Multi-Mask Label Mapping for Prompt-Based Learning

Jirui Qi[1], Richong Zhang[1,2*], Jaein Kim[1], Junfan Chen[1], Wenyi Qin[1], Yongyi Mao[3]

[1]SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada
{qijr,zhangrc,chenjf}@act.buaa.edu.cn, {jaein,wenyiqin}@buaa.edu.cn, ymao@uottawa.ca

Abstract

Prompt-based Learning has shown significant success in few-shot classification. The mainstream approach is to concatenate a template for the input text to transform the classification task into a cloze-type task where label mapping plays an important role in finding the ground-truth labels. While current label mapping methods only use the contexts in one single input, it could be crucial if wrong information is contained in the text. Specifically, it is proved in recent work that even the large language models like BERT/RoBERTa make classification decisions heavily dependent on a specific keyword regardless of the task or the context. Such a word is referred to as a *lexical cue* and if a misleading lexical cue is included in the instance it will lead the model to make a wrong prediction. We propose a multi-mask prompt-based approach with Multi-Mask Label Mapping (MMLM) to reduce the impact of misleading lexical cues by allowing the model to exploit multiple lexical cues. To satisfy the conditions of few-shot learning, an instance augmentation approach for the cloze-type model is proposed and the misleading cues are gradually excluded through training. We demonstrate the effectiveness of MMLM by both theoretical analysis and empirical studies, and show that MMLM outperforms other existing label mapping approaches.

Introduction

With the popularity of pre-trained language models like GPT-3 in the NLP domain (Brown et al. 2020), prompt-based learning has demonstrated its excellent ability to handle numerous few-shot tasks (Liu et al. 2021), such as sentiment classification (Gao, Fisch, and Chen 2021), text classification, and commonsense reasoning (Wei et al. 2022). Among them, prompt-based learning with Cloze-type Language Models (CLMs)[1] have shown their excellence on few-shot classification tasks (Gao, Fisch, and Chen 2021; Hu et al. 2022). Recent works confirm that prompt-based learning significantly outperforms the traditional fine-tuning approaches (Gao, Fisch, and Chen 2021; Hu et al. 2022; Wang, Xu, and McAuley 2022) which adds extra classification networks on the top of CLMs. But the randomly initialized parameters in these classification networks cannot be trained well due to scarce labeled instances (Brown et al. 2020).

Some previous works have demonstrated that prompt-based learning can effectively exploit the rich knowledge in CLM, which is compressed in CLM's parameters during the pre-training process (Trinh and Le 2018; Davison, Feldman, and Rush 2019; Petroni et al. 2019). The vanilla prompt-based approach consists of two components, namely text reformation and label mapping. In the text reformation process, input texts are wrapped by a pre-defined classification template with a {mask} slot. For example, for sentiment classification, the text 'Boring starting but overall ok and worth watching.', is wrapped into a template '{TEXT} It was {mask}.'[2]. After being encoded with CLM, the hidden vector of {mask} is used to calculate the word-occurrence probability that each word in the vocabulary is filled in the {mask} slot based on its context.

Bridging the gap between the word-occurrence probability with the ground-truth label is significant in the label mapping process. To achieve this, verbalizers are proposed in recent work to assign one or multiple representative word(s) to each label (Gao, Fisch, and Chen 2021; Cui et al. 2022; Hu et al. 2022). With the help of verbalizers, the label prediction problem is transferred to comparing the averaged word-occurrence probability of each label at the {mask} slot.

In existing label mapping models, only a single context is considered for filling each {mask} slot. This could lead to a wrong prediction if a misleading lexical cue is contained in the given sentence. In specific, it is studied in recent work that many large language models like BERT or RoBERTa are often heavily dependent on specific lexical cues for decision making (Kavumba, Takahashi, and Oda 2022). For example, in a wrapped sentence 'Boring starting but overall ok and worth watching. It was {mask}.' , the lexical cues are *boring*, *ok* and *worth*. From the human perspective, we can easily judge that the sentence should be classified as a positive label rather than a negative label. However, a language model may consider *boring* as the greatest impact on the

*Corresponding author

[1]More commonly, they are called Masked Language Models (MLMs), but here we use the term CLM to distinguish their abbreviation from MMLM.

[2]We will omit the {TEXT} symbol in the template for the rest of this paper for clarity since we only adopt the concatenation operation in wrapping.

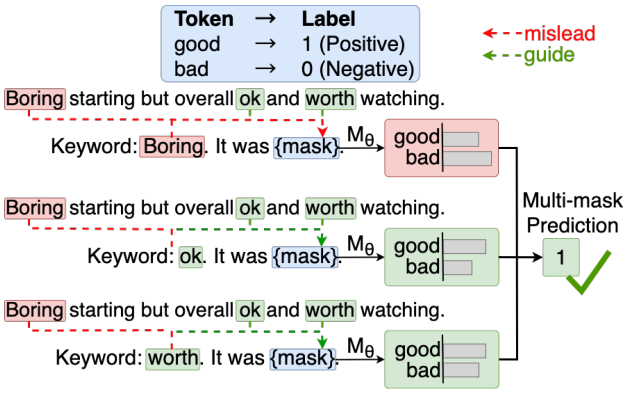

Figure 1: Illustration of 2-way sentiment classification. Vanilla label mapping with a single mask is easy to make a wrong prediction due to the misleading lexical cue 'boring' in the input sentence. In contrast, MMLM generates and utilizes multiple mask slots to alleviate the issue.

classification as it is more directly and emotionally expressive than the other two. In this case, the misleading lexical cue *boring* will lead the model to fill the {mask} slot with a wrong word, which will map the sentence to a negative label and further decrease the classification accuracy.

Believing that more correct information can make the impact of wrong information indistinct, we propose a multi-mask label mapping (MMLM) scheme to enlarge the effect of correct lexical cues to reduce the effect of the misleading cues. MMLM first uses a prompt-based augmentation approach to augment each sentence into a set of augmented texts by automatically extracting lexical cues (keywords) with a stimulation template 'Keyword: {mask}.', filling them in the stimulation template 'Keyword: {cue}.' and concatenating them with the original sentence separately. Given these augmented instances, it next wraps each of them with the classification template 'It was {mask}.' and feed them into perturbed CLM to form multiple prompt-based classifiers. Therefore, each classifier utilizes different contexts with various biases of lexical cues to make relatively independent predictions. Furthermore, the model not only reduces the impact of misleading cues but also optimizes the keyword extractor itself to progressively identify more correct lexical cues during training.

Through experiments, we confirm the effectiveness of MMLM on AG's News, IMDB, Amazon, DBPedia, and Yahoo datasets. The experimental results show that MMLM outperforms existing label mapping methods, some of which leverage external knowledge bases (KBs) outside the scope of the model. In summary, the contributions of this paper can be summarized as follows:

- We propose a multi-mask label mapping method for few-shot classification problem. Theoretical analysis shows the effectiveness of our prompt-based augmentation and multi-mask scheme in the few-shot scenario.
- We demonstrate that the effect of misleading lexical cues in classification can be reduced if the model is allowed to learn multiple context information of different lexical cues with the help of the proposed instance augmentation approach.
- We show that the proposed label mapping model outperforms SOTA by extracting the compressed knowledge in the pre-trained language without needing to involve external KBs.

Related Works

Existing Label Mappings Mainstream label mapping methods are divided into four categories. The first is *Manual Label Mapping* (Schick and Schütze 2021) which manually defines one representative word for each class and uses word-occurrence probability. *Search-based Label Mapping* (Gao, Fisch, and Chen 2021) tries to automatically generate the representative words for each class and also focuses on the fill-in probability of these words at the single mask slot in the classification template. *Soft Label Mapping* (Hambardzumyan, Khachatrian, and May 2021), on the other hand, tries to learn a soft class representative for each class and calculates the label-prediction probability by multiplying the word-occurrence probability with each soft class representative. Finally, *External Knowledge Label Mapping* (Hu et al. 2022) exploits the external KB to find multiple words to represent each class label and calculates the label-prediction probability by averaging the word-occurrence probability of the representative words of each class.

Some search-based label mapping methods (Schick, Schmid, and Schütze 2020; Shin et al. 2020) and external knowledge label mapping try to analyse the context semantics from different aspects by averaging the word-occurrence probability of multiple words to stand for the label-prediction probability. However, they only use the hidden vector of one mask slot, which contains monotonous contexts semantics. If the word-occurrence probability at the mask slot is misled by some ambiguous words, the classification results will also go to a wrong direction.

Lexical Cues While there are many factors for a sentiment classification model to determine a sentence label in label mapping methods, surprisingly, only a few words in the sentence play a major role in decision making. For example, in a sentence with positive sentiment 'The movie is worth watching.', the word 'worth' is a strong cue for the model to predict the sentence as label 'Positive'. In fact, it has been analyzed that even the large language models, such as BERT-based models rely heavily on exploiting such lexical cues to determine the semantic label regardless of the task (Niven and Kao 2019; Kavumba et al. 2019). The features of lexical cues include lexical overlap heuristic (McCoy, Pavlick, and Linzen 2019), frequent words based on statistics (Niven and Kao 2019), or sentence style (Trichelair et al. 2019).

Because the model may make decisions based on the cues regardless of the context or the task, they are in many studies referred to as *superficial cues* (Kavumba, Takahashi, and Oda 2022). While some are misleading, lexical cues still guarantee high performance to some extent as proved in recent work. In this paper, rather than trying to exclude the misleading cues, we extract several lexical cues from the

given text and exploit them to predict the label.

Generation-Based Augmentation With Prompt Precedent works mainly demonstrate the effectiveness of using a proper template to stimulate the knowledge in generation-type pre-trained language models like GPT-3. Typically, an augmenting text x' is generated by feeding the prefix x with a stimulation template t into GPT-3. By processing $x \oplus t \oplus x'$, an enhanced text with additional semantics is formed for the downstream tasks. This approach has been proved effective in multiple few-shot tasks (Wei et al. 2022; Wang et al. 2022; Kojima et al. 2022; Zhou et al. 2022; Li et al. 2022).

It is a parameter-efficient augmentation method as it introduces no extra networks, and only a small amount of new parameters is required for the stimulation template t. However, such an augmentation method is not yet generalized to cloze-type pre-trained language models like BERT and RoBERTa. There are two main reasons for this. First, generation-based models generate the text based on their huge amount of parameters. Some works show that if the amount of parameters is reduced, the capability of generating the text with correct semantics is weakened (Wei et al. 2022). Secondly, cloze-type language models do not have the generation ability like the generative models. Instead, their capability is only to fill the blanks in the sentence. Therefore, it is not easy to extend the augmentation work to cloze-type language models.

Multi-Mask Label Mapping

In this section, we introduce the framework of MMLM. It consists of two interacting modules, cloze-based augmentation with prompt and multi-mask scheme. We first introduce the preliminary definitions and notations for few-shot classification in the setting of prompt-based learning. Then, we describe the probabilistic architecture of MMLM. Finally, we elaborate on these two modules in more detail and demonstrate the effectiveness of the proposed multi-mask model with probabilistic derivation.

Problem Definition

For a N-way K-shot few-shot text classification, the input text is defined as X which contains $N * K$ elements as it consists N classes of instances and each class contains K instances. The corresponding label set is denoted as Y which also contains $N * K$ elements and the label space is defined as $\mathcal{Y}$. For example, for sentiment classification, there is $\mathcal{Y} = \{0, 1\}$ where '0' represents 'negative' and '1' represents 'positive'. The pre-trained cloze-type language model is defined as M_θ where θ stands for its parameters. The vocabulary is defined as $\mathcal{V}$ and the word-occurrence probability $\mathbf{P}$ is defined as a $|\mathcal{V}|$-dimensional vector, where each dimension corresponds to the occurrence probability of a token in $\mathcal{V}$, thus $\sum_{v \in \mathcal{V}} \mathbf{P}[v] = 1$.

The predicting target in few-shot classification task with vanilla prompt-based learning is to maximize

$$\begin{aligned} Q(x) &= \sum_{i=1}^{N*K} P_{M_\theta}(y_i|x_i) \\ &= \sum_{i=1}^{N*K} P_{M_\theta}(\{mask\}_T = r(y_i)|x_i \oplus T) \end{aligned} \tag{1}$$

where $x_i \in X$ and $y_i \in Y$. T is a template and $r(\cdot)$ is a manually designed verbalizer which assigns a representative word to each label. For instance, in sentiment classification it assigns the word 'good' to the label '1' and the word 'bad' to the label '0'.

Probabilistic Architecture

Inspired by the previous exploration on generation-based prompt-based augmentation (Wei et al. 2022; Wang et al. 2022; Kojima et al. 2022; Li et al. 2022), we further attempt to apply an augmentation method on cloze-type pre-trained language models. To enlarge the influence of each lexical cue, we propose a cloze-type prompt-based augmentation with prompt to highlight the different keywords in the given input text x. Specifically, the input text x is concatenated with a stimulation template $t :=' Keyword : \{mask\}_t.'$. Then, the module calculates the word-occurrence probability of the word k at $\{mask\}_t$ slot

$$P_{M_\theta}(k|x \oplus t) = P_{M_\theta}(\{mask\}_t = k|x \oplus t). \tag{2}$$

The words that are semantically relevant to the context are likely to be chosen as keywords. MMLM narrows the vocabulary $\mathcal{V}$ to $\mathcal{V}_x^n$ which contains top-n keywords that have the highest probabilities for text x as

$$\begin{aligned} \mathcal{V}_x^n &= \underset{k \in x}{\text{top-}n}[P_{M_\theta}(\{mask\}_t = k|x \oplus t)] \\ &= \{k_1, \ldots, k_n\}. \end{aligned} \tag{3}$$

The weight w_i of each $k_i \in \mathcal{V}_x^n$ and the set of weights are defined as

$$\begin{aligned} w_i &= \frac{exp(P_{M_\theta}(\{mask\}_t = k_i|x \oplus t))}{\sum_{j=1}^{n} exp(P_{M_\theta}(\{mask\}_t = k_j|x \oplus t))} \\ W_x &= \{w_1, \ldots, w_n\}. \end{aligned} \tag{4}$$

By replacing $\{mask\}_t$ with each $k_i \in \mathcal{V}_x^n$ and concatenating x with t and T, MMLM is able to generate n different augmented instances. Continually, each augmented instance embedding and a set of them are defined as

$$\begin{aligned} e_i &= g_\theta(x \oplus t(k_i) \oplus T) \\ E &= \{e_1, e_2, \ldots, e_n\}, \end{aligned} \tag{5}$$

where $t(k_i)$ is $\{mask\}_t$ being replaced by k_i and $g_\theta(\cdot)$ is the embedding layer of M_θ. Each individual $e_i \in E$ is further perturbed to increase the variability. Then a set of disturbed embeddings E' is obtained via

$$\begin{aligned} e_i' &= D(e_i) \\ E' &= \{e_1', e_2', \ldots, e_n'\}, \end{aligned} \tag{6}$$

where $D(\cdot)$ is the perturbation function.

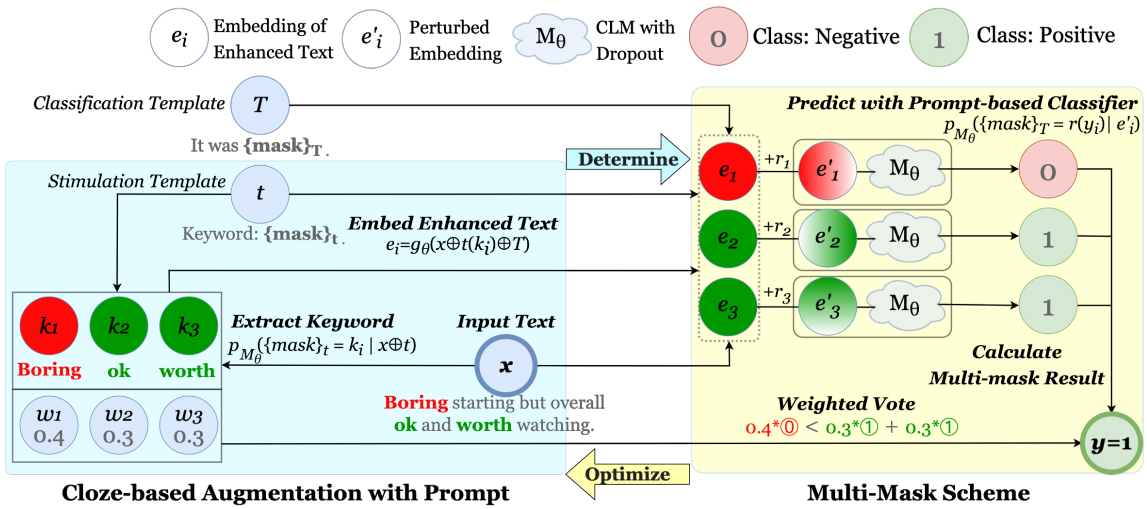

Figure 2: The workflow of Multi-Mask Label Mapping with Cloze-based Augmentation and Multi-Mask Scheme.

To predict y_i as the label of the i-th enhanced text, we use word-occurrence probability of $r(y_i)$ at $\{mask\}_T$ which is calculated by

$$P_{M_\theta}(y_i|e'_i) = P_{M_\theta}(\{mask\}_T = r(y_i)|e'_i). \quad (7)$$

Finally, the overall multi-mask prediction result is composed of n sub-predictions with the weighted votes $\mathcal{W}$ as

$$\begin{aligned} P^{\mathcal{W}}_{M_\theta}(y|x) &= \sum_{i=1}^{n} w_i * P_{M_\theta}(y|e'_i) \\ &= \sum_{i=1}^{n} w_i * P_{M_\theta}(\{mask\}_T = r(y)|e'_i), \end{aligned} \quad (8)$$

or with the plurality votes $\mathcal{P}$ as

$$\begin{aligned} P^{\mathcal{P}}_{M_\theta}(y|x) &= \frac{1}{n}\sum_{i=1}^{n} P_{M_\theta}(y|e'_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} P_{M_\theta}(\{mask\}_T = r(y)|e'_i). \end{aligned} \quad (9)$$

The two modules share the same parameter θ, which is updated during the iterations. Therefore, the number of misleading lexical cues will decrease along with the improvement of classification accuracy.

Cloze-Based Augmentation With Prompt

In order to ensure the stimulation capability of t and meanwhile reduce its influence to $\{mask\}_T$ in T, we set $t =$'$Keyword : \{mask\}_t$.' which only consists of four explicit tokens[3], namely 'Key', 'word', ':' and '.' . We deliberately select the tokens that are relatively neutral in classification but can work as lexical cues for $\{mask\}_t$.

MMLM concatenates x with t, and the hidden state of $\{mask\}_t$ in the last layer is obtained by

$$h_{\{mask\}_t} = M_\theta(x \oplus t). \quad (10)$$

[3]In RoBERTa, 'Keyword' is split into 'Key' and 'word'.

After mapping $h_{\{mask\}_t}$ to a $|\mathcal{V}|$-dimensional vector with pre-trained linear network L_θ, MMLM calculates the word-occurrence probability $\mathbf{P}$ at $\{mask\}_t$ by normalizing the logits with a softmax function

$$\mathbf{P_t} = \text{softmax}(L_\theta(h_{\{mask\}_t})) \quad (11)$$

which is used in Equation 3 for extracting top-n keywords. After separately filling each keyword $k_i \in \mathcal{V}^n_x$ into $\{mask\}_t$ and combining it with x and the classification template T, a set of embedding vector in Equation 5 is determined.

Multi-Mask Scheme

E is a set of the embedding vectors of the augmented texts of x as described in Equation 5. Since e_i is derived using T that already contains a mask slot $\{mask\}_T$ which is used for prediction in prompt-based learning, e_i can be directly handed over to CLM to make prediction. From another perspective, we can treat them as n prompt-based classifiers, each of which can predict the label of x using

$$f_i(x) = M_\theta(e_i) = M_\theta(x, t, k_i, T), \quad i = 1, \ldots, n \quad (12)$$

In order to maximize the independence between different prompt-based classifiers, we introduce the following three methods. First, as Equation 12, each classifier $f_i(\cdot)$ contains a unique keyword k_i, which works as a different lexical cue and guides the M_θ to calculate the corresponding word-occurrence probability as $\{mask\}_T$. Secondly, we introduce a perturbation function to disturb the embedding weight of each M_θ in $f_i(\cdot)$ with normally distributed random variables. For this method, Equation 6 is expanded as

$$e'_i = e_i + \alpha r_i, \quad (13)$$

where $r_i \sim N(0, 1)$ is a normally distributed random perturbation term with a weight α. Thirdly, we adopt two dropout layers in M_θ with a dropout probability of 10% to improve the independence between different prompt-based classifiers. One is for hidden vector in the forward propagation while the other is for attention weights.

Next, the last layer's hidden state of $\{mask\}_T$ is obtained via

$$h_{\{mask\}_T} = M_\theta(e'_i), \tag{14}$$

and the word-occurrence probability of all tokens in $\mathcal{V}$ at $\{mask\}_T$ is

$$\mathbf{P_T} = L_\theta(h_{\{mask\}_T}). \tag{15}$$

We focus on the words that appeared in the representative word set $\mathcal{R}$. For prompt-based classifier $f_i(\cdot)$, the predicting probability in Equation 7 can be further written as

$$\begin{aligned} P_{M_\theta}(y_i|e'_i) &= P_{M_\theta}(\{mask\}_T = r(y_i)|e'_i) \\ &= \frac{\mathbf{P_T}[r(y_i)]}{\sum_{l\in\mathcal{Y}} \mathbf{P_T}[r(l)]} \end{aligned} \tag{16}$$

and the multi-mask prediction of the label y for x can be calculated as n sub-predictions with the weighted votes as

$$\begin{aligned} P^{\mathcal{W}}_{M_\theta}(y|x) &= \sum_{i=1}^{n} w_i * P_{M_\theta}(\{mask\}_T = r(y)|e'_i) \\ &= \sum_{i=1}^{n} w_i * \frac{\mathbf{P_T}[r(y)]}{\sum_{l\in\mathcal{Y}} \mathbf{P_T}[r(l)]} \end{aligned} \tag{17}$$

or with the plurality votes as

$$\begin{aligned} P^{\mathcal{P}}_{M_\theta}(y|x) &= \frac{1}{n}\sum_{i=1}^{n} P_{M_\theta}(\{mask\}_T = r(y)|e'_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{P_T}[r(y)]}{\sum_{l\in\mathcal{Y}} \mathbf{P_T}[r(l)]} \end{aligned} \tag{18}$$

which are optimized during fine-tuning.

Theoretical Analysis of MMLM To further prove the efficiency of the multi-mask scheme, we conduct a theoretical illustration of MMLM with a 2-way sentiment classification using plurality vote as an example. Supposing both the prompt-based classifier $f_\theta(\cdot)$ and y are chosen from label set $\{0, 1\}$, the error rate of each prompt-based classifiers $f^i_\theta(\cdot)$ can be defined as

$$p(f_i(x) \neq y) = \epsilon. \tag{19}$$

Three different methods, different keywords, perturbation function and dropout layers are proposed to maximize the independence between different prompt-based classifiers in few-shot scenarios. For plurality vote, the multi-mask prediction of instance x can be expressed as[4]

$$F(x) = \begin{cases} 1, & \sum_{i=1}^{n} f_i(x) > \lfloor \frac{n}{2} \rfloor \\ 0, & \sum_{i=1}^{n} f_i(x) \leq \lfloor \frac{n}{2} \rfloor. \end{cases} \tag{20}$$

Thus, the error rate of multi-mask classification by n prompt-based classifiers is calculated as

$$P(F(x) \neq y) = \sum_{q=0}^{\lfloor n/2 \rfloor} C^q_n (1-\epsilon)^q \epsilon^{n-q}. \tag{21}$$

[4]Consistent with the previous sections, n is the number of prompt-based classifiers.

Let Z be the number of correct prediction made by the classifiers, the error probability of the multi-mask scheme is calculated as

$$\begin{aligned} P(F(x) \neq y) &= P(Z \leq \lfloor \frac{n}{2} \rfloor) \\ &\leq P(Z - E(Z) \leq -n\frac{(1-2\epsilon)}{2}). \end{aligned} \tag{22}$$

Further with Hoeffding's inequality (Hoeffding 1994), there is

$$P(Z - E(Z) \leq -n\frac{(1-2\epsilon)}{2}) \leq \exp(-\frac{1}{2}n(1-2\epsilon)^2), \tag{23}$$

thus the upper bound of the error probability is

$$P(F(x) \neq y) \leq \exp(-\frac{1}{2}n(1-2\epsilon)^2). \tag{24}$$

Therefore, ideally the error rate of the prompt-based multi-mask model decreases with the raising of n, the number of classifiers. When $n \to +\infty$ the error probability converges to 0. This result implies that exploiting multiple lexical cues is better than one single contextual semantic under the setting of prompt-based learning.

Experiments

Datasets and Implementation Details

We conduct experiments on K=1/5/10/20 in K-shot scenarios on five datasets and average the accuracy over five random seeds for the evaluation. In order to eliminate the performance fluctuation caused by different templates T (Gao, Fisch, and Chen 2021), we use a fixed template provided by OpenPrompt (Ding et al. 2022) to accurately observe the performance of different label mapping methods. For the same reason, we uniformly use RoBERTa-large (Liu et al. 2019) as a pre-training model with a batch size of 2 and 10 fine-tuning epochs.

Considering memory and text-length restrictions, we use $n = 15$ of extracted keywords for AG's News and $n = 5$ for the rest datasets. The memory usage is controlled within 32 GB. The maximum length for truncating each input is 512 for IMDB/Yahoo/Amazon and 128 for DBPedia/AG's News.

Baselines

As mentioned earlier, we mainly compare MMLM with the traditional [CLS] fine-tuning by inputting the hidden vector of [CLS] into a classification layer, as well as the four mainstream label mapping methods. Among them, KPT engages *external knowledge bases* which we use *italics* to highlight. For more convincing results, we employ Manual Label Mapping (vanilla), Search-based Label Mapping, Soft Label Mapping, and External Knowledge Label Mapping (KPT) (Hu et al. 2022) with OpenPrompt (Ding et al. 2022) using PyTorch framework (Paszke et al. 2019). For pre-trained cloze-type language model implementation, we use the interfaces provided by HuggingFace (Wolf et al. 2020) and AdamW optimizer (Kingma and Ba 2015).

Method	AG's News (4-way)				IMDB (2-way)				Amazon (2-way)				DBPedia (14-way)				Yahoo (10-way)			
	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20
CLS FT	22.3	39.2	78.4	84.9	50.8	52.4	79.8	80.3	51.2	53.3	81.7	84.6	10.3	89.0	95.8	96.3	10.8	23.2	46.1	53.7
Manual	78.4	83.1	85.1	86.4	91.5	92.0	92.0	93.6	90.6	93.7	94.0	94.3	93.0	95.2	95.8	96.2	48.2	54.6	57.7	59.6
Search	48.3	74.6	84.1	86.1	67.5	88.3	92.6	92.6	63.1	87.3	93.8	94.3	71.5	93.7	95.7	96.0	21.6	43.0	51.2	57.0
Soft	78.6	83.5	85.4	86.6	90.7	89.6	92.9	93.5	89.1	93.6	93.9	94.1	93.9	95.2	96.2	96.3	48.8	55.3	58.8	60.6
KPT(SOTA)	*82.8*	*85.0*	*86.3*	*87.5*	*92.0*	*92.6*	*93.8*	*94.1*	*92.1*	*93.8*	*94.1*	*94.4*	*94.9*	*95.4*	*96.3*	*96.9*	*53.9*	*57.4*	*59.5*	*60.7*
MMLM(W)	**83.0**	**85.6**	**87.1**	**88.6**	**92.6**	**93.6**	**93.9**	**94.5**	**92.4**	**94.5**	**95.2**	**95.4**	**95.1**	**95.8**	**96.4**	**97.1**	**54.7**	**58.4**	**59.8**	**61.3**
MMLM(P)	82.9	85.5	86.9	**88.6**	92.5	93.5	**93.9**	94.3	92.3	**94.5**	95.1	**95.4**	94.9	95.4	**96.4**	96.6	54.5	**58.4**	59.7	**61.3**

Table 1: The overall classification performance on DBPedia/Yahoo/AG's News (topic) and IMDB/Amazon (sentiment).

Method	K=1	K=5	K=10	K=20
MMLM(W)	**83.0**	**85.6**	**87.1**	**88.6**
(-dropout)	82.8	85.4	87.0	88.2
(-dropout-disturb)	82.7	85.3	86.9	88.0
-Mul	82.4	84.9	86.4	87.2
-Mul-Aug	78.4	83.1	85.1	86.4

Table 2: MMLM ablation studies. -Mul and -Aug represent removing multiple classifiers and augmentation respectively.

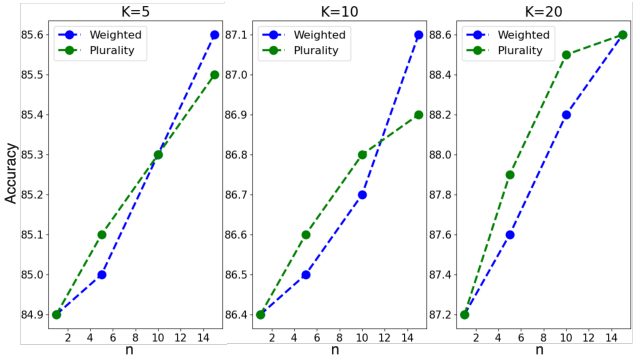

Figure 3: Comparison between different n values on AG's News for K=5, K=10, and K=20.

Overall Performance

As shown in Table 1, traditional CLS fine-tuning works poorly in few-shot scenarios since the number of labeled instances is extremely limited. Especially in the case of K=1, where only 2 instances for IMDB and Amazon, and 4 instances for AG's News are available, CLS fine-tuning almost performs like a random classification. Though, KPT performs better than the other methods because it exploits external knowledge (Hu et al. 2022).

In contrast, MMLM stimulates the compressed knowledge in the pre-trained model only by augmentation and multi-mask scheme. The results demonstrate that MMLM outperforms all other existing label mapping methods on all three sentiment classification datasets. Though KPT can greatly improve on the relatively difficult 4-way classification dataset AG's News, our proposed MMLM still achieves comparable performance when K$\geq$1.

Furthermore, both two voting methods are also shown to be effective while MMLM with weighted vote acquires slightly higher accuracy than plurality vote.

Ablation Study

To separately observe the impact of the prompt-based augmentation and the multi-mask scheme individually, we conduct ablation studies in this section. We compare three main settings of MMLM models, original MMLM, MMLM without multi-mask scheme, and MMLM without multi-mask scheme and augmentation respectively.

For $\mathrm{MMLM-Mul}$, we only use one classifier by only using one lexical cue. For $\mathrm{MMLM-Mul-Aug}$, it becomes equivalent to vanilla label mapping after the two modules are removed.

As shown in Table 2, when K is relevantly small, the prompt-based augmentation can greatly improve the model performance. When K is larger (e.g., K=20), the multi-mask scheme is of more help for performance improvement as it integrates information of multiple instances.

Besides, we also illustrate the effectiveness of disturbed embeddings and dropout layers in the multi-mask scheme. For $\mathrm{MMLM-dropout}$, we set the dropout rates for attention and forward layers to 0. For $\mathrm{MMLM-dropout-disturb}$, we further remove the perturbation term.

The results demonstrate that dropout layers and disturbed embeddings are both effective for performance improvement, on which the former has a slightly higher impact.

Impact of Number of Classifiers

As proved in the earlier section, the misleading lexical cue has less impact when increasing the number of prompt-based classifiers. In other words, ideally, the prediction accuracy will improve if more classifiers are used. Although we propose three methods to enlarge the independence between different prompt-based classifiers, they are still not fully independent since they share the same cloze-type pre-trained language model. Therefore, we further conduct a series of experiments to verify the efficiency of our proof.

Figure 3 illustrates the few-shot classification accuracy raises with the increasing n. This shows MMLM can be ex-

Method	K=1	K=5	K=10	K=20
NO AUG	78.4	83.1	85.1	86.4
SINGLE	81.3	84.9	86.7	87.9
MMLM(W)	**83.0**	**85.6**	**87.1**	**88.6**
MMLM(P)	**82.9**	**85.5**	**86.9**	**88.6**

Table 3: Comparison between vanilla model (NO AUG) and the model with all keywords being stuffed into one classifier (SINGLE) where the number of keywords is 15.

Method	AG's News (NullPrompt)			
	K=1	K=5	K=10	K=20
Manual	65.0	76.5	79.7	85.5
Search	37.1	63.5	77.6	85.6
Soft	66.6	74.9	81.3	85.4
KPT (SOTA)	*76.4*	*82.6*	*85.6*	*86.7*
MMLM (W)	**77.5**	**83.2**	**86.2**	**87.6**
MMLM (P)	77.0	**83.2**	85.8	87.2

Table 4: Comparison with Null Prompt.

pected for more performance improvement if n can be increased with more GPU resources or longer sentence length.

Comparison With Multi-Keywords in One Sentence

To compare with a different form of cloze-based augmentation with prompt, instead of generating n prompt-based classifiers, we stuff all extracted lexical cues into one sentence to form a single prompt-based classifier. For example, an augmented instance can be constructed as "Boring starting but overall ok and worth watching. Keyword: Boring, ok, worth. It was $\{mask\}_T$.".

As shown in Table 3, it achieves lower performance than MMLM on AG's News, which indicates the effectiveness of the multi-mask scheme. But this single-sentence augmentation model still outperforms the vanilla model, further showing the reasonability of our proposed augmentation method.

Alleviating Effect of Templates

To alleviate this concern that templates may bring fluctuation to the classification performance (Gao, Fisch, and Chen 2021; Liu et al. 2021), we follow the idea of NullPrompt (Logan IV et al. 2022) and use '{TEXT} {mask}$_T$' to compare all label mappings. As Table 4 shows, MMLM still outperforms all baselines on AG's News under this fair comparison setting, further demonstrating its effectiveness.

Case Study of Lexical Cue Extracting

We previously demonstrated that employing the proposed multi-mask scheme can reduce the misclassification rate within one iteration. Taking a step further, we expect to reduce the number of extracted misleading lexical cues. As in Figure 4, the top-3 influential cues begin with containing two misleading keywords 'first' and 'Russian'. The model is

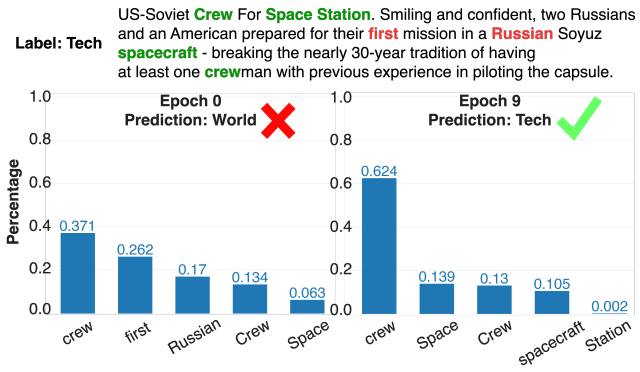

Figure 4: Improvement of the extracted keywords and the prediction result.

Epoch Class		Word with Probability (%)		
Ep 0 Acc 0.86	**W**	nuclear (5.8)	tsunami (4.9)	dealt (4.8)
	S	brawl (4.9)	hamstring (4.7)	Texas (4.5)
	B	Oil (8.9)	GDP (5.0)	Lisbon (4.9)
	T	Google (5.0)	hacking (4.9)	Mac (4.8)
Ep 9 Acc 0.90	**W**	Haiti (5.0)	dealt (4.9)	IRA (4.9)
	S	Olympic (9.9)	Ravens (7.9)	trade (5.0)
	B	inflation (9.4)	Marsh (5.0)	Lisbon (4.9)
	T	IBM (5.0)	Technology (5.0)	hacking (4.9)

Table 5: Performance improvement with the change of word-occurrence probability at $\{mask\}_t$ in the keyword extractor. (W: World; S: Sports; B: Business; T: Technology.)

more likely to make a wrong prediction even if the information of misleading lexical cues becomes dim by the multi-mask scheme. This problem can be minimized by iterating the optimizations. For instance, after 9 iterations, the top-5 keywords are partly replaced. This is because the parameters in the keyword extractor are updated with the parameters in the classifier since they share the same parameter. The word-occurrence probability at $\{mask\}_t$ also changes towards a class-related bias as shown in Tabel 5.

Conclusion

In this paper, we demonstrate how multi-mask approach improves label mapping performance in the prompt-based setting. While existing works focus on data augmentation for generation-type language models, we propose an augmentation method for cloze-type language models to satisfy the conditions of few-shot learning. Further, because lexical cues are proven to play a significant role in large language models like BERT/RoBERTa for classification, containing misleading lexical cues in input text easily leads to wrong predictions. By exploiting multiple instances with multiple classifiers, MMLM is able to reduce the impact of misleading lexical cues. Theoretical analysis shows that exploiting multiple lexical cues is better than one and empirical studies confirm that our proposed model achieves SOTA results in different experimental settings.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0110700, in part by the Fundamental Research Funds for the Central Universities, in part by the State Key Laboratory of Software Development Environment.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cui, G.; Hu, S.; Ding, N.; Huang, L.; and Liu, Z. 2022. Prototypical Verbalizer for Prompt-based Few-shot Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7014–7024. Dublin, Ireland: Association for Computational Linguistics.

Davison, J.; Feldman, J.; and Rush, A. 2019. Commonsense Knowledge Mining from Pretrained Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1173–1178. Hong Kong, China: Association for Computational Linguistics.

Ding, N.; Hu, S.; Zhao, W.; Chen, Y.; Liu, Z.; Zheng, H.; and Sun, M. 2022. OpenPrompt: An Open-source Framework for Prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 105–113. Dublin, Ireland: Association for Computational Linguistics.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830. Online: Association for Computational Linguistics.

Hambardzumyan, K.; Khachatrian, H.; and May, J. 2021. Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, 409–426. Springer.

Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Wang, J.; Li, J.; Wu, W.; and Sun, M. 2022. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2225–2240. Dublin, Ireland: Association for Computational Linguistics.

Kavumba, P.; Inoue, N.; Heinzerling, B.; Singh, K.; Reisert, P.; and Inui, K. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 33–42. Hong Kong, China: Association for Computational Linguistics.

Kavumba, P.; Takahashi, R.; and Oda, Y. 2022. Are Prompt-based Models Clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2333–2352. Dublin, Ireland: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*.

Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; and Chen, W. 2022. On the Advance of Making Language Models Better Reasoners. *arXiv preprint arXiv:2206.02336*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Logan IV, R.; Balazevic, I.; Wallace, E.; Petroni, F.; Singh, S.; and Riedel, S. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2824–2835. Dublin, Ireland: Association for Computational Linguistics.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Florence, Italy: Association for Computational Linguistics.

Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664. Florence, Italy: Association for Computational Linguistics.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. Online: Association for Computational Linguistics.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Trichelair, P.; Emami, A.; Trischler, A.; Suleman, K.; and Cheung, J. C. K. 2019. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3382–3387. Hong Kong, China: Association for Computational Linguistics.

Trinh, T. H.; and Le, Q. V. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Wang, H.; Xu, C.; and McAuley, J. 2022. Automatic Multi-Label Prompting: Simple and Interpretable Few-Shot Classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5483–5492. Seattle, United States: Association for Computational Linguistics.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; and Chi, E. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv preprint arXiv:2205.10625*.