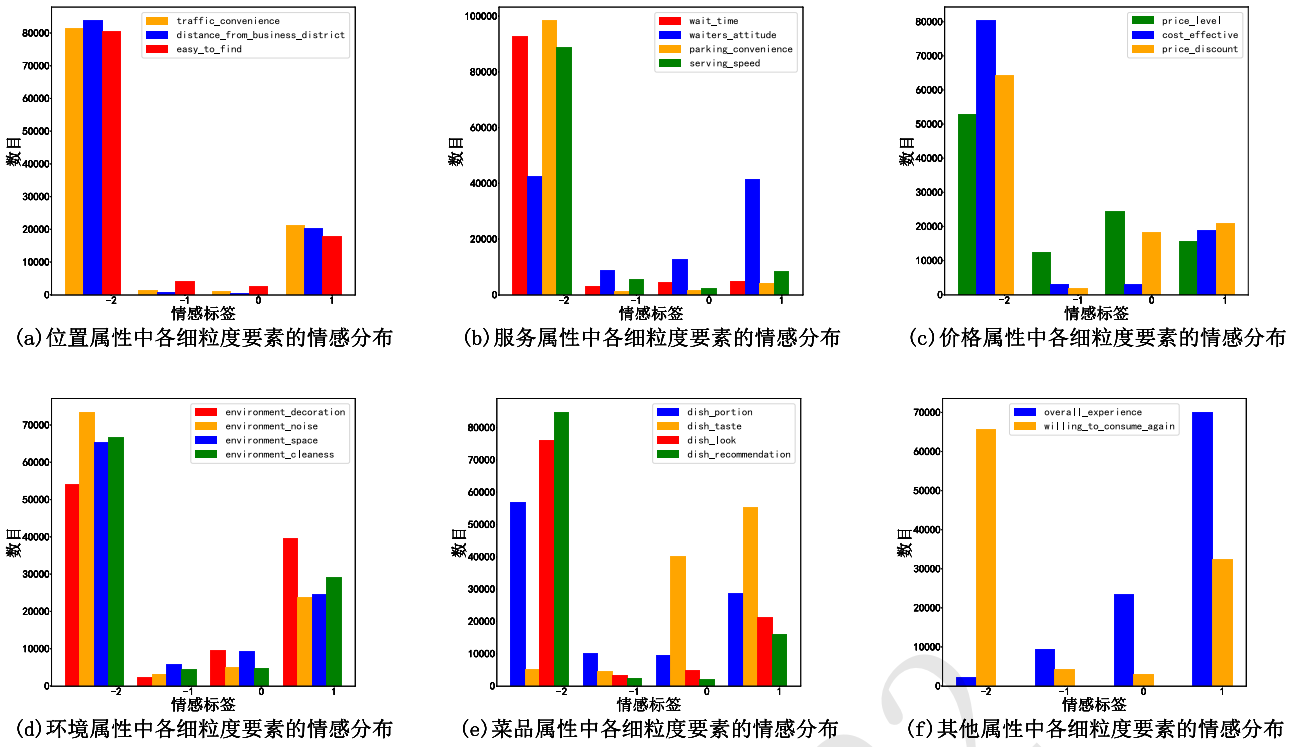
情感标签(labels)	-1	0	1	-2
情感倾向	消极情感(Negative)	中性情感(Neutral)	积极情感(Positive)	情感倾向未提及(Not mentioned)

Table 3: 情感倾向

为了直观了解数据情况、探究数据的内在特征，实验中统计了REST-AI Challenger 2018中六种粗粒度方面下每个细粒度要素在四种情感倾向上的评论文本数目分布，数据统计情况如图 3所示。

从统计结果中可以观察到REST-AI Challenger 2018的分布方式极不平衡，总体上最多的情感倾向为Not mentioned，并且消极和中性的情感倾向普遍相对偏少。这种数据失衡给模型的训练带来了一定挑战，但也启发了后续的标签先验知识的引入。

[1]https://challenger.ai/dataset/fsaouord2018.



(a)位置属性中各细粒度要素的情感分布 (b)服务属性中各细粒度要素的情感分布 (c)价格属性中各细粒度要素的情感分布

(d)环境属性中各细粒度要素的情感分布 (e)菜品属性中各细粒度要素的情感分布 (f)其他属性中各细粒度要素的情感分布

Figure 3: 六种方面的数据统计情况

5.2 实验设置

对于REST-AI Challenger 2018：原训练集随机以9:1的比例分成训练集、验证集，并且构建的提示模板、答案空间及到输出空间的映射、标签先验知识均以字典形式引入。XLNetLMHeadModel加载了Chinese-XLNet-base的预训练权重，具体设置请参考相应网址[2]。

实验采用AdamW优化器，其学习率和批量大小（batch size）分别设置为1e-5和6。最大文本长度设置为512，通过统计分析训练集中的大部分文本长度主要集中在500附近及其之前的区间；另外，长度在250附近区间的文本最多，并随着文本长度的增加，分布逐渐减小，到文本长度超过1000时，如此长度的文本在数据集中就更稀少了。最后，直到模型在验证集上的性能在3个epoch内没有变化，训练过程才结束。

准确率（Precision）、召回率（Recall）和Macro-F1值是评估指标，且Macro-F1值是实验的主要比较指标。在实验中，我们分别计算每个类别下的三种评估指标，并取所有类别的三种指标均值作为最终的评估结果。

对于四个基准数据集：XLNet使用xlnet-base-cased的预训练权重，具体设置请参考相应网址[3]。其他实验设置参照Cai等人 (2020)的工作。

5.3 基线模型

AddOneDim-LSTM (Schmitt et al., 2018)、Hier-GCN-BERT (Cai et al., 2020)是最近关于ACSA任务的两项工作，并被选为我们的主要基线。

AddOneDim-LSTM: 在标签扩维的基础上，词嵌入模型FastText被用来进行评论文本的词嵌入工作，然后通过双向LSTM进行编码，最后编码结果被用于多个分类器，其中分类器的个数取决于给定方面类别的数目，且分类器之间相互独立。为了得到较好的结果，模型参数也进行了以下调整：词嵌入维度设置为100，双向LSTM隐层的维度设置为300，dropout和学习率设为0.5和0.001。另外，训练优化器采用Adam算法，损失函数为交叉熵。

Hier-GCN-BERT: 近年来最先进的工作之一。模型利用BERT进行特征提取，捕获全局情感，并通过多头自注意力进行方面类别的表示；而后，基于类别之间的关联性和类别与情感之间的关联性，图卷积网络被用来建模其中的关系；最后，在类别-情感层次预测结构的基础上，方面类别提取和情

[2]https://huggingface.co/hfl/chinese-xlnet-base
[3]https://huggingface.co/xlnet-base-cased

感分类分别进行，并且方面类别的高优先级被用来进行类似剪枝的操作，以改善模型性能。

全局上，所有基线模型都与AP-LPK的算法流程保持基本一致，最大文本长度也都设置为512。其他未提及的参数，均与模型原始论文中的参数保持基本一致。因此，实验通过复现以上两项工作，得到在REST-AI Challenger 2018数据集上的实验结果，与我们提出的AP-LPK方法进行对比。

5.4 实验结果

本部分将从总体分析、消融分析以及案例分析对实验结果进行讨论。实验结果见表 4、表 5。

Method	P	R	F1
AddOneDim-LSTM	67.73	65.28	65.88
Hier-GCN-BERT	70.83	68.04	69.04
AP-LPK(**Ours**)	72.74	70.15	**71.00**
w/o Label Prior	72.58	70.24	70.77
w/o Label Prior & AutoRegression	74.14	68.38	69.48

Table 4: 在REST-AI Challenger 2018上的实验结果(%)

Method	Restaurant-15			Laptop-15			Restaurant-16			Laptop-16		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AddOneDim-LSTM	54.33	28.44	37.32	-	-	-	61.56	42.82	50.50	-	-	-
Hier-GCN-BERT	71.93	58.03	64.23	71.90	54.73	62.13	76.37	72.83	74.55	61.43	48.42	54.15
AAGCN-BERT	-	-	71.75	-	-	72.39	-	-	80.77	-	-	69.68
AP-LPK(**Ours**)	77.70	77.28	**77.33**	83.03	82.76	**82.72**	77.26	77.60	77.32	80.33	80.22	**80.10**

Table 5: 在四个基准数据集上的主要结果(%)

总体分析：从表 4中可知，在REST-AI Challenger 2018数据集上，AddOneDim-LSTM的Macro-F1值为0.6588。基于类别-情感层次预测结构的Hier-GCN-BERT的Macro-F1值可以达到0.6904。这也论证了Cai等人 (2020)的工作。而相比于通过微调训练的方法，我们提出的带有自回归提示训练和标签先验知识的方法可以达到0.7100的Macro-F1值。这表明我们的提示建模在这个任务中比微调有更好的性能改进，以及标签先验知识的有效性。

在表 5中，AddOneDim-LSTM、Hier-GCN-BERT的结果取自Cai等人 (2020)，AAGCN-BERT的结果取自Liang等人 (2021)。我们的方法AP-LPK在四个基准数据集上的F1值分别可以达到0.7733、0.8272、0.7732、0.8010。其中，因为Restaurant-16预定义的类别较为宽泛，比如类别“FOOD#QUALITY”会涉及更细粒度的方面类别（例如，口感、新鲜度、质地、温度等），提示模板的设计很难精确，由此影响了提示学习的效果，因而我们提出的AP-LPK在F1值方面低于AAGCN-BERT。针对这样的问题，如何进行更合适的提示模板工程，也是我们后续研究、改进的重要内容。

消融分析：在表 4中，我们进一步展示了AP-LPK消融研究的结果。在无标签先验知识的情况下，F1值降低到0.7077。对于无标签先验和自回归的情况，我们将预训练语言模型XLNet替换为BERT (Devlin et al., 2019)，以完成提示训练。它的设置与我们的自回归提示中的设置保持一致。它的F1值从0.71降低到0.6948。从中可见自回归模型在生成高质量标签词方面的能力更强。

此外，在无标签先验和自回归的情况中，基于非自回归的BERT会出现未知、错误标签词的生成。为了解决这个问题，我们使用这些标签词的最后一个字符作为映射参考，以在预定义的字典中寻找最可能的一个答案。由于字典由类别标签组成，因此大量的占比较大的情感极性的映射可以被正确获得，例如情感极性“未提及”。因此，这种情况下平均精度更高，但召回率要低得多。同时，即使使用我们手动添加的字典，在非自回归条件下的Macro-F1值仍然是最低的。

案例分析：如表 6所示， Review 1和Review 2关于装修风格的描述是相似的，但相应的情感标签不同。通过阅读它们，我们可以很容易地知道Review 2在类别*environment_decoration*上的情感标签是训练集中的噪声。而我们的方法AP-LPK通过利用标签先验知识，可以正确预测Review 1关于该方面类别的情感极性为积极。

同时，在图 1中提到的Review 3为训练集带来了噪声标签，其在类别*location_distance_from_business_district*上的真实情感标签应该是积极的。当使用Review 3 作

Review text	Category	Label	w/o Label Prior	AP-LPK
No.1 金殿水库边，靠近云南飞虎队博物馆，位置算很好找的，在一个院子里，整个装修风格很民族风，算是有特点的店.....(**from testing set**)	*environment_decoration*	positive	neutral	positive
No.2 早就听说了这家店，今天在凯德广场转，于是就来尝尝。店铺的装修风格很工厂感，黑色的铁丝网和灰色的墙壁，感觉特别.....(**from training set**)	*environment_decoration*	neutral	-	-
No.3 .....这家店在厦门SM广场二期的三楼店算是不错的，因为环境可以比较新，餐具也比较干净，位置也比较适中.....(**from training set**)	*location_distance_from_business_district*	neutral	-	positive

Table 6: 关于AP-LPK的案例分析

为测试样本时，我们的方法仍然可以正确识别它，这意味着我们的方法可以用来修正训练数据集中的那些噪声标签。这将是未来一项有趣的工作。

另外，从表 7中， 我们也可以直观验证实验中自回归生成与非自回归生成的研究。Review 4在测

Review text	Category	Label	w/o Label Prior & AutoRegression	w/o Label Prior
No.4感谢大众点评，感谢又让我中试吃.....环境，简单干净正经，最值得一题的就是无论你坐在哪里都不会离菜品特别远，不像凯德，每次拿东西都要走十万八千里！！值得表扬(〉ˆωˆ〈).....(**from testing set**)	*environment_decoration*	positive	neutral	positive

Table 7: 关于自回归生成与非自回归生成的案例分析

试集中关于类别*environment_decoration*的真实标签为积极的，从主观分析也可以确认其标注的合理性。而两个消融实验中，只有无标签先验和自回归的实验无法正确预测Review 4在这一方面类别上的情感倾向。这直观地反映了自回归生成的引入对实验是有积极作用的。

通过以上分析，这些案例表明：在ACSA任务中自回归提示和标签先验知识的引入是有效的。

6 结束语

本文着重于缓解基于方面类别的情感分析中噪声标签对分类质量的影响问题，提出了一种自回归提示训练的生成式情感分析方法，从而生成具有更好语义一致性的标签词，并通过伯努利分布引入标签先验知识，以减轻噪声标签的干扰。在五个数据集上的实验结果表明，本文的方法在F1值方面优于最先进的方法。今后的研究将在提示模板的学习上考虑引入连续模式，增加情感分离方法的质量。

参考文献

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics2020*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022*, pages 2778–2788, Virtual Event, Lyon, France. ACM.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained

models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Zehui Dai, Wei Dai, Zhenhua Liu, Fengyun Rao, Huajie Chen, Guangpeng Zhang, Yadong Ding, and Jiyang Liu. 2019. Multi-task multi-head attention memory network for fine-grained sentiment analysis. In *Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing*, pages 609–620, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yujie Fu, Jian Liao, Yang Li, Suge Wang, Deyu Li, and Xiaoli Li. 2021. Multiple perspective attention based on double bilstm for aspect and sentiment pair extract. *Neurocomputing*, 438:302–311.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Xin Guo, Geng Zhang, Suge Wang, and Qian Chen. 2020. Multi-way matching based fine-grained sentiment analysis for user reviews. *Neural Computing and Applications*, 32(10):5409–5423.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4601–4610, Hong Kong, China. Association for Computational Linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. Computation and Language, *arXiv preprint arXiv:2108.02035*. Version 2.

Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. 2021. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346.

Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 208–218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. Computation and Language, *arXiv preprint arXiv:2107.13586*. Version 1.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. Computation and Language, *arXiv preprint arXiv:2103.10385*. Version 1.

Weiran Pan, Wei Wei, and Feida Zhu. 2022. Automatic noisy label correction for fine-grained entity typing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4317–4323. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.