IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 September 2022, accepted 4 October 2022, date of publication 10 October 2022, date of current version 13 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3213027

RESEARCH ARTICLE

Plain Template Insertion: Korean-Prompt-Based Engineering for Few-Shot Learners

JAEHYUNG SEO[1], HYEONSEOK MOON[1], CHANHEE LEE[1], SUGYEONG EO[1], CHANJUN PARK[1,2], JIHOON KIM[3], CHANGWOO CHUN[3], AND HEUISEOK LIM[1,4]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Upstage, Gyeonggi-do 16942, Republic of Korea
[3]Hyundai Motor Group, Seoul 06797, Republic of Korea
[4]Human Inspired Artificial Intelligence Research (HIAI), Korea University, Seoul 02841, Republic of Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the ICT Creative Consilience Program Supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2022-2020-0-01819; in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2021R1A6A1A03045425; and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (MSIT), a Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques under Grant 2020-0-00368.

ABSTRACT Prompt-based learning is a method used for language models to interpret natural language by remembering the prior knowledge acquired and the training objective. Recent prompt-based few-shot learners have achieved superior performance by alleviating the catastrophic forgetting that occurs in pretrained language models. Few-shot learning contributes towards solving the data scarcity problem, an enormous challenge in AI systems and a significant consideration in natural language processing research. In spite of the significance of few-shot learning, research on Korean language-based few-shot learning is insufficient, and whether the prompt-based approach is appropriate for the Korean language has not been thoroughly verified. As a step toward realizing a Korean-prompt-based few-shot learner, we attempt to apply prompt engineering to the Korean language understanding benchmark dataset and introduce plain template insertion to overcome data scarcity in a more practical few-shot setting. The contributions of this study are as follows: (1) presumably, this is the first study to apply prompt-based few-shot learning to Korean benchmark datasets. With 32 few-shot settings, it improves performance by +14.88, +29.04, and +1.81 in the natural language inference, semantic textual similarity, and topic classification tasks. (2) We present prompt engineering, which merely inserts a plain template and increases data efficiency without training example selection, augmentation, reformulation, and retrieval. (3) Our approach is robust to the Korean prompt's contextual information and sentence structure and is applicable to both hard- and soft-prompt.

INDEX TERMS Prompt-based learning, natural language processing, language modeling, Korean language understanding, few-shot.

I. INTRODUCTION

The recent concept of prompt-based learning has been proposed to utilize the vast amount of latent prior knowledge of contained in pretrained language models. Prompt-based learning predicts the correct answer, that is, it solves the task at hand based on linguistic knowledge and contextualized representations memorized during pretraining by reforming the input sequence with a textual prompt. This method enables finetuning of the model with well-aligned pretraining objective, demonstrating exceptionally high performance in using low-resource strategies.

In recent years, research on few-shot learning has been actively conducted to achieve high learning performance with small amounts of training data and light models by using prompt-based learning. According to [9],

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

prompt-based learning is divided into hard-prompts using human-interpretable natural language in discrete spaces and soft-prompts using the continuous embedding space of the model directly. Pattern-exploiting training (PET) [20] and better few-shot finetuning of language models (LM-BFF) [5] based on hard-prompt-based learning demonstrate the interpretability of the model's inference results, achieving superior performance on the GLUE [25] and SuperGLUE [26] benchmark datasets. In soft-prompt tuning, P-tuning [11] achieves outstanding performance on few-shot SuperGLUE dataset through additional prompt encoders' parameters.

Despite the success of prompt-based few-shot learning, research on prompt-based few-shot research learning for Korean has not matured rapidly. Prompt-based learning in other languages cannot guarantee high performance in Korean, which has different word orders and an agglutinative linguistic feature. Further, there is a problem concerning the few-shot settings in previous studies that must be addressed: the earlier prompt-based few-shot learners deviate from the original purpose of overcoming data scarcity. They reformulate the given question for mapping with the prompt, selectively set high-quality training samples, or augment relevant examples to ensure high performance. These approaches require considerable human effort and time during the preprocessing. Moreover, it is challenging to include selective data curation in domains where only a tiny amount of training data exists, and clear distinction for achieving higher quality standards is difficult.

To address these issues, we present a Korean-prompt-based few-shot learner, in which we limit excessive resource consumption using prompt engineering for a more practical few-shot setting. To lighten prompt engineering, we propose plain template insertion (**PTI**) that prohibits augmenting the number of few-shot training examples and avoids human annotators' variations of sentences that do not suit the few-shot purpose. PTI is a method of placing a predefined template containing prompts and a [MASK] token into a specific position without refining the input sample. However, being pathfinders in the field of Korean-prompt-based few-shot learning, we do not intend to use null prompts [13], as these do not aid in grasping the linguistic meaning. Instead, we closely analyze the prompt-based few-shot learner's inherent challenges and reflect the linguistic features of Korean in the proposed PTI.

In this study, we apply the proposed method to a benchmark dataset, referring to as Korean language understanding evaluation (KLUE) [16], and off-the-shelf Korean pretrained language models. We perform comparative analyses by selecting three tasks: natural language inference (NLI) [2], semantic textual similarity (STS) [1], and topic classification (TC) [7], [18], [29], which have critical attributes that require an understanding of Korean with respect to the KLUE benchmark dataset. (i) KLUE-NLI requires an understanding of entailment and contradiction in the context. It is a fundamental evaluation in the subfield research of comprehending semantic representations of natural languages. (ii) KLUE-STS is a task that measures the semantic similarity between sentences. We demonstrate that Korean-prompt-based learning is applicable even when regression is applied to continuous real values rather than discrete classification [5]. (iii) KLUE-TC is a new subtask that does not exist in the GLUE and SuperGLUE datasets. It predicts one of seven predefined news categories (for example, economics and sports) based on a specific news headline. It indicates that prompt-based learning improves the inference capability of the model even for a sequence with a high degree of abstraction.

II. RELATED WORKS

With the advent of language models that can assimilate vast amounts of knowledge, enormous amounts of research has been conducted on the performance of downstream tasks by reforming the language model into cloze-style to maximize its capability. However, considering the low-data regime, recent research has been focused on effectively few-shot learning based on a limited number of examples.

In the few-shot setting, Schick and Schütze [20] proposed PET, which incorporates knowledge distillation and self-training for downstream tasks. They finetuned individual language models for data reformulated with pre-defined patterns and assembled them to annotate soft labels for large unlabeled datasets, which are then applied for classifier learning. An additional method known as iPET was introduced to address the discrepancy and performance gap among trained language models for each pattern. Thereafter, considering that handling only a single token in PET makes answer representation challenging, Schick and Schütze [21] adapted the model to predict multiple tokens.

Several studies have pointed out that leveraging hard-prompts is sub-optimal and demands non trivial labor. To reduce the prompt engineering effort and investigate an optimal prompt, soft-prompting was introduced. In particular, Liu et al. [11] presented P-tuning, which places anchor tokens and automatically searches for prompts by optimization through gradient descent in continuous space using a long short-time memory (LSTM) structure. Logan IV et al. [13] far diminish the design endeavor, as they perform null prompting consisting of the input sentence and masking token only.

Certain studies have also adopted intriguing approaches in the few-shot setting. In terms of boosting performance using only a small amount of data, Wang et al. [27] approached few-shot learning from the perspective of meta-learning. Cross-task transferable knowledge is obtained by performing multi-task finetuning with similar natural language processing tasks before adapting the learner to a specific task. In terms of data efficiency, there have been attempts to determine the prompting data points. Zhao and Schütze [19] investigated an appropriate data quantity per task and verified its effect by comparing it with head-based (or promptless) finetuning. In terms of language extension, XLM-RoBERTabase [30] performed multilingual few-shot learning using prompting. They utilize a multilingual pretrained XLM-RoBERTa-base [3] model along with prompting methods to conduct natural language understanding tasks in 15 languages.

Furthermore, most recent research for prompt-based learning has been applied in various fields of natural language processing including relation extraction [22], commonsense reasoning [8], and complementing weakness of prompt [4], [15], [23]. Son et al. [22] introduced a multitask learning approach for predicting a relation in a dialogue by guiding the model on the relational cues with an MLM-based relational mention prediction and the prior distribution of entity types. Liu et al. [8] proposed generated knowledge prompting to obtain the external knowledge required to solve commonsense reasoning tasks. Cui et al. [4] proposed a soft prototype verbalizer to find a suitable verbalizer within a large vocab. Lu et al. [15] pointed out that model performance can deviate significantly depending on the order of the training samples and the prompt position in prompt-based few-shot learning. Sorensen et al. [23] presented a new approach for choosing generated templates based on mutual information without human-annotated labels or updating models.

However, most studies have mainly experimented on English, and languages with different morphological characteristics have hardly been addressed. Therefore, in this study, we investigate how the performance of few-shot learning varies by applying prompt-based learning to Korean. Moreover, we argue that manipulating data, such as selective exploitation of high-quality data for training, obscures the main purpose of few-shot learning. That is, it is necessary to reconsider whether a scheme adequately alleviates the data scarcity problem. Confronted with this issue, we use data in a more strict and practical manner.

III. PRELIMINARY

In this section, we provide some background knowledge on hard- and soft-prompt tuning. The composition of the template is divided into hard and soft depending on the prompt used to represent the embedding space. We redesign the structure of the model by considering the tuning methods of prompts and determining the form of the template.

A. HARD-PROMPT

A hard-prompt consists of the template T_d with human-interpretable natural language in discrete space. $T_d = d_{0:m}, d_m, d_{m+1:n}$ reflects the purpose of contextual information and the training objective in Korean where each d_i represents natural language tokens including the [MASK] token for d_m. Along with the input sequence X, T_d is fed to the pretrained language model for the intended downstream task. Thereafter, the model is trained to restore the m^{th} position of T_d to the mapped label word $W(y)$ based on linguistic knowledge acquired during the pretraining. The label word $W(y)$ is selected as the natural language token that retains syntactic coherence with T_d and fits the purpose of the original label. In particular, hard-prompt-based training objective of the model θ can be described as follows:

$$\max_{\theta} P_{\theta}(y_{[\text{MASK}]} = W(y)|X, T_d) \tag{1}$$

where $y_{|\text{MASK}|}$ indicates the model output for the m^{th} position of T_d, which is the position of the [MASK] token in the input sequence.

B. SOFT-PROMPT

A soft-prompt consists of a human-uninterpretable template $T_c = c_{0:n}, c_m$ that is composed of trainable embedding vectors in a continuous latent space and a mask token $c_m = $ [MASK]. For the implementation, the prompt template tokens $c_{0:n}$ and distinct prompt embedding $e'(\cdot)$ are initialized. This enables the adaptation of the template to fit the optimal template, which may not be grasped by discrete natural language tokens. In particular, in adopting the soft-prompt, the model θ is trained with the following training objective:

$$\max_{\theta} P_{\theta}(y_{\text{m}} = W(y)|e(X), e'(c_{0:n}), e(c_m)) \tag{2}$$

where $e(\cdot)$ indicates the original token embedding of the pretrained language model, and y_{m} indicates the model output for the m^{th} position of T_c, which is the position of the [MASK] token in the input sequence. The prompt token embedding is trained in a different latent space from the original pretrained language model. Optionally, prompt embedding can further be trained with a Bi-LSTM encoder $E_{bi}(\cdot)$ to impose a direct relational connection between $c_{0:n}$ [10]. Equation 3 represents the encoding process of $E_{bi}(c_{0:n})$. In training the model θ using these, $e'(c_{0:n})$ in equation 2 is replaced with $E_{bi}(c_{0:n})$.

$$E_{bi}(c_{0:n}) = \text{MLP}[\text{LSTM}(c_{0:n}) : \text{LSTM}(c_{n:0})] \tag{3}$$

IV. PLAIN TEMPLATE INSERTION

We propose **plain template insertion (PTI)**, which places a manual template in the fitted position considering minimal Korean contextual information for the given question and the connection between sentences. We integrate PTI with hard- and soft-prompt tuning. PTI can be engineered differently depending on the content, position, and mapping labels, thereby significantly increasing the data efficiency of the few-shot examples.

A. TEMPLATE CONTENT

The template consists of a [MASK] token and prompts combined to determine the content. For example, if the purpose of a given task is to identify the relationship between two input sentences, we can set the template content as follows: "The two sentences are [MASK] related." Special symbol tokens (for example, |, ?) can be used to naturally interpret the context between the input sentences and the template or distinguish the relationship as a separator. Further, the template content represents the randomly initialized embedding space in the soft-prompt. By setting the length of the template, we determine the number of trainable continuous prompts to include.

B. TEMPLATE POSITION

The template is inserted at a specific position in the given sequence. We set the fixed position before and after the input sentence. For example, if two sentences <s_1> and <s_2> are specified as the input sequence, the position of template [t] corresponds to one of "[t] <s_1> <s_2>," "<s_1> [t] <s_2>," or "<s_1> <s_2> [t]." In the hard-prompt, we select the template position such that it does not violate Korean grammar. In contrast, in the soft-prompt, we set the template position regardless of the context. For instance, prompt <p> of template [t] with a specified length can be partially separated to form: "<p> <s_1> [t] <s_2> <p>."

C. MAPPING LABEL

The mapping label is part of the template and represents the answer word to be predicted in the task. Unlike conventional finetuning using the [CLS] token as a model prediction, prompt-based learning infers the answer in the same way as masked language modeling pretrained with self-supervised learning. This approach inhibits the occurrence of catastrophic forgetting caused by the gap between pretraining and finetuning in language models. The hard-prompt maps one of the tokens in the vocabulary of the model, with considering the context to the [MASK] token (that is, we follow the verbalizer in Schick and Schütze [20], [21]). The soft-prompt uses the PTI that meets the training purpose so that the continuous prompt updates the embedding value based on the given sequence and mapping label.

V. PROMPT ENGINEERING

To assess our Korean-prompt-based few-shot learner, we apply PTI to the NLI, STS, and TC tasks, for which the KLUE benchmark dataset is used. We assume that prompt-based learning drives the linguistic knowledge of pretrained language models to understand the contextual representations necessary to solve the downstream task. As explained in Section §IV, a PTI consists of the (i) template content, (ii) template position, and (iii) mapping labels as three variables. PTI is used differently for the hard- and soft-prompt.

A. HARD PLAIN TEMPLATE INSERTION

This section introduces a method of engineering PTI with human-interpretable hard-prompt tuning, which differs for the three downstream tasks.

1) KLUE-NLI

KLUE-NLI is a classification task to train a model to infer the relationship between the premise and hypothesis sentences. The trained model identifies one of three types of relations between two sentences: Entailment, contradiction, or neutrality. The premise and hypothesis sentences include various topics and writing styles in contemporary Korean.

(i) Content of template [t] deliberates the purpose of the task of questioning the relationship between the premise <s_{pre}> and the hypothesis <s_{hyp}>. For example, we make a template "두 문장의 관계는 [MASK]이다. (The two sentences are [MASK] relationships.)" to deduce understanding of textual entailment. (ii) The template position has three types of forms, considering that there are two input sentences: "[t] <s_{pre}> <s_{hyp}>," "<s_{pre}> [t] <s_{hyp}>," or "<s_{pre}> <s_{hyp}> [t]." (iii) KLUE-NLI's annotated three types of relations remain open to interpretation, making it challenging to select flawlessly mapped Korean tokens. Thus, mapping labels use tokens with similar meanings among those from the model's vocabulary. For example, "same" is a token that has a meaning similar to "entailment," and "different" is mapped to "contradiction."

2) KLUE-STS

KLUE-STS is a regression task that the model uses to predict the degree of semantic equivalence between two sentences. This dataset labels semantic similarity as a real value from 0 (indicating no meaning equivalence) to 5 (indicating complete meaning equivalence).

(i) The template content describes the semantic simplicity of the two sentences as [MASK] token and discrete (ii) To recognize the relationships of two sentences <s_1> and <s_2>, we define the position of the template as follows: "[t] <s_1> <s_2>," "<s_1> [t] <s_2>," or "<s_1> <s_2> [t]." For example, our template content is "[MASK] 내용으로, (As the [MASK] content,)." (ii) To recognize the relationships between the two sentences <s_1> and <s_2>, we define the template position as follows: "[t] <s_1> <s_2>," "<s_1> [t] <s_2>," or "<s_1> <s_2> [t]." (iii) Mapping real values to discrete label words is quite difficult. Thus, we use the binary classification tag in KLUE-STS, mapped with similar meaning tokens in the vocabulary. However, with mapping labels such as 같은 (same) and 다른 (different), it is still difficult to predict the real value of the semantic similarity score. Thus, inspired by [3] and [25], we use linear interpolation to alter the range of the real values [0,5] into [0,1] to map the two opposing poles.

3) KLUE-TC

KLUE-TC is a classification task in which the model predicts one of the seven predefined news categories based on a given news headline. It consists of human-annotated news headlines from online articles distributed by the Yonhap News and published from January 2016 to December 2020.

(i) The content of template is constructed by reflecting the attribute of the training objective that infer the one category (or topic) of a given sentence. For example, we use a predefined plain template, "해당 뉴스는 [MASK] 분야에 해당한다. (The news corresponds to the [MASK] field.)" considering the neighboring contextual information. (ii) The position of template is "<s_1> [t]" or "[t] <s_1>" whether a news headline as a single input sequence <s_1> precedes the plain template [t] or <s_1> follows [t]. (iii) The mapping labels consist of the same seven textual categories based on tokens in the vocabulary of the model. As in other tasks, we produce the mapping labels differently

TABLE 1. KLUE-NLI/STS/TC dataset statistics in KLUE benchmark. # class refers to the number of labels.

	Train set	Validation set	Test set
KLUE-NLI (# class = 3)	24,998	3,000	3,000
⊢ **Few-shot 16**	48 (16×3)	48 (16×3)	3,000
⊢ **Few-shot 32**	96 (32×3)	96 (32×3)	3,000
⊢ **Few-shot 64**	192 (64×3)	192 (64×3)	3,000
KLUE-STS (# class = 2)	11,668	519	1,037
⊢ **Few-shot 16**	32 (16×2)	32 (16×2)	519
⊢ **Few-shot 32**	64 (32×2)	64 (32×2)	519
⊢ **Few-shot 64**	128 (64×2)	128 (64×2)	519
KLUE-TC (# class = 7)	45,678	9,107	9,107
⊢ **Few-shot 16**	112 (16×7)	112 (16×7)	9,107
⊢ **Few-shot 32**	224 (32×7)	224 (32×7)	9,107
⊢ **Few-shot 64**	448 (64×7)	448 (64×7)	9,107

according to the subword tokenized separators. For instance, we consider "[과학 (science), 경제 (economy), 사회 (social), 문화 (culture), 세계 (world), 스포츠 (sports), 정치 (politics)]" and "[##과학 (science), ##경제 (economy), ##사회 (social), ##문화 (culture), ##세계 (world), ##스포츠 (sports), 정치 (politics)]" as different mapping labels.

B. SOFT PLAIN TEMPLATE INSERTION

PTI in continuous embedding space consists of uninterpretable prompts but has the advantages of being task-agnostic and independent of contextual information. Therefore, we use the same PTI for three different downstream tasks.[1]

(i) We use randomly initialized continuous prompts to overcome the discreteness, which quickly converges to the local minimum caused by the discrete words [11]. A continuous prompt is not mapped to a natural language of discreteness but determines how much the template contains. The content of such a template with two sentences <s_1> and <s_2> and length of l prompts is "<s_1><p_1>…<p_l> [MASK] <s_2>." (ii) The position of the template is relatively unrestricted. The randomly initialized template is independent of the context, and some prompts can be separated. If a template with l prompt tokens is separated, "<p_0> …<p_{i-1}> <s_1> <p_i> …<p_{l-1}> [MASK] <s_2>" is one of the examples of the position of the soft-prompt template. (iii) The mapping labels consist of the same discrete label words as for the hard-prompt, but there is an additional option to remove the mapping to discrete label words.

VI. EXPERIMENTAL SETTINGS

In this section, we specify the settings for the datasets, models, and evaluation metrics for few-shot learning. We conduct our experiments and analyses based on the experimental settings specified in this section. More details about training environments and hyperparameters are described in an Section VII.

A. DATASET

As described in Table 1, we employed KLEU-NLI/STS/TC dataset as a few-shot learning data considering the number of classes for each task. The recent KLUE dataset does not publicly open a leaderboard[2] and only training sets D_{train} and development sets D_{dev} are disclosed. Therefore, we used the entire D_{dev} as a test sets D_{test}. Inspired by [5] and to achieve the goal of learning from scarce data, we randomly selected the same number of samples as the few-shot size K from the entire D_{train} and used them as the few-shot training sets D^K_{train} and few-shot development sets D^K_{dev} (that is, $|D^K_{\text{train}}| = |D^K_{\text{dev}}|$). Each label in the D^K_{train} and D^K_{dev} has K samples and is composed differently based on the five seeds: {42, 52, 62, 72, 3407}.

B. MODELS

We choose KLUE-RoBERTa-large [16], which is pretrained with the RoBERTa [12] architecture and Korean corpora and demonstrates the best performance in the KLUE benchmark as a few-shot the solely masked language model pretraining of RoBERTa is acceptable for prompt-based learning to predict the [MASK] token. To recognize the improvement gap based on the model size, we supplemented the experiments for KLUE-RoBERTa-base. In soft-prompt, we optionally attach a Bi-LSTM head to the RoBERTa and conduct comparative analysis in Table 5. Bi-LSTM is a lightweight natural network that prevents the problem of rapidly converging continuous prompts to discreteness caused by a pretrained language model's embedding layers and sets randomly initialized prompts to have contextual dependencies [11].

C. METRICS

We measure performance based on the evaluation metrics presented by the KLUE [16]. KLUE-TC estimates performance using the F1-score, the harmonic mean of precision and recall, to prevent false-positive and false-negative problems arising from the imbalance of the seven labels. KLUE-NLI is uniformly divided into three labels and measures performance with accuracy. KLUE-STS measures performance using the Pearson correlation representing a linear correlation between the predicted and label values.

VII. EXPERIMENTAL DETAILS

We implemented the Huggingface [28] and Pytorch-lightning[3] framework for language modeling on an 18-core Intel Xeon Gold 6230 CPU and an NVIDIA Quadro RTX A6000 GPU. We trained Bi-LSTM (embedding size 300, 2 layers, model parameters 3.8M), KLUE-RoBERTa-base (embedding size 768, 12 layers, 12 heads, model parameters

[1] Whether there are one or two input sentences leads to different number of choices, but the rule of template insertion is effectively the same.

[2] The performances of baseline models are presented on the leaderboard, but updates to additional models are not activated.

[3] https://github.com/PyTorchLightning/pytorch-lightning

TABLE 2. We set the few-shot size to 16/32/64 and compare the performance between finetuning and our prompt-based few-shot learner. We report mean performance over five different seeds. Majority selection refers to predicting the answer with the largest number of labels. PTI(H) and PTI(S) indicate plain template insertion with hard- and soft-prompt engineering, respectively. ↑ presents performance improvements for the same-sized finetuned model. ↓ indicates a decline in performance for the same size finetuned model. The best models for the same few-shot size are formatted in bold and the second-best ones are underlined.

	Few-shot size 16			Few-shot size 32			Few-shot size 64		
	KLUE-NLI (Accuracy)	KLUE-STS (Pearsonr)	KLUE-TC (F1-score)	KLUE-NLI (Accuracy)	KLUE-STS (Pearsonr)	KLUE-TC (F1-score)	KLUE-NLI (Accuracy)	KLUE-STS (Pearsonr)	KLUE-TC F1-score
Majority selection	33.33	-	40.16	33.33	-	40.16	33.33	-	40.16
Bi-LSTM	33.49	16.67	14.32	34.20	21.25	20.72	34.88	26.66	27.21
KLUE-RoBERTa-base	34.34	28.72	77.11	39.31	55.40	79.60	43.71	63.70	79.71
KLUE-RoBERTa-large	37.84	24.06	74.14	43.23	41.43	78.24	45.50	73.07	78.50
KLUE-RoBERTa-base + PTI(H)	44.27 ↑	36.45 ↑	**77.35** ↑	49.72 ↑	57.79 ↑	**80.81** ↑	57.46 ↑	73.57 ↑	**81.45** ↑
KLUE-RoBERTa-base + PTI(S)	45.31 ↑	36.89 ↑	72.23 ↓	53.68 ↑	57.20 ↑	76.35 ↓	55.31 ↑	72.92 ↑	78.46 ↓
KLUE-RoBERTa-large + PTI(H)	44.90 ↑	**53.60** ↑	76.81 ↑	50.55 ↑	**70.47** ↑	80.05 ↑	**60.07** ↑	**84.11** ↑	81.07 ↑
KLUE-RoBERTa-large + PTI(S)	**49.69** ↑	47.67 ↑	76.57 ↑	**58.11** ↑	64.78 ↑	76.48 ↓	58.76 ↑	75.98 ↑	75.92 ↓

110M) and KLUE-RoBERTa-large (embedding size 1024, 24 layers, 16 heads, model parameters 377M) model as a few-shot learner. In addition, we assigned the template content, template position, and mapping labels in the process of loading data modules. We conducted the validation during the training step and saved the best checkpoint with the highest performance monitoring with evaluation metrics for each task.

A. HARD-PROMPT PTI

We set the hyperparameters in the training process as follows: batch size 8, maximum length of sequence 128, 20 epochs, warmup ratio 0.1, AdamW optimizer [14] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$), initial learning rate of 5×10^{-5}, drop out rate 0.4, and five different seeds (42, 52, 62, 72, 3407).

B. SOFT-PROMPT PTI

We set the hyperparameters in the training process as follows: batch size 8, maximum length of sequence 128, 20 epochs, warmup ratio 0.1, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$), initial learning rate of 3×10^{-5}, drop out rate for the prompt encoder as 0.3, and five different seeds (42, 52, 62, 72, 3407).

C. BI-LSTM

We set the hyperparameters in the training process same as soft-prompt PTI. Also, we utilized FastText [6] with 300-dimension embeddings based on common crawl[4] corpus to train Bi-LSTM.

VIII. EXPERIMENTAL RESULTS

A. MAIN RESULTS

We conduct a quantitative analysis of our PTI for the three types of Korean natural language understanding tasks. As presented in Table 2, it is difficult for the model to achieve stable performance because few-shot training examples do not include the selection of quality, and the number of development examples is extremely limited. In particular, Bi-LSTM model without pre-trained knowledge has a problem that can perform less than the majority selection for multi-class classification problems. However, few-shot learning, which assumes strict data scarcity, still reveals biased features with a small amount of data. Even though it is difficult for the model to guarantee the best result caused by the typical problem of few-shot learning, PTI outperforms the traditional promptless finetuning method in 31 of the 36 comparative experiments except soft-prompt PTI in KLUE-TC. These results indicate that PTI has the advantage of performance improvement while increasing data efficiency by adding three variables to the input sequence.

PTI incorporated into hard-prompt engineering exhibits performance improvement of +14.57, +29.54, and +2.67 in KLUE-NLI, STS, and TC at most, respectively. PTI incorporated into soft-prompt engineering enhances the performance by +14.88, +23.61, and +2.43 in KLUE-NLI, STS, and TC at most, respectively, but a maximum performance reduction of −4.88 in KLUE-TC. There is no previous research on whether the topic classification is valid for soft-prompting in the English benchmark datasets. However, based on the achievements made by hard-prompt engineering, we interpret that the continuous prompt rather increases the difficulty of the tasks while as well as the uncertainty. Moreover, few-shot learners do not necessarily guarantee higher performance of larger models because of the uncertainty of data scarcity. Although complete improvement has not been made, the PTI approach to the same few-shot examples in the KLUE-STS task relieves the uncertainty with consistent results in the model size.

B. FEW-SHOT AND FULL-SHOT SETTINGS

We conduct a case study with KLUE-STS as an instance to evaluate the change in performance based on the number of training examples. As depicted in Figure 1, PTI illustrates high data efficiency in the strict few-shot setting with a small number of training and development examples. In the KLUE-STS task, we show that 1% of the training examples (128/11,668) can achieve performance close to that with the full-shot setting. However, as the number of training and development examples increases, the performance gap

[4] https://commoncrawl.org/

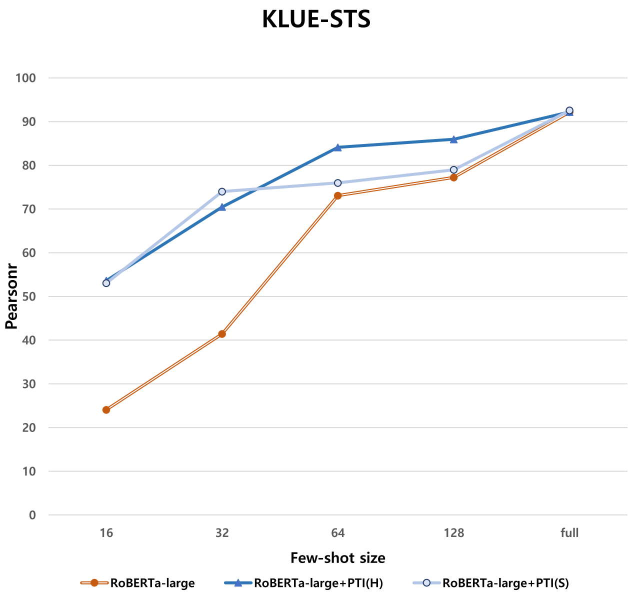


FIGURE 1. Performance on KLUE-STS with few-shot (16/32/64/128) and full-shot dataset. We average over five seeds to compute the score. We apply hard/soft PTI and promptless finetuning to the RoBERTa-large model.

between PTI and promptless finetuning gradually converges. Experiments with the full-shot setting indicate that the gap between the two approaches is marginal. This reason for this is that the development examples of the full-shot settings offset the advantages of PTI with high efficiency in limited resources. We reaffirm that PTI is a method to achieve near-full-shot performance by significantly improving data efficiency in low-resource settings rather than dramatically boosting the performance of state-of-the-art models in the large-resource environment.

C. OUT OF CONTEXTUAL UNDERSTANDING

The PTI represented in the discrete prompt can be an inexplicable sentence that is not consistent with the natural language depending on the setting of the three variables. We manipulate templates that are not contextually proper for each task objective and evaluate the performance. Table 3 presents the case for the counterexamples from the expected result in prompt-based learning. Overall, the implausible template violates the Korean sentence structure and contextual information but achieves better performance than promptless finetuning. In the KLUE-NLI, there is a case in which the implausible template, which consists of unrelated mapping labels, is slightly superior in performance to the plausible template. In another case, an implausible template improves the performance to a larger extent, even when the content and position of the template are completely unrelated to the training purpose of the task and the input sequence in the KLUE-STS. Finally, even in the KLUE-TC, the template with the implausible content can enhance the performance as much as that with the plausible template. We interpret these results for two reasons as follows: (1) Robustness - PTI does not need to strictly follow Korean grammar, sentence structure, and contextual information in the production of templates. In addition, we demonstrate that it is unnecessary to put too much effort into considering natural language in generating the manual template. That is, it is enough for prompt-based few-shot learners to train through roughly engineered PTIs. (2) Forgetting - The flip side is that it is difficult for prompt-based learning to preserve the knowledge of the language learned during the pretraining to address the downstream task. Considering the superior performance of the human-uninterpretable template in the soft-prompt, language models still do not fully understand natural language and are not entirely free from catastrophic forgetting.

TABLE 3. The PTI is manipulated based on the contextual understanding for each task. To estimate the performance, the particular few-shot size (for example, f32) and model (for example, base or large) are set. Except for the plausible template in KLUE-NLI (-0.02), all other cases outperform the promptless finetuning. Based on the three variables in PTI, the explicable component is denoted as ✓ and the inexplicable component as ✗.

KLUE-NLI	0: entailment, 1: neutral, 2: contradiction
Plausible template	Accuracy (f32/base): 39.29 (-0.02)
- Template content	두 문장은 [MASK] 관계이다. ✓
	(The two sentences are [MASK] relationships.)
- Template position	<s_1> <s_2> [t] ✓
- Mapping labels	0: 같은, 1: ##중간 2: ##반대 ✓
	(0: same, 1: ##neutral, 2: ##opposite)
Implausible template	Accuracy (f32/base): 39.62 (+0.31)
- Template content	두 문장은 [MASK] 관계이다. ✓
	(The two sentences are [MASK] relationships.)
- Template position	<s_1> <s_2> [t] ✓
- Mapping labels	0: 딸기, 1: 당근 2: 수박 ✗
	(0: strawberry, 1: carrot, 2: watermelon)
KLUE-STS	0: equivalence, 1: no equivalence
Plausible template	Pearsonr (f32/large): 67.82 (+26.39)
- Template content	[MASK] 내용으로, ✓
	(As the [MASK] content,)
- Template position	<s_1> [t] <s_2> ✓
- Mapping labels	0: 같은, 1: 다른 ✓
	(0: same, 1: different)
Implausible template	Pearsonr (f32/large): 69.31 (+27.88)
- Template content	이 모래알은 [MASK]다. ✗
	(This grain of sand is [MASK].)
- Template position	<s_1> [t] <s_2> ✗
- Mapping labels	0: 동의, 1: 모순 ✓
	(0: agreement, 1: contradiction)
KLUE-TC	0: science, 1: economy, 2: social, 3: culture, 4: world, 5: sports, 6: politics
Plausible template	F1-score (f32/base): 79.90 (+0.30)
- Template content	해당 뉴스는 [MASK] 분야에 해당한다. ✓
	(The news corresponds to the [MASK] field.)
- Template position	<s_1> [t] ✓
- Mapping labels	0: 과학, 1: 경제, 2: 사회, 3: 문화, 4: 세계, 5: 스포츠, 6: 정치 ✓
	(0: science, 1: economy, 2: social, 3: culture, 4: world, 5: sports, 6: politics)
Implausible template	F1-score (f32/base): 79.87 (+0.27)
- Template content	내가 제일 좋아하는 과일은 [MASK]이다. ✗
	(My favorite fruit is [MASK].)
- Template position	<s_1> [t] ✓
- Mapping labels	0: ##과학, 1: ##경제, 2: ##사회, 3: ##문화, 4: ##세계, 5: ##스포츠, 6: 정치 ✓
	(0: ##science, 1: ##economy, 2: ##social, 3: ##culture, 4: ##world, 5: ##sports, 6: politics)

D. OUTSTANDING TEMPLATE

As described in Table 4, we present the outstanding template, which demonstrates the highest performance for all models and few-shot sizes for each task. We consider that the

TABLE 4. Case study for the outstanding template. This table averages the performance of two models (KLUE-RoBERTa-base and KLUE-RoBERTa-large) and three few-shot sizes (16/32/64), identifying the template with the highest improvement. All outstanding templates are contextually plausible (✓) and achieve overall enhancements (+9.91, +12.38, and +0.94).

KLUE-NLI	0: entailment, 1: neutral, 2: contradiction
Outstanding template - Template content - Template position - Mapping labels	Total accuracy: 50.56 (+9.91) [MASK], ✓ ([MASK],) $<s_1>$ $[t]$ $<s_2>$ ✓ 0: 같은, 1: ##중간 2: ##반대 ✓ (0: same, 1: ##neutral, 2: ##opposite)
KLUE-STS	0: equivalence, 1: no equivalence
Outstanding template - Template content - Template position - Mapping labels	Total pearsonr: 60.11 (+12.38) 두 문장은 [MASK] 관계이다. ✓ (The two sentences are [MASK] relationships.) $<s_1>$ $<s_2>$ $[t]$ ✓ 0: 동의, 1: 모순 ✓ (0: agreement, 1: contradiction)
KLUE-TC	0: science 1: economy 2: social 3: culture 4: world 5: sports 6: politics
Outstanding template - Template content - Template position - Mapping labels	Total F1-score: 78.82 (+0.94) 주제는 [MASK]다. ✓ (Topic is [MASK].) $<s_1>$ $[t]$ ✓ 0: 과학, 1: 경제, 2: 사회, 3: 문화, 4: 세계, 5: 스포츠, 6: 정치 ✓ (0: science, 1: economy, 2: social, 3: culture, 4: world, 5: sports, 6: politics)

TABLE 5. Ablation study to validate prompt encoder. A comparative analysis was conducted to decided whether to use Bi-LSTM as a prompt encoder (w Bi) or not (w/o Bi). Further, training was implemented using the model parameters of RoBERTa freeze and only the additional parameters of the Bi-LSTM head (Freeze+w Bi).

KLUE-NLI	Few-shot size 16	Few-shot size 32	Few-shot size 64
RoBERTa-large	37.84	43.23	45.50
RoBERTa-large (w Bi)	49.25	54.93	**58.76**
RoBERTa-large (w/o Bi)	**49.69**	**58.11**	53.84
RoBERTa-large (Freeze+w Bi)	44.43	48.33	51.79
KLUE-STS	Few-shot size 16	Few-shot size 32	Few-shot size 64
RoBERTa-large	24.06	41.43	73.07
RoBERTa-large (w Bi)	35.06	60.15	**75.98**
RoBERTa-large (w/o Bi)	**47.67**	**64.78**	72.51
RoBERTa-large (Freeze+w Bi)	24.00	43.49	46.22
KLUE-TC	Few-shot size 16	Few-shot size 32	Few-shot size 64
RoBERTa-large	74.14	78.24	78.50
RoBERTa-large (w Bi)	54.48	69.10	57.79
RoBERTa-large (w/o Bi)	**76.57**	**76.48**	**75.92**
RoBERTa-large (Freeze+w Bi)	67.62	74.12	75.18

empirical experiments and potentially infinite number of discrete word combinations cannot guarantee the best performance. In addition, considering the results presented in Table 3, plausible templates do not necessarily guarantee higher performance improvements. However, Table 4 indicates that contextual information is not a negligible attribute and can be advantageous in improving performance.

E. PROMPT ENCODER

The soft-prompt PTI in the main result shows the best performance among the various data points that can be combined. As shown in Table 5, we can extend the data points as the setting of the prompt encoder or freezing pretrained language models. Prompt encoder is known as an effective method of maximizing the performance of soft-prompt tuning by addressing problems caused by discreteness as well as association. Contrary to the expected results, the models with Bi-LSTM attached do not consistently outperform the case of simply attaching the pooling layer head without Bi-LSTM. To determine the cause for these results, we evaluate the performance by freezing the learning parameters of the RoBERTa model and training only the light parameters of Bi-LSTM. Using only the model parameters of Bi-LSTM leads to lower performance than promptless finetuning in most cases. Through ablation studies, we find that the Bi-LSTM as a prompt encoder has low capacity to understand contextual representations, offsetting the robustness of PTI. Therefore, the prompt encoder with Bi-LSTM appears to hinder the pretrained model's learning and memory of prior knowledge.

F. MASKED LABEL PREDICTION IN ZERO-SHOT

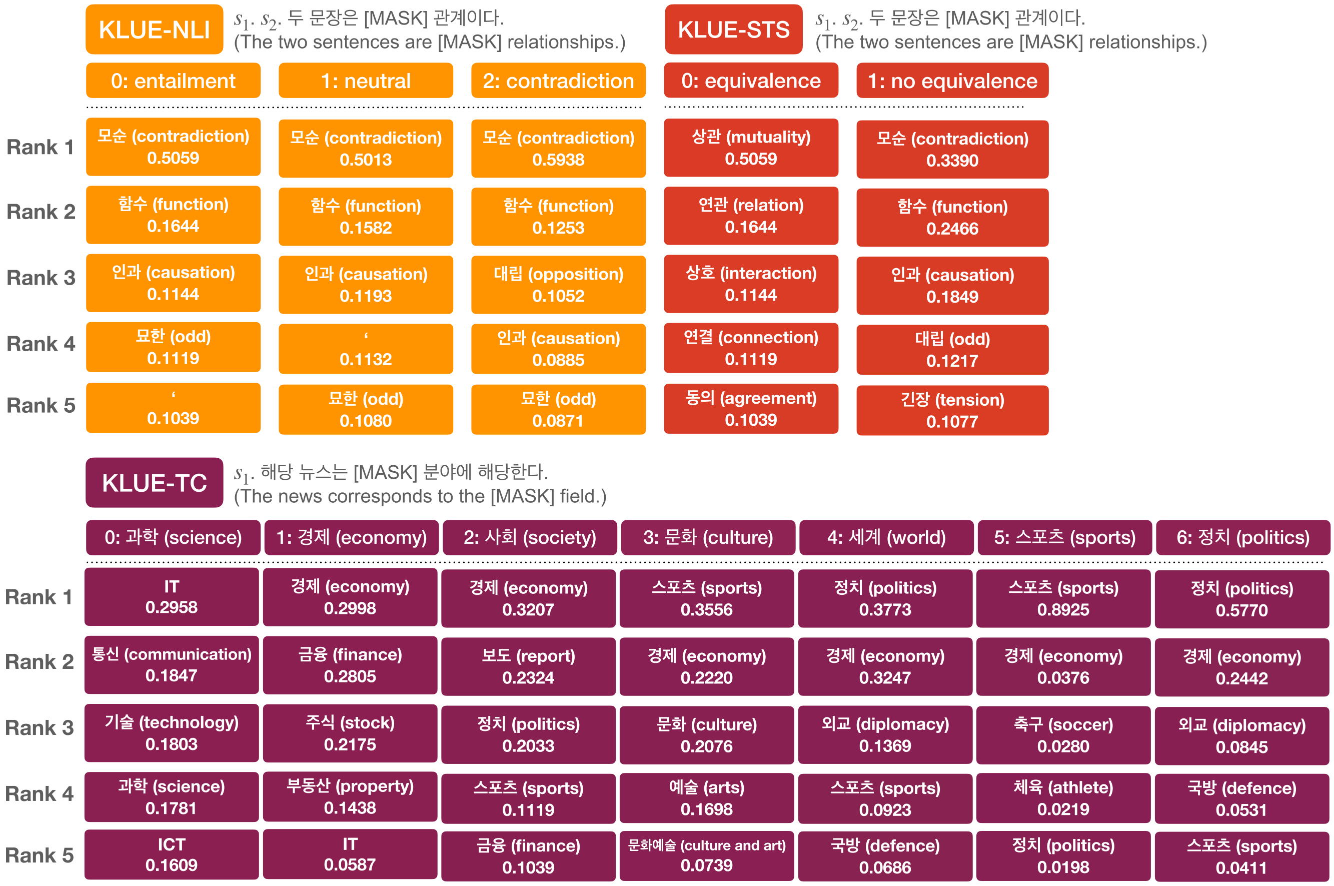
We rank the results of KLUE-RoBERTa-large's prediction of [MASK] of template in the zero-shot setting to track the optimal mapping label for template content and position. Even though the template content and position make sense in human-interpretable, mask token prediction can converge an unrelated word such as "모순 (contradiction)." Figure 2 demonstrates that the pretrained language model already struggles to infer proper mapping labels. Thus, Table 3 presents that the models forget the incomplete inferences of pre-acquired knowledge and newly recognized patterns for mapping labels, showing robust results for mapping labels that are irrelevant to the context. Additionally, prompt-based learning of the PTI method aids the inference that is difficult to complete with only pretrained knowledge and significantly improves the performance.

IX. DISCUSSION AND ERROR ANALYSIS

Although the proposed prompt-based learning outperforms the promptless finetuned baselines, it needs to be noted that there are still unreasonable and burdensome to interpret results in the evaluation. There are three types of limitations and we present the direction of the study to be revealed later.

Firstly, the bias problem that a small amount of data can cause is inherent in our rigorous few-shot learning. Few-shot training examples make the performance gap between seeds large, and the model is challenging to find the optimization point caused by a small number of development examples. For this uncertainty, we have focused on improving performance without compromising the intrinsic purpose of few-shot but have not presented a precise solution to overcome the deviation.

Secondly, Table 5 shows the prompt encoder with auto-prompting, Bi-LSTM, does not significantly impact performance improvement. This result is slightly different from what was claimed in the previous study in soft-prompt tuning [10]. Additionally, it is not clear whether the prompt encoder is capable of resolving discreteness and



KLUE-NLI

s_1. s_2. 두 문장은 [MASK] 관계이다.
(The two sentences are [MASK] relationships.)

	0: entailment	1: neutral	2: contradiction
Rank 1	모순 (contradiction) 0.5059	모순 (contradiction) 0.5013	모순 (contradiction) 0.5938
Rank 2	함수 (function) 0.1644	함수 (function) 0.1582	함수 (function) 0.1253
Rank 3	인과 (causation) 0.1144	인과 (causation) 0.1193	대립 (opposition) 0.1052
Rank 4	묘한 (odd) 0.1119	‘ 0.1132	인과 (causation) 0.0885
Rank 5	‘ 0.1039	묘한 (odd) 0.1080	묘한 (odd) 0.0871

KLUE-STS

s_1. s_2. 두 문장은 [MASK] 관계이다.
(The two sentences are [MASK] relationships.)

	0: equivalence	1: no equivalence
Rank 1	상관 (mutuality) 0.5059	모순 (contradiction) 0.3390
Rank 2	연관 (relation) 0.1644	함수 (function) 0.2466
Rank 3	상호 (interaction) 0.1144	인과 (causation) 0.1849
Rank 4	연결 (connection) 0.1119	대립 (odd) 0.1217
Rank 5	동의 (agreement) 0.1039	긴장 (tension) 0.1077

KLUE-TC

s_1. 해당 뉴스는 [MASK] 분야에 해당한다.
(The news corresponds to the [MASK] field.)

	0: 과학 (science)	1: 경제 (economy)	2: 사회 (society)	3: 문화 (culture)	4: 세계 (world)	5: 스포츠 (sports)	6: 정치 (politics)
Rank 1	IT 0.2958	경제 (economy) 0.2998	경제 (economy) 0.3207	스포츠 (sports) 0.3556	정치 (politics) 0.3773	스포츠 (sports) 0.8925	정치 (politics) 0.5770
Rank 2	통신 (communication) 0.1847	금융 (finance) 0.2805	보도 (report) 0.2324	경제 (economy) 0.2220	경제 (economy) 0.3247	경제 (economy) 0.0376	경제 (economy) 0.2442
Rank 3	기술 (technology) 0.1803	주식 (stock) 0.2175	정치 (politics) 0.2033	문화 (culture) 0.2076	외교 (diplomacy) 0.1369	축구 (soccer) 0.0280	외교 (diplomacy) 0.0845
Rank 4	과학 (science) 0.1781	부동산 (property) 0.1438	스포츠 (sports) 0.1119	예술 (arts) 0.1698	스포츠 (sports) 0.0923	체육 (athlete) 0.0219	국방 (defence) 0.0531
Rank 5	ICT 0.1609	IT 0.0587	금융 (finance) 0.1039	문화예술 (culture and art) 0.0739	국방 (defence) 0.0686	정치 (politics) 0.0198	스포츠 (sports) 0.0411

FIGURE 2. **The results of mask token prediction (i.e., [MASK] is mapping to science) with the KLUE-RoBERTa-large model with the zero-shot setting. We set plausible templates (e.g., The news corresponds to the [MASK] field.) and normalize the score (e.g., 0.2958) for the prediction results up to top-rank 5th (e.g., IT, communication, technology, science, and ICT).**

association. It exhibits low performance when only using additional prompt encoders, and lightweight prompt encoders become less influential as the pretrained model embedding layers become deeper.

Thirdly, as described in Table 3 and 4, we cannot determine which template has the best content, position, and mapping label caused by the limitations of empirical experiments. Moreover, we find that even with contextual meaning and labels in the template inexplicable for training purposes, the performance is sometimes increased and vice versa. These results show the benefits of robustness and easy reproducibility of PTI. However, it is difficult to suggest detailed interpretations of whether language models are possible to overcome catastrophic forgetting entirely and understand natural language like a human.

To address these issues, we can conduct future work in the direction of enhancing the verification scheme within a given low-resource data, such as [17] or removing the bias of the few-shot data. In addition, soft-prompt needs to be studied to sufficiently prove its effectiveness for the prompt encoder. Furthermore, we should research few-shot learners with higher performance improvements and interpretability, minimizing the catastrophic forgetting of language models.

X. IMPLICATIONS OF THE STUDY

We propose Korean-prompt-based few-shot learnings and apply PTI as the prompt engineering method. PTI is a few-shot learning method close to a more practical manner and maximizes the efficiency of given data. Our study is applicable to research and industries that use datasets from domains (for example, fraud and personal information) with considerable collection constraints. In particular, the implications are expected to be more significant in the Korean language, which has strict social norms related to personal information and data collection. Moreover, in developing detection models with class imbalance problems, our study can be utilized to replace traditional sampling methods with prompt-based few-shot learning.

XI. CONCLUSION

As pathfinders in the field of Korean-prompt-based few-shot learning, we conducted an in-depth analysis considering the Korean sentence structure and overcoming data scarcity through rigorous few-shot learning. In this paper, we proposed PTI, which sets a manual template in the suitable position considering Korean contextual information and consists of template content, template position, and mapping labels.

Our prompt engineering method is powerful in the context of Korean and applicable to both hard- and soft-prompt tuning. PTI is robust to the uncertainty of low resources, achieving significant performance improvements in the few-shot learning KLUE-NLI, STS, and TC tasks. To reconsider whether a scheme sufficiently relieves the data scarcity problem, PTI also adheres to using data more practically. In future work, we plan to study whether we can dynamically determine the optimal template for the input sequence through abductive reasoning and contrastive learning within limited resources. We will also attempt to produce the best prompt that can be a guide to achieving performance close to full-shot learning in a few-shot setting. We hope that our proposed PTI and analyses will be a fundamental resource for research on Korean-prompt-based few-shot learners.

REFERENCES

[1] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "*SEM 2013 shared task: Semantic textual similarity," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics, Main Conf. Shared Task, Semantic Textual Similarity*, vol. 1, 2013, pp. 32–43.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 632–642.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[4] G. Cui, S. Hu, N. Ding, L. Huang, and Z. Liu, "Prototypical verbalizer for prompt-based few-shot tuning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7014–7024. [Online]. Available: https://aclanthology.org/2022.acl-long.483

[5] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, vol. 1, Aug. 2021, pp. 3816–3830.

[6] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.Zip: Compressing text classification models," 2016, *arXiv:1612.03651*.

[7] D. Kakwani, A. Kunchukuttan, S. Golla, A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 4948–4961.

[8] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, and H. Hajishirzi, "Generated knowledge prompting for commonsense reasoning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2022, pp. 3154–3169.

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2022, *arXiv:2107.13586*.

[10] X. Liu, K. Ji, Y. Fu, W. Lam Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning V2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," 2021, *arXiv:2110.07602*.

[11] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," 2021, *arXiv:2103.10385*.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[13] R. L. Logan IV, I. Balažević, E. Wallace, F. Petroni, S. Singh, and S. Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," 2021, *arXiv:2106.13353*.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[15] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. [Online]. Available: https://aclanthology.org/2022.acl-long.556

[16] S. Park, "Klue: Korean language understanding evaluation," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–76.

[17] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with checklist," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4902–4912.

[18] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: A literature survey," *Proc. Comput. Sci.*, vol. 189, pp. 19–28, Aug. 2021.

[19] T. Le Scao and A. M. Rush, "How many data points is a prompt worth?" 2021, *arXiv:2103.08493*.

[20] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 255–269.

[21] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Aug. 2021, pp. 2339–2352.

[22] J. Son, J. Kim, J. Lim, and H. Lim, "GRASP: Guiding model with relational semantics using prompt for dialogue relation extraction," 2022, *arXiv:2208.12494*.

[23] T. Sorensen, J. Robinson, C. Rytting, A. Shaw, K. Rogers, A. Delorey, M. Khalil, N. Fulda, and D. Wingate, "An information-theoretic approach to prompt engineering without ground truth labels," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2022, pp. 819–862.

[24] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat, "Sift-based Arabic sign language recognition system," in *Proc. 1st Int. Afro-Eur. Conf. Ind. Adv.* Addis Ababa, Ethiopia: Springer, Nov. 2015, pp. 359–370.

[25] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 353–355.

[26] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "SuperGlue: A stickier benchmark for general-purpose language understanding systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[27] C. Wang, J. Wang, M. Qiu, J. Huang, and M. Gao, "TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, May 2021, pp. 2792–2802.

[28] T. Wolf, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.

[29] L. Xu, "CLUE: A Chinese language understanding evaluation benchmark," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4762–4772.

[30] M. Zhao and H. Schütze, "Discrete and soft prompting for multilingual models," 2021, *arXiv:2109.03630*.

JAEHYUNG SEO received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in computer science and engineering. Currently, he is a part of the Natural Language Processing & Artificial Intelligence Laboratory, under an integrated master's and Ph.D. courses. His research interests include language generation and decoding strategy, where he attempts to find inspiration from how humans do it and build generative model based on commonsense reasoning.

HYEONSEOK MOON received the B.S. degree from the Department of Science in Mathematics and Engineering, Korea University, Seoul, South Korea, in 2021, where he is currently pursuing the Ph.D. degree in computer science and engineering. Currently, he is a part of the Natural Language Processing & Artificial Intelligence Laboratory, under an integrated master's and Ph.D. courses. His research interests include natural language processing, neural machine translation, automatic post editing, and parallel corpus filtering.

CHANHEE LEE received the B.S. degree in computer science and engineering from Sogang University, Seoul, South Korea, in 2013. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul. Currently, he is a part of the Natural Language Processing & Artificial Intelligence Laboratory, under an integrated master's and Ph.D. courses. He is currently working as an AI Research Engineer at NAVER Search U.S. His research interests include language understanding and neural network pruning, where he tries to find inspiration from how humans do it, and build computational models based on this.

SUGYEONG EO received the B.S. degree in linguistics and cognitive science, language and technology from the Hankuk University of Foreign Studies, Yongin, South Korea, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea. Currently, she is a part of the Natural Language Processing & Artificial Intelligence Laboratory, under an integrated master's and Ph.D. courses. Her research interests include neural machine translation and quality estimation, where she tries to predict machine translation quality that minimizes human labor.

CHANJUN PARK received the B.S. degree in natural language processing and creative convergence from the Busan University of Foreign Studies, Busan, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Korea University, Seoul, South Korea. From June 2018 to July 2019, he worked at SYSTRAN as a Research Engineer. He is also working as an AI Research Engineer at Upstage. His research interests include machine translation, grammar error correction, simultaneous speech translation, and deep learning.

JIHOON KIM received the B.S. degree in computer science and engineering from Sungkyunkwan University, Seoul, South Korea, in 2018, and the M.S. degree in computer science and engineering from Seoul University, Seoul, in 2020. He is currently a Research Engineer at the Research and Development Division, Hyundai Motor. His research interests include representation learning, spoken language understanding, and text classification.

CHANGWOO CHUN received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2011 and 2013, respectively. He is currently a Senior Researcher at the Research and Development Division, Hyundai Motor. His research interests include natural language processing, machine learning, and artificial intelligence.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor at the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.