Prompt-Learning for Short Text Classification

Yi Zhu, Ye Wang, Jipeng Qiang, Member, IEEE, and Xindong Wu, Fellow, IEEE

Abstract—In the short text, the extremely short length, feature sparsity, and high ambiguity pose huge challenges to classification tasks. Recently, an effective method for tuning Pre-trained Language Models for specific downstream tasks, prompt-learning has attracted a vast amount of attention and research. The main intuition behind the prompt-learning is to insert the template into the input and convert the tasks into equivalent cloze-style tasks. However, most prompt-learning methods only consider the class name and monotonous strategy for knowledge incorporating in cloze-style prediction, which will inevitably incur omissions and bias in short text classification tasks. In this paper, we propose a short text classification method with prompt-learning. Specifically, the top M concepts related to the entity in the short text are retrieved from the open Knowledge Graph like Probase, these concepts are first selected by the distance with class labels, which takes both the short text itself and the class name into consideration during expanding label word space. Then, we conducted four additional strategies for the integration of the expanded concepts, and the union of these concepts are adopted finally in the verbalizer of prompt-learning. Experimental results show that the obvious improvement is obtained compared with other state-of-the-art methods on five well-known datasets.

Index Terms—Prompt-Learning, short text classification.

I. INTRODUCTION

With the rapid development of web service, short-texts, which are posted at unprecedented rates, accentuate both the importance of learning tasks and the challenges posed by the inherent properties such as the extremely short length, feature sparsity, and high ambiguity. In recent decades, short text classification has attracted a vast amount of attention and research from multiple disciplines [1], [2], the advances in short text data processing have far-reaching consequences in practical applications like Twitter [3], Facebook [4], and Microblog [5].

Existing methods for short text classification can roughly be categorized into two classes: sole source and external knowledge base methods. The methods based on sole source extend the feature space by the rules or statistical information hidden in the current short texts [6], [7]. The sole source methods still face the severe data sparsity problem, the methods based on external knowledge are applied widely in recent years, which extends the feature space by external open knowledge base [8], [9]. However, most existing methods based on external knowledge rely on large-scale training instance data to formalize the model, which leads to high costs in collecting eligible training data and performance degradation in few-shot learning.

Recently, Pre-trained Language Models (PLMs) have attained much attention and remarkable improvements in a series of downstream Natural Language Processing (NLP) tasks, such as text classification [10], question answering [11], machine translation [12], and lexical simplification [13]. PLMs can learn syntactic [14], semantic [15], and structural [16] information about language. To adapt the versatile knowledge contained in PLMs to various NLP tasks, the fine-tuning method with an extra classifier has been applied widely to stimulate and exploit rich knowledge in PLMs and has achieved excellent performance in downstream tasks [17].

Among all the fine-tuning approaches, inspired by GPT-3 that uses the information provided by the prompts in few-shot learning and achieves substantive results [18], prompt-learning fills the input statements into the natural language template and adapts the masked model, which formalized regarded the downstream NLP task as cloze-style tasks. For example, to classify the topic of sentence $x$ as “Call them the ‘Nightmare Team’. " into the “Sports” category, the template can be noted as “the [MASK] question”, and prompt-learning predict the probability that the word “sports” is filled in the “[MASK]”. Compared with the previous fine-tuning PLMs approaches, no additional neural layer is needed in prompt-learning, and excellent performance has been achieved even in the scenario of few-shot or zero-shot learning. Moreover, in the prompt-learning, the mapping from label words (e.g., sports, association, basketball et al.) to the category (e.g., SPORTS) can effectively alleviate the discrepancy between text and label space, which is called the automatic selection of label words [19] or the verbalizer [20]. This strategy of constructing label word mapping has been proven to be effective in achieving more desirable text classification performance [21].

Due to the inherent characteristics of short texts, contemporary classification methods cannot achieve satisfactory results for feature sparsity and high ambiguity. There are two main problems that prevent the further development of existing short text classification methods: (1) Feature sparsity: most short texts only contain dozens of words, such as in Twitter or Microblog, which inevitably leads to feature sparsity and performance degradation. (2) High ambiguity: due to the short length introduced the low
word co-occurrence and concept confusion, how to accurately find the corresponding concepts and labels of short texts is still a great challenge.

To address these problems, in this paper, we present an intuitive and innovative idea for short text classification, we exploit recent advances in the prompt-learning model [22] based on knowledgeable expansion. For the first problem, taking the feature sparsity of short text into consideration, the proposed Prompt-Learning approach for Short Text classification (PLST) incorporated both the short text itself and external knowledge from open Knowledge Graph like Probase to extend label word space. For the second problem, the top \(N\) concepts concerning the entities in the short text are first retrieved from Probase, and the integration of label words from four strategies is then conducted, and the union of these concepts is adopted finally in the verbalizer of prompt-learning. The advantage of our method is that it generates more effective label words by considering the short text itself, not just the class name.

Here, we give an example shown in Fig. 1 to illustrate the advantage of our method PLST. For one sentence “Ford cuts production while Chrysler’s sales rise,” and template “This topic is about [mask]”, the label word space in Prompt-tuning [22] only includes the class name ‘business’, which refers to that only predicting the word “business” for the [MASK] token is regarded as correct regardless of other relevant words. The expanded label words generated by Knowledgeable Prompt-tuning [23] contain plenty of words as {commerce, trade, market, antique, purchase,...}, which only related to the class name ‘business’ without paying attention to the original sentence. The expanded label words generated by our PLST are not related to the class name, but also can expand knowledge from the original sentence. Then, by considering the candidate substitutions ranking with multiple strategies, ‘company’ and ‘manufacturer’ are selected as the expansion for the class business. In this case, the expanded label word space {business, company, manufacturer,...} is more accurate and more reliable than other label word space.

The contributions of our paper are summarized as follows:

1) Our PLST is a novel prompt-learning-based method for short text classification, which can take full advantage of prompt-learning to bridge the objective form gap between pre-training and fine-tuning. Compared with existing methods, our PLST can achieve more desirable performance, since it considers both the short text and class name during expanding label word space.

2) Our PLST is a simple, effective, and scalable short text classification method. 1) Simple: many steps used in existing short text classification methods have been eliminated from our method, e.g., co-occurrence words searching and the entire model tuning. 2) Effective: it obtains new state-of-the-art results on five benchmarks. 3) Scalability: the strategies of the expanded words integration can be extended by multiple kinds of features and strategies, the union of these expanded words from different strategies can obviously improve the performance of short text classification.

3) To our best knowledge, this is the first attempt to apply prompt-learning models on short text classification tasks.

The code to reproduce our results is available at https://github.com/zhuyiYZU/PLST.

II. RELATED WORK

The last decades have witnessed a vast amount of interest and research on short text, which has played a crucial role in many real-world scenario applications. Recently, PLMs and prompt-learning have attained substantial performance in a series of downstream NLP tasks, and a prompt-learning strategy that incorporated knowledge for verbalizer is proposed in this paper. In this section, we review the literature pertaining to short text classification, prompt-learning, and verbalizer construction respectively.

A. Short Text Classification

Short text classification has provoked a vast amount of attention and research in recent decades [8], which has played an important role in many practical applications like sentiment analysis [24], dialogue systems [25], and user intent understanding [26]. Short text classification aims to process texts with very short length, usually no more than 100 characters, such as blog content [5], online comments [27], news title [28], and so on.

Existing short text classification methods can be roughly divided into two categories: sole source methods and external knowledge methods. The methods based on sole source expand the feature space by disentangling explanatory factors of variations behind the current short text [6]. For example, Lai et al. proposed a recurrent convolutional neural network to capture contextual information and the key components for text classification [7]. Bollegala et al. proposed a ClassiNet network to predict missing features for addressing feature sparseness problems, unlabeled data are utilized to generalize word co-occurrence graphs, and the relations between features and short text are explored [29]. Hao et al. proposed a Mutual-Attention Convolutional Neural Networks for Chinese short text classification, which integrates word and character-level features without losing feature information [30].

However, the sole source methods still face the severe data sparsity problem, the methods based on external knowledge are applied widely in recent years, which extends the feature space by the external open knowledge base. For example, Chen et al. proposed to retrieve knowledge from external knowledge source to enhance the semantic representation of short texts, and attention mechanisms are introduced in this method to acquire the weight of concepts [8]. Xu et al. proposed a hybrid model to incorporate context-relevant knowledge into a convolutional neural network for short text classification [9]. Yang et al. proposed a heterogeneous information network to incorporate additional information and their relations from open Knowledge Base, which can address the semantic sparsity problem in short text classification [31]. Although these methods can obtain fairly good results in short text classification, they all rely on the large-scale training instance data to formalize the model, which leads to high costs in collecting eligible training data and performance degradation in few-shot learning.
B. Prompt-Learning

Recently, the fine-tuned Pre-trained Language Models have achieved tremendous success in various NLP tasks, such as question answering [11, 32], text classification [10, 33], machine translation [12, 34], and lexical simplification [13, 35]. PLMs can learn syntactic [14], semantic [15] and structural [16] information about language. For example, PromptEM focused on the low-resource Generalized Entity Matching (GEM) [36], which addressed three challenging issues including designing GEM-specific prompt, selecting high-quality pseudo-labels, and avoiding expensive self-training in low-resource GEM. Ye et al. constructed a graph convolutional network for short text, and the word and document nodes trained by the GCN and vector generated by the BERT are input into the BiLSTM classifier for short text classification [37]. Sun et al. conducted extensive experiments to investigate the different approaches to fine-tuning BERT and achieved state-of-the-art performance on short text classification tasks [38].

However, the huge gap between pre-training and fine-tuning still prevents downstream tasks from fully utilizing pre-training knowledge. To this end, inspired by GPT-3 [18], prompt-learning has been proposed to transfer downstream tasks as some cloze-style objectives and achieved superior performance, especially in few-shot learning [39]. Along this line, many hand-crafted prompts have been made in various tasks, such as knowledge probing [40, 41], relation extraction [42], entity typing [43] and text classification [23, 44]. For example, Han et al. applied logic rules to construct prompts with several sub-prompts on relation classification, which consistently outperforms existing state-of-the-art baselines without introducing any additional model layers, manual annotations, and augmented data [42]. Ding et al. proposed a prompt-learning method on fine-grained entity typing, the entity types are extracted without over-fitting by performing distribution level optimization [43]. Schick et al. proposed Pattern Exploiting Training (PET) method, which consists of defining pairs of cloze question patterns and verbalizers to stimulate rich knowledge in PLMs [45]. The models in this method are fine-tuned for all pattern-verbalizer pairs and then applied to create large annotated datasets on which standard classifiers can be trained. Furthermore, to avoid time-consuming and labor-intensive prompt design, a series of automatic prompt generation methods have been explored recently [46], [47]. For example, Shin et al. proposed an automatically generated prompts method to elicit knowledge from language models on sentiment analysis and textual entailment [48]. Hambardzumyan et al. proposed an automatic prompt generation method to transfer knowledge from large PLMs to downstream tasks by appending embeddings to the input text, which showed good performance in a few-shot setting [49]. Although automatic prompt generation methods can avoid intensive labor, most of these methods cannot achieve comparable performance to manually selected prompt methods in the real-world scenario [42].

C. Verbalizer Construction

In the prompt-learning, the verbalizer refers to a projection from label words (e.g. sports, basketball et al.) to the category (e.g. SPORTS), which has been proven to be an important and effective strategy for alleviating the discrepancy between text and the label space [20]. The hand-crafted verbalizers have achieved sound performance in text classification and other NLP tasks. For example, Schick et al. proposed to use pairs of cloze question patterns and manually designed verbalizers for leveraging the knowledge contained within PLMs for downstream tasks [21]. However, the manually designed verbalizers are highly impacted by prior knowledge and incur omissions and bias for knowledge expansion.

Since the hand-crafted verbalizers require enough training and validation sets for optimization, a series of automatic verbalizer construction methods in the prompt-learning are presented [19], [20]. For example, Wei et al. proposed a prototypical network to generate prototypical embeddings for different labels in the feature space, which summarized the semantic information of labels to form a prototypical prompt verbalizer [50]. However, synonyms of category names are more likely to be expanded instead of diverse and comprehensive label words in these methods. To denoise expanded label words in automatic verbalizer, some other works try to select related words from external knowledge base [23]. Such a method can greatly enhance the semantics of labels, but due to a large number of useless words being extracted during the stage of verbalizer construction, the verbalizer is difficult to use directly and leads to unsatisfied results in downstream tasks like text classification. Compared to the previous methods, in this paper, we proposed a prompt-learning strategy to significantly improve the performance of short text classification. Not only class name but also the short text itself are taken into consideration for knowledgeable expansion, and the integration of label words is conducted with four additional strategies for verbalizer.

III. METHODOLOGY

In this section, we briefly summarize the motivation of our PLST, and then the overall architecture, label words set construction, cluster, and integration, and text classification are described successively in detail.

A. Motivation

Short-texts, which are posted at unprecedented rates, accenuate both the importance of learning tasks and the challenges posed by such extremely short length. Recently, fine-tuning PLMs approaches have achieved excellent performance in short text classification. However, most PLMs-based methods, such as BERT, T5, and Prompt, still face several problems. The first problem of which is the feature sparsity. Despite some methods have introduced popular taxonomy knowledge base (such as Wikipedia) for feature extension, e.g. Knowledgeable Prompt-tuning method (KPT) [23], the Probase1 used in our paper provides probability for the super-concepts of an instance or concept, which is also one order of magnitude larger than Wikipedia. Therefore, the introduced concepts can effectively and better address the problem of data sparsity.

The second problem is the high ambiguity of short texts. Due to the short length introduced the low word co-occurrence

1http://research.microsoft.com/en-us/projects/probase/release.aspx
and concept confusion, most existing methods are difficult to find the corresponding concepts accurately. For example, KPT introduced the Related Words,\(^3\) an open knowledge base, to extend the concepts for short text, and the method proposed a contextualized calibration method for denoising. However, the role of the short text itself in reducing conceptual confusion is not taken into consideration, which may lead to poor noise removal and performance degradation. In contrast, first, the concepts concerning the entities in the short text are retrieved and refined by the distance between concepts and class labels in our method. In the experiments on the AG’s News dataset, the average number of expanded concepts on four classes is 344, which is much larger than the \((M + 4 \times N_a)\) in our method, where \(M\) is the number of concepts retrieved from Probase by taking the short text itself into consideration, and \(N_a\) is the number of expanded words for the verbalizer by the other four strategies.

The third problem is the selection of the expanded concepts. Although most existing methods extended the feature space by the external open knowledge base, the accompanying noise is still detrimental to the performance and incurs a loss with non-trivial magnitude. The selection of the expanded concepts can not only alleviate the negative effects of noise, but also reduce the overall running time and improve the efficiency. In our method, besides the distance between retrieved concepts and class labels in embedding space is calculated, we conducted four additional strategies for the integration of the expanded concepts, the union of these concepts is adopted finally in the verbalizer of prompt-learning.

B. Overall Architecture

The proposed Prompt-Learning approach for Short Text classification (PLST) stems from the recent success of the prompt-tuning model. As shown in Fig. 2, the method consists of three main components: concept retrieval, verbalizer construction, and text classification. First, taking the extremely short length of the short text into consideration, we incorporated both the short text itself and external knowledge from Probase for concept retrieval. The top \(N\) concepts concerning the entities in the short text are retrieved from Probase to extend label word space. Compared to other methods like KPT, we retrieve concepts from short text itself, instead of finding concepts directly in large knowledge base by class name. Second, besides the distance between retrieved concepts and class labels in embedding space is calculated for verbalizer construction, we introduce other four strategies for the furthermore integration of the expanded concepts, the union of these concepts is adopted in the verbalizer. Finally, the constructed verbalizer is utilized in prompt-learning to predict the probability of each label word to a special class for text classification.

C. Concept Retrieval

In the prompt-learning method, the input statements are formalized as the natural language template, and the text classification tasks are regarded as the close-style tasks. For example, in topic classification, assuming we need to classify the sentence \(x^*\) “yukos cuts output to save money” into the label \(y_1 = BUSINESS\) or \(y_2 = SPORTS\), the template can be noted as:

\[
x_p = [CLS]x, a \text{ [MASK] question}
\]

given the input \(x = \{x_1, \ldots, x_n\}\), which is classified into a category with label \(y \in Y\), the label word set is denoted as \(V_y = \{v_1, \ldots, v_m\}\), where \(V_y\) is a subset of the whole vocabulary \(V\), i.e., \(V_y \subseteq V\), and \(V_y\) is mapped into a category with label \(y\). In PLMs \(M\), the probability that each word \(v\) in \(V_y\) is filled in the [MASK] can be shown as \(p([MASK] = v \mid x_p)\). Thus, the text classification tasks can be transferred into a probability calculation problem of label words, which can be computed as

\[
p(y \mid x) = p([MASK] = v \mid x_p) \tag{1}
\]

In the above-mentioned example, if the calculated probability of \(V_y = \{business\}\) for \(y_1 = BUSINESS\) is larger than \(V_y = \{sports\}\) for \(y_2 = SPORTS\), it indicates that the sentence \(x\) is classified into the category BUSINESS. In the scenario of automatic expansion of label words or verbalizer construction, the \(V_y\) related to a special category with label \(y\) is expanded, such as \(V_y = \{business\}\) can be expanded as \(V_y = \{business, commerce, company, \ldots\}\) in the above example, which can obviously improve the short text classification performance with prompt-learning.

In short text classification, the crucial problem for label words expansion is the hierarchical label space, which refers to multiple aspects and granularities. For example, “country”, “province”, and “city” are multi-level and related words, and they may be all fit the predicting masked words in the template of prompt-learning. To this end, we select Probase\(^3\) as the external knowledge source, which is an open Knowledge Graph constructed by Microsoft. Probase specifies the probability of each instance belonging to the concept, and the concepts with relatively small correlations will be abandoned in the process of label words expansion, which reduces the difficulty in verbalizer refinement.

To address the problem of high ambiguity of short texts mentioned above, different from the existing methods that extended the concepts directly from the knowledge base, we extract entities from the short text itself. In the experiments, given a short text \(s_i\), we first extract \(e_i\) entities from \(s_i\). Then, the \(e_i\) entities are input into Probase to retrieve \(N_i\) concepts. Since the verbalizer is a projection between the label word set and the category label space, in the embedded space, the distance \(\text{dist}(N_i, y)\) between each expanded label word set and each label name \(y\) is first calculated. Thus, the top \(N_j\) words are extracted whose vectors are closer in terms of cosine similarity with the label name \(y\). For all the short texts, the \(\{N_1, \ldots, N_j, \ldots, N_n\}\) are ranked by the similarity and the top \(M\) words are selected for each category as verbalizer refinement, excluding the morphological derivations of \(y\). In this way, the verbalizer is first refined to each topic, which not only considers the hierarchical concept itself but also can fit each label word into a special category. For instance,
in the above-mentioned example Ford cuts production while Chrysler’s sales rise. The concepts of multiple entities, such as Ford and Chrysler, are both retrieved from Probase. Then in the embedded space, the distance between concepts of Ford and Chrysler and the label name Business or Sports is calculated, and \( N_1 \) words are ranked by the similarity. Finally, for all the short texts, \( M \) words are selected from \( \{N_1, \ldots, N_j, \ldots, N_n\} \) for each category as verbalizer refinement.

It is worth mentioning that the number of instances can be very huge in large datasets. For example, there are over 20000 short texts in the OHSUMED dataset. In the experiments, we only randomly select some instances to retrieval concepts instead of picking all the short texts in the datasets, and we call this selection as “support set”. We believe that a certain amount of short text can already cover the extension words of a certain topic, and the related experiments for “support set” are shown in Section IV-H Influence of the Support Set.

D. Verbalizer Construction

Although the expanded label words set is filtered by concepts probabilities, there are still many useless and noisy words since the gap between the pre-training model and concepts in the knowledge graph. To expand high-quality words for the verbalizer, we conduct additional four strategies as follows. The versatile knowledge in PLMs, the higher-probability words for classification, the frequency-based features and the context information are taken into consideration, each of the strategies captures one aspect of characteristics of the expanded word and the union of these words is adopted in the final verbalizer.

- Prediction by BERT: To utilize the versatile knowledge contained in PLMs, the probability distribution of masked word in the template, such as “yukos cuts output to save money, a [MASK] question” in the above example, is obtained. The probability refers to the relevance for the corresponding class, we rank all the words according to the probability for the top \( N_a \) words selection.

- FastText Semantic similarity: The semantic similarity is calculated between the fastText vector of each class label name \( y \) and the fastText vector of expanded label words. Same as the first strategy, the words with the lower similarity are discarded and the top \( N_a \) words are selected.

- Frequency words selection: Roughly speaking, the higher the frequency, the more application in reality. The words frequency estimation is conducted based on Probase, and the top \( N_a \) concepts are selected to match the size.

- Context information: The expanded words should take the sequence of words preceding and following the masked word into consideration, which refers to context information. In our method, we propose to replace the masked word into the expanded word. Then, we mask one word around the masked word from front to back, and feed it into BERT to calculate the cross-entropy loss. All the expanded words are sorted according to the average loss. In the experiments, we only take a symmetrical window of size five around the masked word, and the words with the higher loss are discarded.

It is worth mentioning that we take \( N_a = 15 \) on all four strategies in the experiments. As the result, we finally select \( (M + 4 + N_a) \) for verbalizer in our method, which is much more smaller than the verbalizer in existing methods. To intuitively show the results, we list the expanded concepts and label words of KPT and our PLST in Table I, and we can clearly observe that our PLST extends fewer and more accurate concepts than KPT.

E. Text Classification

After the final verbalizer construction with different strategies, we need to map the predicted probability on each label word to a special class, which can be noted as an objective function \( g \) for verbalizer utilization. Due to that each word in the final verbalizer can be assumed to contribute equally for predicting, the average of the predicted scores is used for text classification, i.e., \( g \) can be calculated as (2):

\[
\arg\max_{y \in Y} \frac{1}{|V_y|} \sum_{v \in V_y} p([MASK] = v|x_p) \quad (2)
\]

Suppose that there is a sentence “Ford cuts production while Chrysler’s sales rise.” and the corresponding topic is BUSINESS, we can obtain the label words in verbalizer as \{company, manufacturer, industry, investor, competitor,...\}. The whole framework of our proposed PLST is illustrated in Fig. 2, and we can see the expanded label words not only have a strong correlation with the category BUSINESS but also hold hierarchical and multi-granularities properties of the topic. If we adopt the existing state-of-the-art method based on Knowledgeable Prompt-tuning proposed by [23], the expanded label words are \{commerce, trade, market, antique, purchase,...\}. Very obviously, our method generates a better label words set.
ZHU et al.: PROMPT-LEARNING FOR SHORT TEXT CLASSIFICATION

Fig. 2. Illustration of our PLST for short text classification. MLM is short for Masked Language Modeling. The left part refers to the input text: “Ford cuts production while Chrysler’s sales rise.” with the template: “This topic is about [mask].” The center part is the concept retrieval by the short text itself and class name from Probase and verbalizer construction by four additional strategies. The right part is the final text classification. The red part in Probase refers to the concept node corresponding to ‘Chrysler’ in the input text.

TABLE I
EXAMPLES OF THE EXPANDED CONCEPTS ON AG’S NEWS DATASET

| Method | Label | Label Words Set |
|--------|-------|-----------------|
| KPT    | BUSINESS, SPORTS | vice, sector, planning, act, arts, transportation, barter, public, chaffer, self-sufficiency, ... |
| PLST   | BUSINESS, SPORTS | industry, company, manufacturer, investor, provider, giant, fund, insurance, ... |

Fig. 3. Parameter Influence of N on all three datasets.

IV. EXPERIMENTS

In this section, we conduct experiments on five datasets to evaluate the effectiveness of our proposed method in short text classification.

A. Datasets and Templates

The experiments are conducted on five well-known benchmark short text datasets, which are described as follows, and the statistics of each dataset are listed in Table II.

AG’s News\(^4\): We conducted the experiments on AG’s news topic classification dataset, which is constructed by selecting the four largest classes from the original AG’s News corpus. AG’s news topic classification dataset contains 120K English news from more than 2000 news sources with four categories as World, Sports, Business, and Sci/Tec. Notably, we only pick the news title of AG’s News as short text in the experiments.

Snippets\(^5\): Three parts, i.e., URL, title, and text description, are consisted of Snippets datasets. For each query phrase entered into the Google search engine, the top 20 or 30 web search snippets are collected. Then, the class labels of the collected search snippets are assigned the same class labels as the uttered phrases.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 4. Parameter influence of $M$ on all three datasets.

Fig. 5. Impact of Support Set on AG’s News and Snippets dataset.

TMN News\textsuperscript{6}: Similar to AG’s News dataset, TMN News dataset is also a collection of English news from popular newspaper Websites, which contains 32 K English news with seven categories. It is worth mentioning that, like the AG’s News dataset, only news titles are selected in the experiments.

Twitter\textsuperscript{7}: The sample tweets are downloaded from the Natural Language Toolkit (NLTK) package, which contains 20000 tweet samples retrieved from the Twitter Streaming API. The dataset contains 2 main categories including 5000 tweets with negative sentiments, and 5000 tweets with positive sentiments.

Ohsumed\textsuperscript{8}: The OHSUMED dataset is a set of 348,566 references from medline, the online medical information database, consisting of titles and/or abstracts from 23 medical journals over five years (1987–1991). The dataset contains 23 categories such as Virus Diseases, Parasitic Diseases, Neoplasms, etc.

In the experiments, the manual templates are used due to they are proven to be competitive with or better than the auto-generated templates [42], and the construction of manual templates is simpler. To test the influence of different topic templates, further experiments are conducted on some datasets in Section IV-E.

B. Compared Methods

To verify the effectiveness of our PLST, we conduct the baseline approaches as the BERT model with fine-tuning, regular prompt-tuning, Knowledgeable Prompt-tuning for short text classification.

- **Label-Name-Only Text Classification (LOTClass) [51]**: LOTClass aims to find category-indicative words by PLMs from the whole unlabeled corpus, then the masked category prediction task is conducted based on these obtained words.

- **Heterogeneous Graph Attention Networks (HGAT) [31]**: HGAT is a heterogeneous information network for incorporating additional information and their relations from the open Knowledge Base, which can address the semantic sparsity problem in short text classification.

- **BERT + Fine-tuning (BERT + FT) [52]**: This method first obtains the hidden embedding of the [CLS] token through Bert, and then inputs this hidden embedding into the classification layer to predict.

- **Regular prompt-learning (PL) [39]**: The regular prompt-learning fills the input statements into a hand-crafted template, and only the category name is used to form the label word space. For fairness, the hand-crafted templates in PL are consistent with our method.

- **Pattern Exploiting Training (PET) [45]**: This method consists of defining pairs of cloze question patterns and verbalizers to stimulate rich knowledge in PLMs. The models are fine-tuned for all pattern-verbalizer pairs and then applied to create large annotated datasets on which standard classifiers can be trained.

- **Unified Prompt Tuning (UPT) [53]**: This method explicitly captured general prompting semantics from non-target datasets. The unified paradigm Prompt-Options-Verbalizer is first proposed for joint prompt learning, then a auxiliary Knowledge-enhanced Selective MLM is proposed to capture task-invariant prompting knowledge.

- **Knowledgeable Prompt-tuning method (KPT) [23]**: The KPT method expands the verbalizer in prompt-tuning with the external Knowledge Base. The verbalizer construction, refinement, and utilization are used to incorporate external knowledge for prediction.

- **KPT_Pro**: For fairness comparison of KPT and our PLST, we introduce the same external knowledge graph for KPT, i.e., the Probase is used in KPT to replace with the Wikipedia in the original paper, we call it KPT_Pro in the experiments.

- **PLST_sup**: In the verbalizer refinement, we conduct additional four strategies for capturing more aspects of the suitability of the expanded words. Therefore, PLST_sup is a simple version of PLST, which only take the distance $\text{dist}(V_y, y)$ between each expanded label words set and each label name $y$ into account.

\textsuperscript{6}http://acube.di.unipi.it/tmn-dataset/
\textsuperscript{7}https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/twitter_samples.zip
\textsuperscript{8}https://huggingface.co/datasets/ohsumed
C. Experiment Settings

In our experiments, we conduct 5, 10, 20-shot as k-shot learning experiments for all the prompt learning-based models, such as PL, PET, P-Tuning, UPT, KPT, KPT_Pro, PLST_sup, and PLST. Specifically, k and 1000 instances of each category are sampled from the original training set to form a new training set and a support set respectively. To address the potential impact of limited training samples on baseline methods, we created different training sets for neural network-based (HGAT) and fine-tuning methods (BERT+FT). The number of training samples was set to 1000/2000/4000, 800/1600/3200, 800/1600/3200, 800/1600/3200, and 1150/2070/4050 as corresponding to k-shot for the datasets AG’s News, Snippets, TMN News, Twitter, and Ohsumed, respectively.

Furthermore, for the BERT model, we utilized the bert-base-uncased version and employed a batch size of 32 for training over 10 epochs. Concerning the HGAT model, we followed the model parameters of the original paper, and we established a hidden dimension to 512, graph neural network layers as 2, a learning rate to 0.005, and a dropout rate to 0.8. Regarding PET, we adopted the parameters of the original model, we encompassed a learning rate to 1e-5, a batch size to 16, and a maximum sequence length to 256. To account for the disparities of certain datasets, we doubled the training samples in each iteration to enhance the diversity of the training data.

For prompt learning-based models (PL, PET, P-Tuning, UPT, KPT, KPT_Pro, PLST_sup, and PLST), we utilized the xlm-roberta-large [54] as PLMs. During training, we set the dropout to 0.5 for preventing overfitting, the learning rate is set to 5e-5, and the batch size is set to 32 with a weight decay of 1e-5. We conducted validation steps to fine-tune the hyper-parameters of the model. In the experiments, we set the epoch to 10 to ensure full training and stable results. The Adam optimizer was used to optimize the model’s parameters. Notably, in the KPT_Pro method, we leveraged the Probase knowledge graph to acquire external concepts for constructing the verbalizer.

It is worth mentioning that all experimental results are obtained by repeating five experiments and taking the average value. The OpenPrompt [22] is adopted for the implementation of all prompt-based methods such as PL and KPT. For LOTClass, we conducted the source code and used the default parameters as reported in [51].

Moreover, all experimental results were conducted on a server with an NVIDIA GeForce RTX 3090 Founders Edition, an Intel(R) Core(TM) i9-10980XE CPU running at 3.00 GHz, and 125 GB of memory. For conducting the experiments, we employed Python version 3.9.16 in conjunction with pytorch-cuda version 11.7. The accuracy and F1-score are adopted in all the experiments as the classification metrics.

D. Experimental Results

All the experimental results on five datasets are recorded in Table III. The following insightful observations can be listed from the experimental results:

(1) Generally, as the experiments vary from 5-shot to 20-shot, the performance of all fine-tuning methods has improved, which reveals the increase of labeled instances number can improve the results of short text classification.

(2) The prompt-learning methods can achieve better results than HGAT in most cases on datasets besides Ohsumed dataset. The results validated the effectiveness and superiority of the prompt-tuning model on text classification even in the scenario of few-shot learning. It is worth mentioning that HGAT can achieve substantial performance in Ohsumed dataset, we believe that the graph attention networks can address the semantic sparsity problem in multi-class short text classification.

(3) Even though we only employ few-shot learning instead of large unlabeled training data, our PLST can still achieve substantial performance than LOTClass in all five datasets. The results indicated that the concept retrieval from Probase and verbalizer construction by four additional strategies can find more accurate and more reliable expanded words than LOTClass.

(4) Our method and KPT achieve more desirable performance than other fine-tuning methods in most cases, such as regular prompt-learning, PET, P-Tuning, and UPT. The results indicate the effectiveness of expanding the verbalizer in prompt-tuning with the external Knowledge Base, and the introduced concepts can effectively address the problem of data sparsity for short text. It is worth mentioning that P-Tuning can achieve fairly good performance in all the datasets, which will be the further research direction in our further works.

(5) The results of the KPT are not stable especially in the TMN News dataset, which shows that KPT cannot incorporate appropriate knowledge in some datasets for short texts. Moreover, we can observe that the performance of KPT is even worse than regular prompt-learning on the TMN News dataset, it further demonstrates that the expanded label words by KPT may have some bias and do not have enough coverage to address the high ambiguity problem in short texts.

(6) Even with the same knowledge graph as Probase, we can observe that our PLST still achieved better results than KPT_Pro on all five datasets. The results validated the effectiveness of our proposed concept retrieval methods and verbalizer construction strategies. Notably, the utilization of Probase is one of the contributions of our PLST, the experiments on KPT_Pro further demonstrated the contribution and effectiveness of our proposed PLST.

(7) In all the cases on five datasets, the performance of PLST consistently outperforms PLST_sup, the results demonstrate that the proposed four strategies can find more high-quality words in verbalizer, and the union of these words can effectively address the feature sparsity and high ambiguity problems.

(8) Overall, our PLST performs best in most cases, which validates the effectiveness of incorporating external knowledge for extending label word space to address the inherent problems of short text classification. It is worth mentioning that the Ohsumed dataset is a multi-class classification dataset including 23 main categories, which leads to a lower performance in text classification. Furthermore, it should be noted that our method can achieve stable performance in all five datasets, which shows the effectiveness of knowledgeable expansion with concept retrieval and verbalizer construction.
TABLE III
ACCURACY AND F1-SCORE RESULTS ON FIVE DATASETS

| Shot | Method  | AG’s News | Snippets | TMN News | Twitter | Observed |
|------|---------|-----------|----------|----------|---------|----------|
|      |         | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| LOTClass | HGAT | 63.25 | 61.66 | 59.69 | 58.11 | 76.05 | 77.32 | 83.90 | 83.89 | 29.97 | 28.60 |
|       | BERT + FT | 63.46 | 63.04 | 72.10 | 71.33 | 60.46 | 59.80 | 55.18 | 54.99 | 26.70 | 26.32 |
|       | PET | 32.30 | 31.19 | 35.30 | 34.17 | 31.31 | 30.07 | 45.13 | 45.13 | 11.52 | 10.15 |
|       | PL | 74.48 | 74.31 | 72.54 | 71.56 | 62.31 | 61.66 | 69.60 | 68.66 | 17.74 | 16.73 |
|       | P-Tuning | 47.30 | 73.38 | 72.01 | 71.04 | 66.54 | 66.96 | 64.93 | 62.86 | 24.73 | 22.66 |
|       | UPT | 72.03 | 71.67 | 70.78 | 70.58 | 65.42 | 62.91 | 62.63 | 60.04 | 18.58 | 18.25 |
|       | KPT | 75.56 | 75.30 | 73.91 | 73.84 | 67.96 | 67.96 | 72.46 | 72.68 | 15.34 | 13.16 |
|       | KPT_Pro | 75.86 | 74.71 | 82.36 | 81.93 | 66.11 | 66.31 | 75.93 | 75.81 | 27.37 | 26.65 |
|       | PLST | 77.14 | 77.11 | 82.80 | 82.25 | 68.96 | 68.13 | 77.06 | 76.92 | 30.41 | 29.49 |
|       | PLST_sup | 75.83 | 74.81 | 82.03 | 83.09 | 71.88 | 72.03 | 71.86 | 71.77 | 29.11 | 26.37 |
|       | PLST | 82.59 | 82.55 | 82.72 | 83.10 | 75.53 | 75.48 | 80.96 | 80.96 | 31.57 | 30.22 |

The best results are marked in bold.

TABLE IV
RESULTS OF THE INTERSECTION AND UNION ON DIFFERENT STRATEGIES IN THREE DATASETS

| Shot | Methods  | AG’s News | Snippets | TMN News |
|------|----------|-----------|----------|----------|
|      | intersection | 77.46 | 83.68 | 70.11 |
|       | union     | 82.59 | 83.72 | 75.53 |
|      | intersection | 81.47 | 84.14 | 74.85 |
|       | union     | 83.07 | 84.45 | 77.64 |
|      | intersection | 82.23 | 86.13 | 76.38 |
|       | union     | 84.70 | 87.55 | 78.07 |

The best results are marked in bold.

E. Influence of Verbalizer Construction Strategies

In this part, we investigate the influence of the different verbalizer construction strategies to the performance of our PLST. Table IV shows the results of PLST with intersection and union on different strategies, the intersection refers to the selection of expanded words that appear simultaneously in different strategies. The union refers to putting all the expanded words selected by different strategies into the verbalizer, which is conducted in our PLST. Compared to the intersection method, the observed results from Table IV show that the union of different strategies can improve the performance of short text classification, which demonstrated the intersection may remove some helpful expanded words from the verbalizer and the effectiveness of the union in retaining high-quality words.

Meanwhile, we tried different variants on these strategy integrations in verbalizer construction. BP, FT, FS, CI, and CR are denoted for Prediction by BERT, FastText Semantic similarity, Frequency words selection, Context Information, and Concept Retrieval, respectively. The comparative impact of the different variants is shown on Table V. From these results, we can observe that the strategy of Concept Retrieval has played a crucial role in the experiments, which validated the main contribution of our paper, i.e., the expanded label words by taking both the short text itself and category names into consideration is more accurate and more reliable than other label word space in the existing methods. Moreover, the results further show that the model can obtain more comprehensive information in the prediction with pre-selected label words, and the less bias introduced in the verbalizer, the more accurate the detection results.
TABLE V
INFLUENCE OF DIFFERENT VARIANTS ON THE STRATEGY INTEGRATIONS OF LABEL WORD SPACE EXPANSION

| CR | BP | FT | FS | CI | AG’News | Snippets | TMN News | Twitter | Observed |
|----|----|----|----|----|---------|----------|----------|---------|----------|
| ✓  | ✓  | ✓  | ✓  | ✓  | 80.71   | 83.06    | 74.96    | 86.13   | 33.57    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 81.75   | 83.55    | 75.35    | 88.47   | 31.06    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 82.51   | 85.08    | 74.78    | 86.84   | 32.71    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 81.39   | 84.95    | 75.82    | 87.91   | 32.09    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 80.77   | 82.50    | 74.89    | 83.24   | 32.85    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 83.51   | 86.84    | 76.96    | 93.22   | 34.92    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 84.26   | 86.59    | 76.42    | 93.34   | 35.13    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 83.77   | 86.40    | 77.10    | 92.18   | 35.30    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 82.48   | 85.26    | 77.28    | 89.05   | 34.61    |
| ✓  | ✓  | ✓  | ✓  | ✓  | 82.32   | 84.42    | 76.55    | 91.25   | 34.09    |

The experiments are conducted with accuracy(%) over five datasets. The best results are marked in bold.

TABLE VI
DIFFERENT TEMPLATES ON THREE DATASETS

| Dataset | id | Templates |
|---------|----|-----------|
| AG news | 1  | X [mask] news: X |
|         | 2  | X This topic is about [mask] |
|         | 3  | The category of X is [mask] |
|         | 4  | Topic[mask] X |
| Snippets| 1  | X is about [mask] |
|         | 2  | X This topic is about [mask] |
|         | 3  | The category of X is [mask] |
|         | 4  | The topic of X is [mask] |
| TMN News| 1  | A [mask] news: X |
|         | 2  | X This topic is about [mask] |
|         | 3  | X is about [mask] |
|         | 4  | The topic of X is [mask] |

F. Influence of the Templates

The templates have always been one of the important factors that affect the effectiveness of prompt-learning methods, we list all the templates used in our experiments in Table VI. The 5, 10, and 20-shot experimental results of regular prompt-learning (PL), knowledgeable Prompt-tuning method (KPT), PLST_sup, and PLST with four templates on three datasets are reported in Tables VII, VIII, and IX, the best results of different method in each dataset are marked in bold. The results reveal that our method has a significantly better performance on three datasets than PL and KPT, our method can achieve substantial and consistent performance in all templates. In addition, we have observed that the experimental results of KPT are even worse than PL in the TMN News and Snippets datasets, which shows that KPT cannot keep stable with the change of templates.

G. Parameter Sensitivity

In this section, we investigate the influence of parameters in our proposed method, including top N in concept retrieval and the selected top M words in verbalizer refinement. It is worth mentioning that M indicates the number of expanded words closest to the label name, which does not take the additional four strategies into account. When we change one parameter, the rest others are fixed in the experiment. N and M are sampled from the set {1, 2, 3, 4, 5, 6, 7} and {30, 40, 50, 60, 70} respectively. All the results conducted on 5-shot are reported in Figs. 3 and 4 respectively, and we set N = 5 and M = 50 to get the best and most stable results. It is worth mentioning that the results are declining when M = 80, which indicates that more label words with weak relevance incur a performance loss, that is also why KPT achieved unsatisfying results in some datasets.

H. Influence of the Support Set

In our experiments, 1000 instances of each category are sampled from the original training set to form a support set. In this section, to verify the impact of the support set, we altered the size from the set {0, 256, 512, 752, 1000, 1256, 1512}, the 10-shot experimental results with the first template on three datasets can be seen in Fig. 5. From the observation of these results, our method keep the encouraging result on different sizes of support sets, which demonstrated the stability and effectiveness of our method.
In this paper, we propose a prompt-learning method for short text classification. Taking the special characteristics of short text into consideration, the method can consider both the short text itself and the class name during expanding label word space. The proposed method retrieves top $N$ concepts from the open Knowledge Graph and conducts additional four strategies for verbalizer construction. The experiments show the effectiveness of our method.

In the future, we will extend our research work from the following two directions. One is exploring better methods for automatic template construction on short text. The other is to incorporate more auxiliary information from external knowledge for some other tasks.

**REFERENCES**

[1] A. Sun, “Short text classification using very few words,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 1145–1146.

[2] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, “Pre-training with whole word masking for chinese bert,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3504–3514, 2021.

[3] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 841–842.

[4] M. Faqeeh, N. Abdulla, M. Al-Ayyoub, Y. Jararweh, and M. Quwaider, “Cross-lingual short-text document classification for facebook comments,” in *Proc. Int. Conf. Future Internet Things Cloud*, 2014, pp. 573–578.

[5] Z. Liu, W. Yu, W. Chen, S. Wang, and F. Wu, “Short text feature selection for micro-blogging,” in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, 2010, pp. 1–4.

[6] Y. Kim, “Convolutional neural networks for sentence classification. corr abs/1408.5882 (2014),” 2014, arXiv:1408.5882.

[7] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2267–2273.

[8] J. Chen, Y. Hu, J. Liu, Y. Xiao, and H. Jiang, “Deep short text classification with knowledge powered attention,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6252–6259.

[9] J. Xu and Y. Cai, “Incorporating context-relevant knowledge into convolutional neural networks for short text classification,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 10067–10068.

[10] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning–based text classification: A comprehensive review,” *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2021.

[11] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, “Data augmentation for bert fine-tuning in open-domain question answering,” 2019, arXiv:1904.06652.

[12] R. Weng, H. Yu, S. Huang, S. Cheng, and W. Luo, “Acquiring knowledge from pre-trained model to neural machine translation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9266–9273.

[13] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, and X. Wu, “Lexical simplification with pretrained encoders,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8649–8656.

[14] Y. Goldberg, “Assessing BERT’s syntactic abilities,” 2019, arXiv:1901.05287.

[15] X. Ma, Z. Wang, P. Ng, R. Nallapati, and B. Xiang, “Universal text representation from BERT: An empirical study,” 2019, arXiv:1910.07973.

[16] G. Jawahar, B. Sagot, and D. Seddah, “What does bert learn about the structure of language?,” in *Proc. Ann. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3651–3657.

[17] H. Xu et al., “Pre-trained models: Past, present and future,” 2021, arXiv:2106.07139.
Ye Wang received the BS degree from the Jiangxi Institute of Applied Science and Technology. He is currently working toward the graduate degree with the School of Information Engineering, Yangzhou University. His research interests are short text classification and fake news detection.

Jipeng Qiang (Member, IEEE) received the PhD degree in computer science and technology from the Hefei University of Technology, in 2016. He is currently an associate professor with the School of Information Engineering, Yangzhou University. He was a PhD visiting student with the Artificial Intelligence Lab, University of Massachusetts Boston from 2014 to 2016. He has published more than 40 papers, including AAAI, IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Knowledge Discovery from Data, and IEEE/ACM Transactions on Audio, Speech, and Language Processing. His research interests mainly include natural language processing and data mining.

Xindong Wu (Fellow, IEEE) received the BS and MS degrees in computer science from the Hefei University of Technology, China, in 1984 and 1987 respectively, and his PhD degree in artificial intelligence from the University of Edinburgh, Britain, in 1993. He is a senior research scientist with the Research Center for Knowledge Engineering, Zhejiang Lab, China. He is a foreign member of the Russian Academy of Engineering, and a fellow of AAAS. His research interests include data mining, knowledge engineering, Big Data analytics, and marketing intelligence.