frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 07 March 2024
DOI 10.3389/fphy.2024.1361695

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Guodong Du,
Yanshan University, China
Peng Gui,
Wuhan University, China

*CORRESPONDENCE
Yan Xiang,
sharonxiang@126.com

RECEIVED 26 December 2023
ACCEPTED 20 February 2024
PUBLISHED 07 March 2024

CITATION
Xian Y, Qin Y and Xiang Y (2024), Auto-verbalizer filtering for prompt-based aspect category detection.
Front. Phys. 12:1361695.
doi: 10.3389/fphy.2024.1361695

Auto-verbalizer filtering for prompt-based aspect category detection

Yantuan Xian[1,2], Yuan Qin[1,2] and Yan Xiang[1,2]*

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China

Aspect category detection (ACD) is a basic task in sentiment analysis that aims to identify the specific aspect categories discussed in reviews. In the case of limited label resources, prompt-based models have shown promise in few-shot ACD. However, their current limitations lie in the manual selection or reliance on external knowledge for obtaining the verbalizer, a critical component of prompt learning that maps predicted words to final categories. To solve these issues, we propose an ACD method to automatically build the verbalizer in prompt learning. Our approach leverages the semantic expansion of category labels as prompts to automatically acquire initial verbalizer tokens. Additionally, we introduce an indicator mechanism for auto-verbalizer filtering to obtain reasonable verbalizer words and improve the predicting aspect category reliability of the method. In zero-shot task, our model exhibits an average performance improvement of 7.5% over the second-best model across four ACD datasets. For the other three few-shot tasks, the average performance improvement over the second-best model is approximately 2%. Notably, our method demonstrates effectiveness, particularly in handling general or miscellaneous category aspects.

KEYWORDS

aspect category detection, prompt learning, few-shot, verbalizer, sentiment analysis

1 Introduction

Aspect category detection (ACD) is a subtask of sentiment analysis that aims to detect the categories contained in reviews from a predefined set of aspect categories. For example, the sentence "Nevertheless the food itself is pretty good" contains the aspect category "Food," and the sentence "But the staff was so horrible to us" contains the aspect category "Service." Most of the existing excellent methods [1–3] finetune the pre-trained language models to solve ACD tasks, and their effects largely depend on the size of labeled data. However, as online reviews are updated quickly, the aspect categories will also be updated. It is difficult to provide sufficient label data for newly emerging categories. Therefore, the performance of the above methods will drop significantly when there are only few labeled samples.

In order to stimulate pre-trained language models (PLMs) to exhibit a greater performance under the conditions of few-shot and zero-shot, the researchers were inspired by GPT-3 [4] and LAMA [5] and proposed to use prompt to convert the classification task into a cloze task, which unified the downstream task and PLMs into the same schema to maximize the use of prior knowledge of PLMs. Prompt learning obtains the probability of each token filled in the [MASK] position in the PLM vocabulary through the prompt and then uses the verbalizer to map it to the final category. As one of the

TABLE 1 Examples of verbalizer words in the "miscellaneous" category. Bold indicates that it appears in the other categories, and "xx" indicates that it does not appear in the PLM vocabulary.

Method	Prompt words
Manual	Miscellaneous,. . .
Search based	Bryan, anonymous, Wes, noise, LM, KH, Ethan, Wayne, dark, iii, YOU. . .
KBs	Sundry, assorted, heterogeneous, multifarious, extraneous, mixed. . .

important components of prompt learning, the verbalizer contains the mapping relationship between prompt words and the final aspect category. Therefore, constructing a high-quality prompt word set can greatly improve the performance of the verbalizer.

The current methods of constructing prompt word sets can be roughly divided into three types: manual construction [6], search based [7], and continuously learnable [8]. Table 1 shows the prompt words selected for the "miscellaneous" category of the restaurant dataset by different methods. It can be seen that the main problems are as follows: 1) These methods either require intensive manual work or require the support of external knowledge bases and labeled data and, thus, cannot handle zero-shot tasks at a small cost. At the same time, many words searched from external knowledge will not appear in the PLM vocabulary. For example, for the words highlighted in red in the third row of Table 1, statistics show that 11 of the first 50 prompt words obtained for this category cannot be recognized by PLMs. This is because the vocabulary of the external knowledge base is different from the PLMs, due to which the overlap between the two is lacking. 2) The manually constructed prompt word sets only contain category words themselves, so the diversity of prompt words is not enough. For example, the first row in Table 1 only contains the word "miscellaneous" itself. Search-based methods do not consider the specificity of prompt words, where a word may appear in different word sets. For example, words such as "anonymous" in the second row of Table 1 also appear in prompt word sets of other categories at the same time. In addition, these methods do not consider the characteristics of the ACD task, such as categories are basically represented by nouns.

In response to the first type of problems mentioned above, we propose to use the semantic expansions of category labels as prompts to directly search for the initial prompt words from the internal vocabulary of PLMs so that the prompt words in the verbalizer conform to the PLM vocabulary. For the second type of problem, we propose a filtering mechanism to select prompt words. Specifically, we first consider the task characteristics; that is, the ACD task is to detect predefined aspect categories contained in sentences which should be represented by words with actual meaning. Therefore, we start from the parts of speech and select nouns, verbs, and adjectives. Second, we consider diversity and select words with high semantic similarity to the category. Finally, in terms of specificity, choosing words that are much more similar to the category to which they belong than to other categories as prompt words can avoid confusion in the mapping process. The main contributions of this article are as follows:

1) Auto-verbalizer filtering methods are proposed for prompt-based aspect category detection, which alleviates the limitations of the detection performance caused by unreasonable verbalizer design in existing prompt-based ACD methods.
2) The semantic extension of category labels is used as prompts to construct an initial verbalizer and eliminate dependence on labeled data and external knowledge bases. At the same time, an automatic filtering mechanism is introduced for the verbalizer to select prompt words related to aspect categories.
3) Experiments show that the proposed method can achieve optimal performance under zero-shot and few-shot conditions compared with existing prompt-based learning methods.

2 Related work

2.1 Aspect category detection

Semeval proposed the ACD subtask in 2014. Under the condition of sufficient labeled data, most of the previous ACD methods are based on machine learning, such as the classic SVM [9] and maximum entropy [7,10] which handcrafts multiple features such as n-grams and lexical features to train a set of classification devices. In recent years, methods based on deep neural networks [2] have been widely adopted. In [11], the output of CNN training as a type of feature and other POS tags and other features was sent into the one-vs-all classifier. The one-vs-all classifier used in [3] consists of a set of CNN network layers above the LSTM layer, which implements aspect category detection and aspect term extraction in parallel.

2.2 Prompt verbalizer construction

In the case of insufficient labeled data, researchers detect categories by mining association rules [12] or calculating word co-occurrence frequencies [13], but this requires obtaining reasonable rules in advance. Since the release of GPT-3, prompt learning has provided new ideas for ACD when labels are insufficient. The way of using prompts to stimulate internal knowledge of PLMs and avoiding the introduction of a large number of parameters to be trained usually includes two important parts: templates and a verbalizer. According to the manually created cloze template provided by the LAMA dataset, the previous templates are all artificially created auxiliary sentences which are human-understandable. For example, manually designed prefix-type prompts [4] had achieved good results in some NLP tasks, such as text question answering and neural machine translation. However, although this type of template has the advantage of being intuitive, it requires a lot of experience and a lot of time to obtain good performance prompts and cannot be optimized to the best. To solve these problems, automated template-based methods are proposed [14–17], which automatically search for natural language phrases in discrete space to form prompt templates. Later, scholars discovered that the prompts were constructed to allow PLMs to better understand the task rather

than humans. Therefore, they proposed that templates do not need to be limited to human understandability. In [18–21], continuous templates were directly constructed in the model embedding space. The template is no longer restricted by additional parameters and can itself be trained and optimized along with downstream tasks.

When working on templates, researchers are also focusing on exploring another important component of prompt learning—the construction of the verbalizer. The most straightforward method is to use manually selected words to construct prompt word sets, and it has been proven to be effective [7]. However, this type of method involves personal biases, so the coverage of the vocabulary is insufficient. Based on these problems, some automatically search-based methods have been proposed. The work in [22] searched for label words in the pruned candidate space and redefined the k classification problem as a binary classification problem of "1 vs. k-1" so that PLMs can distinguish category y from other categories. In [21], a two-stage gradient-based automatic search method was used to calculate the representation of each category in the first stage and train a classifier. The second stage uses this classifier to select words that are close to the category representation to construct a verbalizer. In [23–25], relevant words were selected from the external knowledge base and then refined to align with the PLM vocabulary. However, such automatically search-based methods require the assistance of sufficient training data or external knowledge. In contrast to the discrete verbalizer, the continuous verbalizer [8,20] represents categories in word embedding space and can be trained and optimized. In [8], vector form was used to represent categories, carry out a dot product between the token vector predicted by PLMs and the category vector, and select the corresponding category that obtains the maximum dot product as the prediction result. In [26], the filled-in token vectors of all sentences under each category were averaged to obtain the prototype representation of this category, and this prototype was continuously optimized. Similarly, continuous vectors also require a large amount of data for training and optimization, so they cannot be directly applied to zero-shot learning.

3 Prompt-based aspect category detection with auto-verbalizer filtering

3.1 Task definition

ACD is to identify aspect categories $y \in \{1, 2, \ldots C\}$ for a given sentence, where C is the number of aspect categories. The basic process of prompt-based ACD is as formula (1): the ith sentence x_i is packed into x_i^p with a template, which is a natural language text with the "[MASK]" token:

$$x_i^p\text{: } x_i \text{ [sep] It is about [MASK] category.} \tag{1}$$

We obtain the probability p ([MASK]= $v|x_i^p$) of each token v in the vocabulary $V \in R^D$ filling in the [MASK] position by PLMs. The probability distribution vector of the entire vocabulary for the ith sentence is $P_i^V \in R^{1\times D}$. Finally, the probability of category y can be calculated as formula (2)

$$p\left([\text{MASK}] = v|x_i^p\right) = f\left(p\left([\text{MASK}] = v|x_i^p\right)|v \in V_p\right), \tag{2}$$

where V_p is the prompt word set of the verbalizer and f is a function transforming the probability of prompt words into the probability of the category.

3.2 Initial construction of the verbalizer based on label semantic extension

When evaluating aspect categories of reviews, the most important consideration is the semantic similarity between the review and the label categories [27–29]. Consequently, the specific category label itself serves as valuable prior knowledge that can be utilized. Following this idea, we propose to utilize category labels as prompts to construct the verbalizer.

Specifically, as shown in Figure 1A, we use task-specific templates such as "[x]. This is about the [MASK] category," where [x] is the definition statement of the corresponding category label j in Wikipedia (see Table 2). The definition statement is encapsulated into a natural language text x_j^c with [MASK] tokens and is sent to PLMs to obtain the probability that each token in the vocabulary V is filled to the [MASK] position. In this way, the probability distribution vector $P_j^V \in R^{1\times D}$ for a given label category j can be obtained. As shown in Figure 1B, this is carried out for different label categories, and a complete verbalizer initial probability matrix $P \in R^{C\times D}$ is constructed.

3.3 Indicator mechanism for verbalizer filtering

We propose an indicator-based filtering mechanism to improve the verbalizer. Specifically, we set an indicator value b_{ji} for each probability p_{ji} in the probability matrix P representing the correlation of token i with a specific category j. A value of 1 signifies that the token is highly important for the corresponding category, whereas a value of 0 signifies the opposite. Initially, all indicator values are set to 1, forming the indicator matrix $B \in R^{C\times D}$. Next, as shown in Figure 1C, we refine the indicator matrix to obtain more reasonable prompt words by considering three parts.

(1) In order to be more consistent with the characteristics of the ACD task, we use the pos_tag method from the nltk package to define the set of tokens in the vocabulary V that match nouns, verbs, and adjectives as $\{pos\}$ and then adjust the corresponding element values in the indicator matrix B to get a new indicator matrix B^{pos} according to formula (3):

$$b_{ji}^{pos} = \begin{cases} b_{ji} & if\ v_i \in \{pos\} \\ 0 & else \end{cases} \tag{3}$$

(2) In order to retain prompt words with more highly semantic similarity to a specific category, we further modify the element values in the matrix B^{pos} based on category semantic relevance and obtain B^{sem} according to formula (4):

$$b_{ji}^{sem} = \begin{cases} b_{ji}^{pos} & if\ p_{ji} > MAX\_M\left(P_j^V\right) \\ 0 & else \end{cases}, \tag{4}$$

10.3389/fphy.2024.1361695

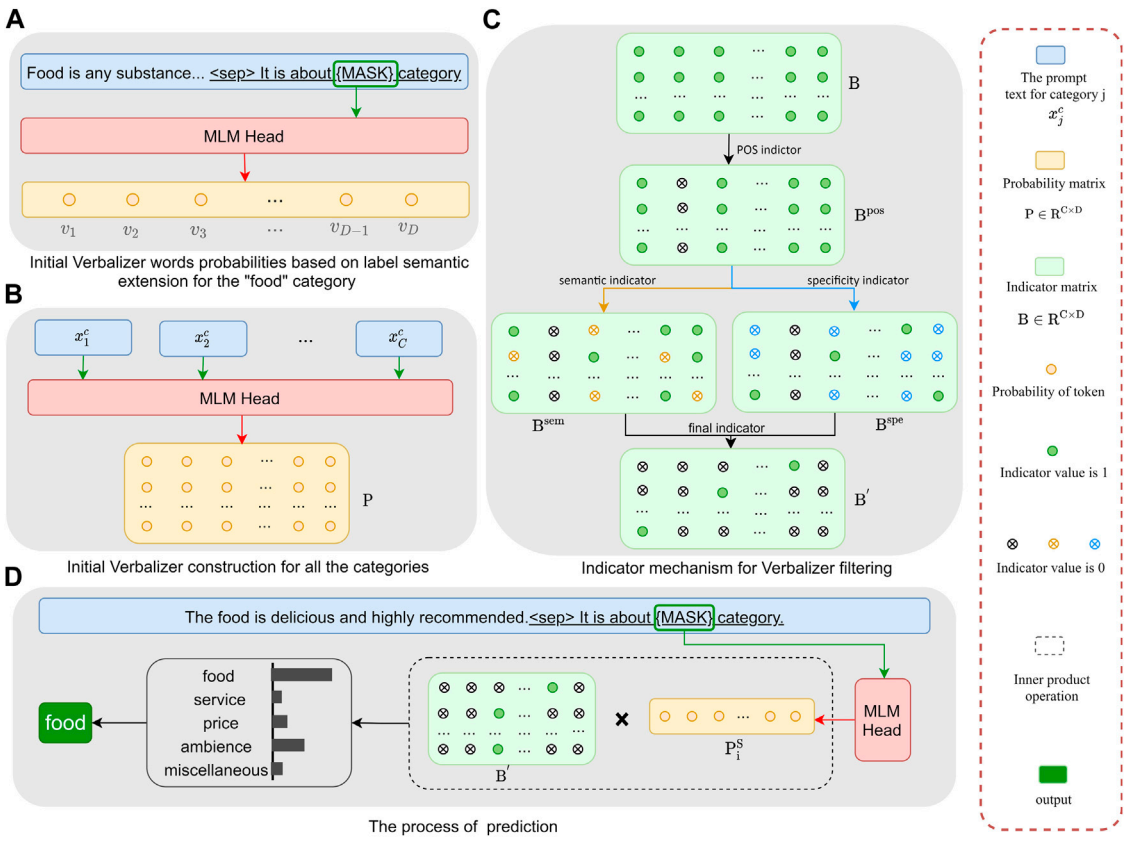

FIGURE 1
Illustration of our model. **(A)** Initial Verbalizer words probabilities based on label semantic extension for the "food" category. **(B)** Initial Verbalizer construction for all the categories. **(C)** Indicator mechanism for Verbalizer filtering. **(D)** The process of prediction.

TABLE 2 Semantic extensions of categories. The semantic extensions are derived from Wikipedia or Baidu Encyclopedia. We take the first one or two sentences of the definition as the semantic extensions.

Label	Semantic extension
Food	Food is any substance consumed by an organism for nutritional support
Service	Customer service refers to the provision of assistance to customers or clients
Price	Price is the quantity of payment or compensation given by one party to another in return for goods or services
Ambience	Ambience which is also known as atmospheres or backgrounds
Miscellaneous	Miscellaneous refers to a collection of writings on various subjects or topics
Comfort	Comfort is the physical and psychological sense of ease
Size	Clothing size in general is the magnitude or dimensions of a thing
Quality	Quality is a product or service free of deficiencies
Layout	Keyboard layout is an arrangement of the keys on a typographic keyboard
Connection	Connection refers to a communication link between two or more devices
Service	Customer service is the assistance and advice provided by a company to those people who buy or use its products or services
Image	A digital image is an image composed of picture elements which is also known as pixels
Sound	The sound is the loudness of the sound and the characteristics of the timbre

where $MAX\_M(.)$ represents the Mth largest probability value in the probability distribution vector of the label category.

(3) In order to select the prompt words with specificity, we adjust the element values in the matrix B^{pos} based on the following formula to obtain the updated indicator matrix B^{spe} according to formula (5):

$$b_{ji}^{spe} = \begin{cases} b_{ji}^{pos} & if \frac{p_{ji}}{\sum_{j=1}^{C} p_{ji}} > \alpha \\ 0 & else \end{cases}, \tag{5}$$

where α is a threshold indicating that the words exceeding this threshold are class-specific.

Also, the modified matrix B' is calculated as formula (6)

$$B' = B^{sem} \circ B^{spe}, \tag{6}$$

where $\circ$ represents the Hadamard product.

Finally, the prompt words of each category are composed of tokens whose indicator value is 1 in the matrix B' under this category.

3.4 Aspect category prediction

During category prediction, we package the review x_i into a natural language text like in Figure 1D and send it to PLMs to obtain a probability distribution vector $P_i^S \in R^{1\times D}$ and finally map it to the aspect category label by the constructed verbalizer.

For the zero-shot scenario, we assume that all prompt words in the verbalizer contribute equally to the prediction of the corresponding category, so we calculate the category probability $\widehat{Y_{ij}}$ of the sentence x_i with respect to category j using the following formula (7):

$$\widehat{Y_{ij}} = P_i^S \left(B_j'\right)^T. \tag{7}$$

For few-shot scenario, we set a weight parameter for each token, and the probability $\widehat{Y_{ij}}$ of the sentence x_i with respect to category j is calculated as formula (8)

$$\widehat{Y_{ij}} = \left(P_i^S \circ W\right)\left(B_j'\right)^T, \tag{8}$$

where $W \in R^{1\times D}$ is the parameter vector to be trained, which can be optimized using the cross-entropy loss as formula (9). The objective function is the loss between the final predicted label and the true label:

$$loss = -\frac{1}{C} \sum_{i\in|D_{train}|} \sum_{j\in C} \hat{y} log p\left(y_j|x_i\right), \tag{9}$$

where $\hat{y}$ is the true label of input x_i.

4 Experiment

4.1 Datasets

We conducted experiments on four ACD datasets, including Restaurant-2014, Boots, Keyboards, and TV of the Amazon dataset. In the few-shot experiment, following most few-shot learning settings, we adopt the N way K shot mode, randomly selecting K samples of each category for the validation set and the training set, and the remaining samples are used as the test set. The size of the training set and validation set are $|D_{dev}|=|D_{train}|=N * K$.

4.2 Baselines

We selected several advanced models for comparative experiments. Same as this model, all prompt learning methods adopt the most basic prompt learning method: templates were used to convert the input into a natural language text with the [MASK] token, and the vocabulary token probability output by the model is mapped to class labels by the verbalizer. All models use the same template, so only the verbalizer is constructed differently.

Finetuning: The traditional finetuning methods add a classification layer after the PLM model, obtaining the hidden vector of [CLS] and making predictions via the classification layer.

Manual: The manually constructed verbalizer contains limited category prompt words. In this experiment, we use the category word itself to represent the only prompt word of this category.

WARP [8]: The model uses continuous vectors instead of discrete words to represent the categories. The output of the [MASK] position also obtains its hidden vector, and the two calculate the probability of belonging to different categories through the dot product. In the experiment, we use the word embedding of the category word as the initialization of the category vector.

PETAL [22]: The model uses labeled data and unlabeled data to automatically search for prompt words from PLM pruned vocabularies. By maximizing the likelihood function, it ultimately prefers to select words with higher frequency.

Auto-L [17]: The model sequentially prunes the search space through the initial probability distribution of the vocabulary and maximizing the accuracy in the zero-shot task and finally uses reordering to search for the best top n prompt words on the validation set. We fixed the automatic template generation part of the model and only use the construction part of the verbalizer.

KPT [23]: This method expands the verbalizer with the help of external knowledge and then refines the selected prompt words in various ways on the support set.

4.3 Experiment settings

The PLMs in the model adopt RoBERTa large. For zero-shot experiments, since there are no trainable parameters, we use the results of one experiment as the experimental data. For few-shot experiments, we use five different seeds to randomly select data, and the final experimental data are obtained by averaging the results from these five experiments. This setting ensures that the experimental findings are not overly influenced by a specific random initialization and provide a more robust and reliable assessment of the model's performance. Macro F1 is used as the test indicator in the experiment.

4.4 Main results

Table 3 contains all the experimental results on the four datasets, where AVG represents the average performance of each model of the

TABLE 3 Macro F1 (%) of different models on the four datasets.

K	Dataset	Finetuning	Manual	WARP	PETAL	Auto-L	KPT	Ours
0-shot	Restaurant	5.4	28.8	–	–	–	38.1	**74.4**
	Boots	15.9	32.4	–	–	–	28.8	**34.7**
	Keyboards	11.3	22.7	–	–	–	20.4	**25.1**
	TV	3.4	18.7	–	–	–	16.3	**26.2**
	AVG	9.0	25.7	–	–	–	25.9	**33.4**
5-shot	Restaurant	40.3	67.6	70.9	63.2	71.2	67.9	**73.5**
	Boots	23.7	55.4	60.1	48.6	57.2	55.6	**60.9**
	Keyboards	22.2	39.6	39.7	42.8	44.3	40.1	**45.4**
	TV	25.9	47.9	44.1	46.3	49.7	47.8	**51.2**
	AVG	28.0	52.6	53.7	50.2	55.6	52.9	**57.8**
10-shot	Restaurant	66.5	70.3	72.2	76.5	78.0	77.3	**78.8**
	Boots	43.2	61.6	60.2	48.4	58.3	66.1	**67.2**
	Keyboards	30.2	49.6	**51.4**	43.2	45.6	46.3	51.3
	TV	43.7	48.6	47.5	46.8	50.6	49.2	**52.6**
	AVG	45.9	57.5	57.8	53.7	58.1	59.7	**62.5**
20-shot	Restaurant	78.4	79.2	76.6	80.3	80.6	81.2	**82.8**
	Boots	55.7	68.3	65.3	64.4	64.3	65.2	**69.2**
	Keyboards	44.1	60.1	58.8	56.2	56.5	57.4	**60.9**
	TV	51.9	50.9	52.2	50.1	51.8	51.1	**53.5**
	AVG	57.5	64.6	63.2	62.8	63.3	63.7	**66.6**

Bold values indicate the best performance.

four datasets and bold represents the optimal performance. As shown in the table, our model achieves almost the best results under all settings. Compared with the second-best model, our model increased by 9.3%, 5.9%, 4.7%, and 9.9%, respectively, on the four datasets under the zero-shot setting, and the growth rate was particularly obvious. It shows that the prompt words searched from the PLM vocabulary using our method can better represent the category labels. Under different K values of the few-shot task, our method maintains a certain degree of performance growth in different field datasets which indicates that our model has a certain degree of generalization. Using the average performance AVG for comparison, our model increased by 2.2%, 2.8%, and 2.0%, respectively, under different K-shots compared to the second-best model. This shows that introducing weights for each prompt word and further training are beneficial to the optimization of the mapping process.

When further comparing different prompt learning methods, it can be found that our model almost achieved the best results under all K value settings, which proved the effectiveness of the design of this method. When the K value is small, the effect of the PETAL model is lower. According to the construction method of the verbalizer, it is speculated that PETAL needs training data to search for prompt words. So, when the labeled data are less, the deviation of the searched prompt words is greater. Auto-L may not consider the word confusion problem, so the effect is still lacking. As the K value continues to increase, KPT becomes the best model among all baselines, proving that the model requires training data to reduce the impact of noise words to a certain extent.

In addition, it is observed that the finetuning method is lower than all cue learning models in both zero-shot and few-shot tasks, so prompt learning is an advantageous method when there is less labeled data. However, as the training data increases, that is, as the K value increases, the gap between the two results decreases. It can be speculated that when the K value increases to a certain value, the finetuning method will still show comparable results.

4.5 Ablation study

To evaluate the impact of some designs in the model on the final performance, we conduct ablation experiments. We tested the influence of the three parts of the indicating filtering mechanism on the four datasets, respectively, and the results are shown in Figure 2. "w/o pos," "w/o spe," and "w/o sem" mean not to use B^{pos}, B^{spe}, and B^{sem}, respectively for verbalizer filtering.

Compared with the complete model, the significant decrease in experimental results of three ablation models illustrates that these three parts of the indicator mechanism can greatly ensure

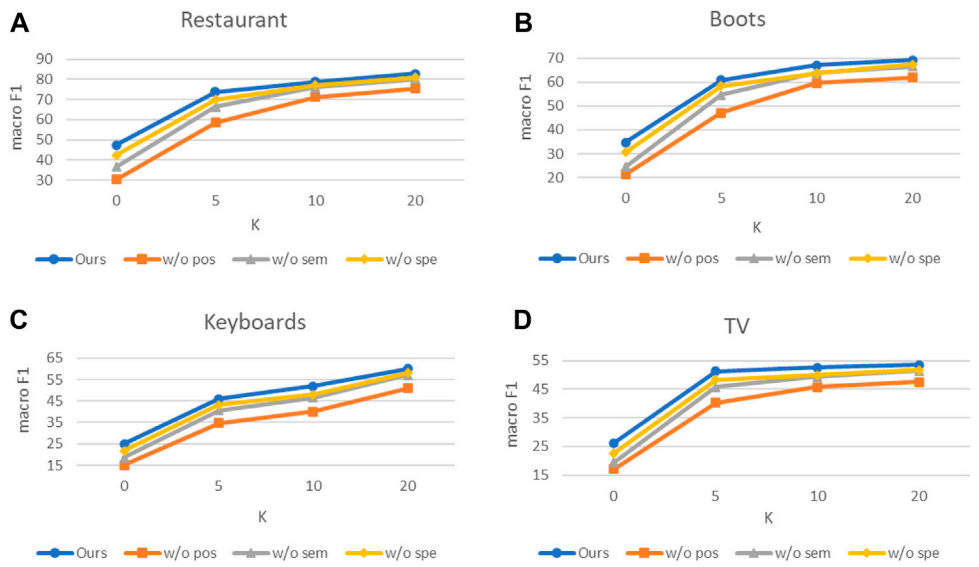

FIGURE 2
Ablation study on four datasets. **(A)** Ablation Study on "Restaurant" dataset. **(B)** Ablation Study on "Boots" dataset. **(C)** Ablation Study on "Keyboards" dataset. **(D)** Ablation Study on "TV" dataset.

that the most reasonable prompt words are searched for each category, thereby ensuring model performance. In addition, the following can be clearly observed: 1) The "w/o pos" model performs the worst on all four datasets, and the growth rate is lower than that of other models. This shows the prompt word set that has not been denoised contains more meaningless tokens, and these tokens have a higher prediction probability when filling in the [MASK] position, resulting in a decrease in the mapping performance of the verbalizer. 2) The performance of the "w/o sem" and "w/o spe" models is similar, indicating that category specificity and category semantic similarity are equally important when searching for prompt words. The common constraints of the two make each prompt word set not only have as many prompt words as possible and avoid mapping contradictions between different categories, which is beneficial to the subsequent mapping process.

4.6 Comparison of the miscellaneous category

This section quantitatively and qualitatively studies the effects of different models on the "miscellaneous" and "general" categories. The Amazon dataset contains the "general" category. For convenience of presentation, the two labels are collectively referred to as "miscellaneous" below. Figure 3 shows the results of each model under zero-shot and few-shot conditions, respectively. Table 4 shows the prompt words of "miscellaneous" obtained by different models. As shown in Figure 3, our model showed excellent results in different settings; especially in the zero-shot task, the improvement effect is obvious. On the zero-shot task, our method demonstrates improvements of 14.1%, 10.6%, 6.9%, and 11.1% compared to the second-best model across four datasets. Additionally, for the 10-shot task, our method exhibits enhancements of 3.0%, 4.8%, 6.3%, and 3.8% on the same datasets, respectively.

Referring to the data in Table 4, we speculate that because the sentences of the "miscellaneous" category have no obvious characteristics and the range of semantic expression is wide, the manual method only uses category word as the prompt word, which obviously cannot cover all data of this category, so the results are not ideal. Although KPT has expanded the scope of mapping, most of the prompt words searched from the external knowledge base for this category are uncommon and cannot be recognized by PLMs, resulting in poor performance in this category. Although the search-based model does not suffer from these two problems, it ignores the confusion between categories and can easily cause misjudgments during the prediction process. In addition, our model focuses on and solves the above problems, and the obtained prompt words have a high correlation with the category and can show good prediction ability on semantically ambiguous sentences.

4.7 Impact of the semantic extension

Our method still has certain prediction ability in the case of zero-shot because of using the semantic extended sentences of category labels as prior knowledge. This section studies the impact of the semantic extended sentences of category labels. Figure 4 shows the effect of the length of the semantic extension sentence on the final results. Wikipedia has a very detailed explanation for category words, usually from different aspects, so the optional range of semantic expansion sentences is long. The length "len" is calculated based on the number of tokens. In addition to using the category word itself with "len" as 1, the length of the semantic extended sentence is changed by continuously increasing the number of tokens in the definition statement.

10.3389/fphy.2024.1361695

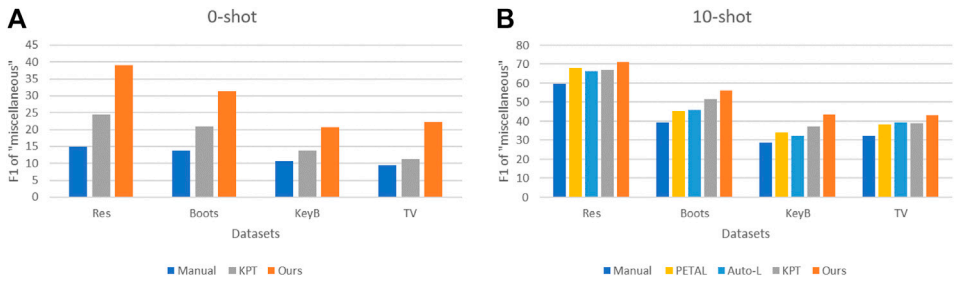

FIGURE 3
Experiments on the "Miscellaneous" category. **(A)** Zero-shot experiments on "Miscellaneous" category. **(B)** Few-shot experiments on "Miscellaneous" category.

TABLE 4 Prompt words for the "Miscellaneous" category.

Dataset	Method	Verbalizer token
Restaurant	Manual	miscellaneous
	PETAL	darkness, fiction, opinions, academia, interests, sociology, links,. . .
	Auto-L	Bryan, anonymous, Wes, noise, LM, Ethan, Wayne, dark,. . .
	KPT	heterogeneous, diverse, dissimilar, disparate, different, unlike,. . .
	Ours	same, general, main, particular, whole, specific, various, primary,. . .
Amazon Boots	Manual	general
	PETAL	remembered, Articles, arrived, finished, published, instructed,. . .
	Auto-L	produced, systems, published, female, default, quoted, . . .
	KPT	army, officer, brigadier, military, air, commander, field,. . .
	Ours	interesting, done, closed, clear, true, possible, established, like,. . .
Amazon Keyboards	Manual	general
	PETAL	votes, remarks, guy, Subject, excerpt, speakers, policy,. . .
	Auto-L	god, voice, journal, Jackson, guy, James, blogger, admin, hi,. . .
	KPT	lieutenant, cosmopolitan, universal, ecumenical, consumable,. . .
	Ours	included, fix, changed, summary, various, basic, likely,. . .
Amazon TV	Manual	general
	PETAL	url, AUTHOR, Hannah, username, starred, published, Votes,. . .
	Auto-L	controversy, followers, community, ranking, Society, twitter,. . .
	KPT	generality, rank, oecumenical, commander, admiral, full. . .
	Ours	titled, defined, concluded, called, summarized, cited, listed,. . .

Combining the results of the four datasets, it can be observed that the experimental results have improved with the increase in extended sentence tokens. This may be because the semantics of sentence expressions are rich, and PLMs can better understand the meaning of category labels to search for more reasonable prompt words. However, when the length is too long, the performance decreases instead. We speculated that it may be because the meaning contained in the semantic extensions is too complex leading to understanding deviation, which is not conducive to the model to choose more accurate prompt words. The semantic extended sentences used in the best experimental results of the method can be found in Table 2.

4.8 Impact of the templates

Template is another important component that affects prompt learning performance, so in this section, we tested the impact of different prompt templates on the proposed method. Table 5 lists all

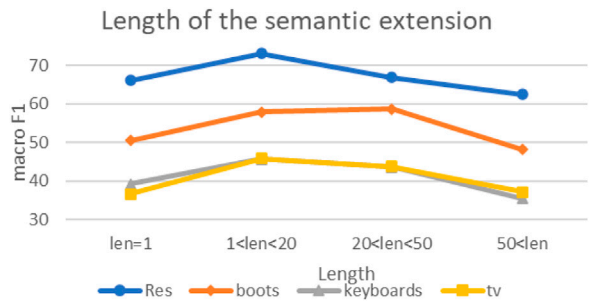

FIGURE 4
Impact of semantic extension length.

TABLE 5 Templates used in experiments.

ID	Templates
1	The "mask" category is discussed
2	The sentence discusses the "mask" category
3	It is about the "mask" category
4	Category: "mask"

the templates used in experiments. Figure 5 shows the results of using different templates on the Restaurant and Boots datasets under the 10-shot setting. As can be seen from the figure, our model not only maintains excellent performance in both datasets but also has a relatively gentle change curve compared with some other methods, indicating that it has a certain degree of robustness to different templates.

4.9 Impact of hyperparameters

In this section, we explore the impact of hyperparameters on experimental results and conduct grid searches on the Restaurant and Boots datasets for the two hyperparameters of "taking the first M words" and "taking the specificity probability greater than the threshold α." For the parameter M and parameter α, we set them to $\{50, 100, 300, 500, 1000\}$ and $\{0.90, 0.80, 0.75, 0.70, 0.60\}$, respectively. We use grid search to find the optimal values within the ranges of two parameters. The experimental results for parameter M are shown in Figure 6A. The results show that as the M value continues to increase, the model performance increases. However, when M increases to 1,000, the performance decreases, indicating that at too large M, it may select some low-quality prompt words, resulting in the reduction of the final classification results. Similar to the M value, as shown

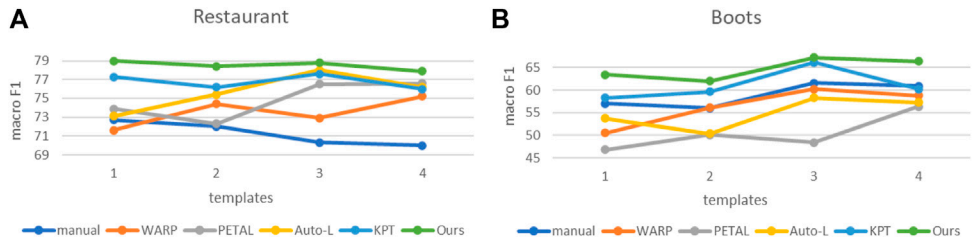

FIGURE 5
Impact of templates. **(A)** Experiment on "Restaurant" dataset. **(B)** Experiment on "Boots" dataset.

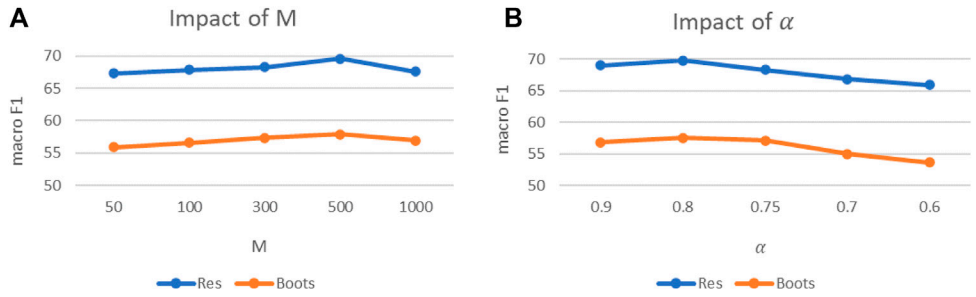

FIGURE 6
Impact of hyperparameters. **(A)** Experiment of parameter M. **(B)** Experiment of parameter α.

TABLE 6 Case study.

Examples	Manual prompt	KPT	Ours
I highly recommend this restaurant!! (Restaurant dataset)	food (✗)	food (✗)	miscellaneous (✓)
If you turn backlight all the way down it gets better (TV dataset)	sound (✗)	image (✓)	image (✓)

in Figure 6B, the α experimental curve also shows a trend of increasing first and then decreasing. This is because a too small α value will also introduce low-quality words and affect the model performance. The best experimental results in this article were obtained when M = 800 and α=0.8.

4.10 Case study

Table 6 shows some examples from different test sets. For example 1, the meaning expressed by this sentence does not belong to the categories "food," "service," "price," and "ambience," but to "miscellaneous." We speculate that due to the word "restaurant" in the sentence, the prompt word "restaurant" in the "food" category from the KPT model is easy to obtain a higher probability, and thus, it is mapped to the "food" category. The manual method detected errors in both examples. This may be because the category words themselves cannot better summarize the meaning of the example sentences, and it is easy to be misjudged as other categories.

5 Conclusion

In this paper, we propose a simple and effective method for aspect category detection based on prompt learning. To address the challenge of lack of labeled data and external knowledge, the semantic expansion of category labels is exploited to build the initial verbalizer. Additionally, we employ an indication mechanism to construct an appropriate verbalizer for category mapping. We conduct experiments on zero-shot and few-shot settings, respectively, and the results demonstrated the superiority of the proposed method. In our article, the verbalizer is constructed under a predefined manual template. In recent years, there has been a lot of work exploring the design of templates, but in most cases, the construction of the two is still separated, and both require certain labeled data. Therefore, in future work, we plan to further explore how to build prompt templates and verbalizers simultaneously to find the best combination of these two components.

.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

YX: conceptualization, formal analysis, methodology, validation, and writing–review and editing. YQ: software, validation, and writing–original draft. YX: conceptualization, formal analysis, methodology, validation, and writing–review and editing.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62162037 and U21B2027), Yunnan provincial major science and technology special plan projects (Grant Nos. 202302AD080003 and 202303AP140008), and Yunnan provincial major science and technology special plan projects (Grant No. 202301AT070444).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hu M, Zhao S, Guo H, Xue C, Gao H, Gao T, et al. *Multi-label few-shot learning for aspect category detection* (2021). *arXiv preprint arXiv:2105.14174*.

2. Zhou X, Wan X, Xiao J. Representation learning for aspect category detection in online reviews. In: Proceedings of the AAAI conference on artificial intelligence, 29 (2015). p. 1547–52. doi:10.1609/aaai.v29i1.9194

3. Xue W, Zhou W, Li T, Wang Q. Mtna: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (2017). p. 151–6. 2: *Short Papers*.

4. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. *Language models are few-shot learners advances in neural information processing systems* (2020). 33.

5. Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, et al. *Language models as knowledge bases?* (2019). *arXiv preprint arXiv:1909.01066*.

6. Schick T, Schütze H. *Exploiting cloze questions for few shot text classification and natural language inference* (2020). *arXiv preprint arXiv:2001.07676*.

7. Kiritchenko S, Zhu X, Cherry C, Mohammad S. Nrc-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th international workshop on semantic evaluation (2014). p. 437–42. SemEval 2014.

8. Hambardzumyan K, Khachatrian H, May J. *Warp: word-level adversarial reprogramming* (2021). *arXiv preprint arXiv:2101.00121*.

9. Xenos D, Theodorakakos P, Pavlopoulos J, Malakasiotis P, Androutsopoulos I. Aueb-absa at semeval-2016 task 5: ensembles of classifiers and embeddings for aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (2016). p. 312–7. *SemEval-2016*.

10. Hercig T, Brychcín T, Svoboda L, Konkol M. Uwb at semeval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (2016). p. 342–9. Sem Eval-2016.

11. Toh Z, Su J. Nlangp at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features. In: Proceedings of the 10th international workshop on semantic evaluation (2016). p. 282–8. *SemEval-2016)*.

12. Su Q, Xiang K, Wang H, Sun B, Yu S. Using pointwise mutual information to identify implicit features in customer reviews. In: International Conference on Computer Processing of Oriental Languages. Springer (2006). p. 22–30.

13. Schouten K, Van Der Weijde O, Frasincar F, Dekker R. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE Trans cybernetics* (2017) 48:1263–75. doi:10.1109/tcyb.2017.2688801

14. Jiang Z, Xu FF, Araki J, Neubig G. How can we know what language models know? *Trans Assoc Comput Linguistics* (2020) 8:423–38. doi:10.1162/tacl_a_00324

15. Haviv A, Berant J, Globerson A. *Bertese: learning to speak to bert* (2021). *arXiv preprint arXiv:2103.05327*.

16. Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S. *Autoprompt: eliciting knowledge from language models with automatically generated prompts* (2020). *arXiv preprint arXiv:2010.15980*.

17. Gao T, Fisch A, Chen D. *Making pre-trained language models better few-shot learners* (2020). *arXiv preprint arXiv:2012.15723*.

18. Li XL, Liang P. *Prefix-tuning: optimizing continuous prompts for generation* (2021). *arXiv preprint arXiv:2101.00190*.

19. Lester B, Al-Rfou R, Constant N. *The power of scale for parameter-efficient prompt tuning* (2021). *arXiv preprint arXiv:2104.08691*.

20. Qin G, Eisner J. *Learning how to ask: querying lms with mixtures of soft prompts* (2021). *arXiv preprint arXiv:2104.06599*.

21. Han X, Zhao W, Ding N, Liu Z, Sun M. Ptr: prompt tuning with rules for text classification. *AI Open* (2022) 3:182–92. doi:10.1016/j.aiopen.2022.11.003

22. Schick T, Schmid H, Schütze H. *Automatically identifying words that can serve as labels for few-shot text classification* (2020). *arXiv preprint arXiv: 2010.13641*.

23. Hu S, Ding N, Wang H, Liu Z, Wang J, Li J, et al. *Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification* (2021). *arXiv preprint arXiv:2108.02035*.

24. Gu Y, Han X, Liu Z, Huang M. *Ppt: pre-trained prompt tuning for few-shot learning* (2021). *arXiv preprint arXiv:2109.04332*.

25. Zhu Y, Wang Y, Qiang J, Wu X. Prompt-learning for short text classification. *IEEE Trans Knowledge Data Eng* (2023) 1–13. doi:10.1109/tkde.2023.3332787

26. Cui G, Hu S, Ding N, Huang L, Liu Z. *Prototypical verbalizer for prompt-based few-shot tuning* (2022). *arXiv preprint arXiv:2203.09770*.

27. Meng Y, Zhang Y, Huang J, Xiong C, Ji H, Zhang C, et al. *Text classification using label names only: a language model self-training approach* (2020). *arXiv preprint arXiv: 2010.07245*.

28. Liu H, Zhang F, Zhang X, Zhao S, Sun J, Yu H, et al. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022). p. 1079–87.

29. Zhang W, Song X, Feng Z, Xu T, Wu X. *Labelprompt: effective prompt-based learning for relation classification* (2023). *arXiv preprint arXiv:2302.08068*.