Knowledge Stimulated Contrastive Prompting for Low-Resource Stance Detection

Kai Zheng **Qingfeng Sun** **Yaming Yang*** **Fei Xu**
Microsoft, Beijing, China
{zhengkai,qins,yayaming,fexu}@microsoft.com

Abstract

Stance Detection Task (SDT) aims at identifying the stance of the sentence towards a specific target and is usually modeled as a classification problem. Backgound knowledge is often necessary for stance detection with respect to a specific target, especially when there is no target explicitly mentioned in text. This paper focuses on the knowledge stimulation for low-resource stance detection tasks. We firstly explore to formalize stance detection as a prompt based contrastive learning task. At the same time, to make prompt learning suit to stance detection, we design a template mechanism to incorporate corresponding target into instance representation. Furthermore, we propose a masked language prompt joint contrastive learning approach to stimulate the knowledge inherit from the pre-trained model. The experimental results on three benchmarks show that knowledge stimulation is effective in stance detection accompanied with our proposed mechanism.

1 Introduction

Stance detection is one of the most challenging NLP tasks, which aims to identify the stance of a piece of text towards a given target. Recent years have witnessed rapid progress on learning an intent classification model for target stance detection (Rasooli and Tetreault, 2015; Mohammad et al., 2016; Baly et al., 2018; Stefanov et al., 2020).Though such models in advanced neural architectures (Kaushal et al., 2021) are capable of replying with responses regarding to specific target in a piece of text, to exactly detect the stance of target, background knowledge is often necessary, especially when there is no target explicitly mentioned in text. Take the tweet "I'm confused to hear that Trump win the poll today." as example, the stance detection system should detected the attitude "Against" when given target "Trump". However, when target is "Biden", the stance detected should became "Favour" correspondingly. Furthermore, deep neural networks often require large-scale high-quality labeled training data to achieve state-of-the-art performance (Bowman et al., 2015). However, the cost of labeled data collection is not trival. In this paper, we study the low-resource stance detection tasks, including target-specific stance detection and cross-target stance detection tasks, where only small labeled data is available. Previous works often introduce external data for the models to learn. He et al. (2022) propose to utilize external training knowledge data which could be expensive to obtain.

To solve the above dilemma, many existing stance detection works (Hardalov et al., 2021; He et al., 2022) resort to fine tuning Pre-trained Language Models (PLMs). However, we argue that, in the low-resource stance detection tasks, directly finetuning all parameters of PLMs with small training data (especially when there are less than 100 samples) could result in over-fitting and PLMs simply memorizes the training instances. Recently, several works propose prompt tuning (Schick and Schütze, 2020; Gao et al., 2020a), which only backpropagates the error to Soft Prompts (i.e., a sequence of continuous vectors prepended to the input of PLMs) instead of the entire model. They show that prompt tuning is sufficient to be competitive with full model tuning while significantly reducing the amount of parameters to be tuned. Thus, the prompt tuning is quite suitable to tackle the above over-fitting issue in low-resource finetuning.

Unfortunately, we find a major limitation of task-based prompts that the prompts are too coarse-grained and ignore fine-grained information in input data. In state-of-the-art methods, all input data within a tuning task shares an identical prompt implying the task domain. However, besides the task, the specific content in the input data also contains

* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1168–1178
December 7-11, 2022 ©2022 Association for Computational Linguistics

context that can help the PLM retrieve more relevant knowledge, for example, the specific object being talked about. Such knowledge behinds the input data should be fully exploited to unleash the potential of prompts. The key is that there is a gap between prompt and pre training, because the template in prompt is absent during pre training. In order to make up for this defect, we propose to use cloze task that more familiar with PLM, that is, only do the mask method in the original input to stimulate the richer knowledge learned by the original PLMs. To make PLMs better understand the stance detection task, we further propose a novel contrastive learning objective to boost the PLM to distinguish the different stance.

We conduct experiments with three benchmarks of knowledge-grounded stance detection that are constructed by crowd-sourcing. Evaluation results in terms of both automatic metrics and human judgment indicate that our model not only achieves comparable performance with the state-of-the-art model that is learned from crowd-sourced training sets, even with a small amount of data, our model surpasses the performance of the state-of-the-art model that using whole data and external knowledge, but also exhibits a good generalization ability over different targets and different datasets.

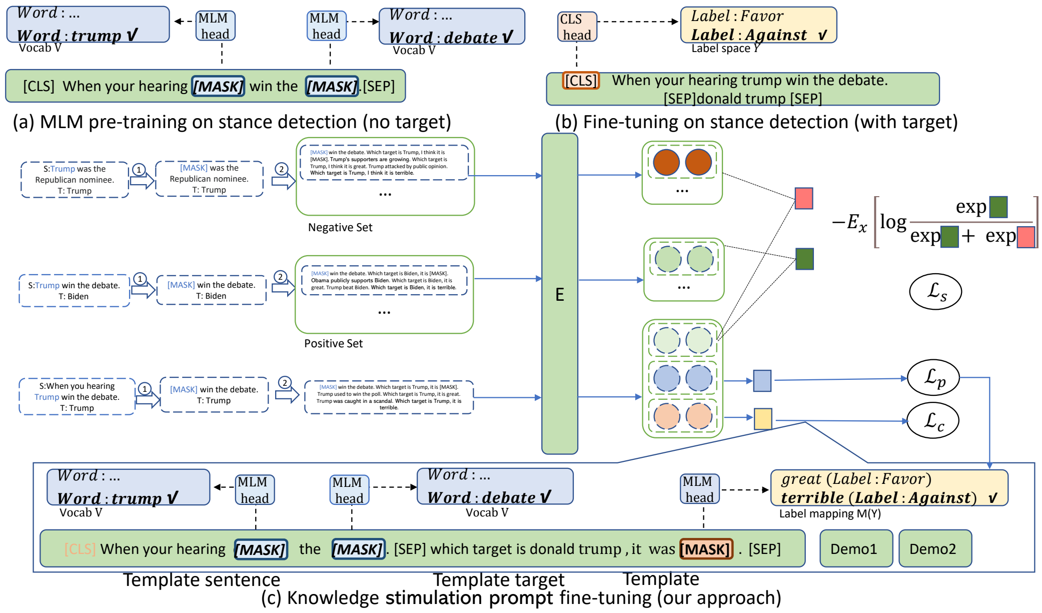
Our contributions are four-fold: (1) exploration of knowledge-grounded stance detection under a low-resource setting; (2) proposal of a knowledge stimulated method that bridge the gap between PLM and prompt tuning; (3) proposal of a unified cloze contrastive prompting learning approach; and (4) empirical verification of the effectiveness of the proposed approach on three benchmarks of knowledge-grounded stance detection.

2 Related Work

Stance Detection Early neural networks based methods have achieved good performances in stance detection task, and they attempt to introduce attention mechanisms (Xue and Li, 2018; Allaway and McKeown, 2020) to capture relationships between sentence and target. And previous works on stance detection (Hardalov et al., 2021) fail to incorporate knowledge in modeling stance. Recently, He et al. (2022) propose to utilize background knowledge from Wikipedia about the target as an external knowledge, the results show that knowledge enhancement is a key factor to improve the performance of stance detection. The wikipedia information couldn't align with the relationship between sentence and target. And the crawled information couldn't cover each target precisely, also could lost sentence information when exceed the max sequence length of model. Furthermore, expert label knowledge is very expensive, and can not be easily obtained. Pre-trained Language Models (PLMs) have shown impressive performance in various NLP tasks, these models usually trained with enormous data to learning the knowledge of each token, word, or entities (Sun et al., 2019). Intuitively, without using the repeated external contextual data to ingest knowledge, we force model to stimulate it's knowledge when training. To the best of our knowledge, no neural networks based methods utilizes the previous potential knowledge to improve their detection performance directly.

Prompt-tuning for PLMs There is an emerging interest in using prompts to extract knowledge from large language models (Chen et al., 2022; Scao and Rush, 2021; Su et al., 2021; Ye et al., 2022; Zhou et al., 2022). Prompt-tuning is a signicant research in the past years. The GPT-3 (Brown et al., 2020) fueled the development of prompt-based learning, which relies on handcraft prompts and achieves impressive per-formance. To further construct automatic prompt, AutoPrompt (Shin et al., 2020) and LM-BFF (Gao et al., 2020b) propose automatic prompt construction through generate discrete tokens. Recently, Prefix-tuning (Li and Liang, 2021), P-tuningV2(Liu et al., 2021), PTR (Han et al., 2021), and soft-prompts (Qin and Eisner, 2021) propose continuous prompt construction, which introduce differential parameters to solve the issue. Unlike previous work (Hardalov et al., 2021), our prompt based framework mainly focuses on SDT, which is the novel exploration of this challenging task in low resource setting.

Contrastive learning Contrastive learning aims to learn representations with self-supervision, so that similar samples are embedded close to each other (positive pair) while pushing away samples that are dissimilar (negative pairs). Such representations have been shown to capture the semantic information of the samples by maximizing the lower bound of the mutual information between two augmented views (Bachman et al., 2019; Tian et al., 2020b,a). Several methods for contrastive learning have been developed so far (Oord et al., 2018; Chen et al., 2020; Dwibedi et al., 2021; He et al., 2020). SupCon (Khosla et al., 2020) is a spe-


(a) MLM pre-training on stance detection (no target)

(b) Fine-tuning on stance detection (with target)

(c) Knowledge stimulation prompt fine-tuning (our approach)

Word : ...
Word : trump √
Vocab V
MLM head
MLM head
Word : ...
Word : debate √
Vocab V
[CLS] When your hearing [MASK] win the [MASK].[SEP]
CLS head
Label : Favor
Label : Against √
Label space Y
[CLS] When your hearing trump win the debate. [SEP]donald trump [SEP]
S:Trump was the Republican nominee. T: Trump
1
[MASK] was the Republican nominee. T: Trump
2
[MASK] win the debate. Which target is Trump, I think it is [MASK]. Trump's supporters are growing. Which target is Trump, I think it is great. Trump attacked by public opinion. Which target is Trump, I think it is terrible.
...
Negative Set
S:Trump win the debate. T: Biden
[MASK] win the debate. T: Biden
[MASK] win the debate. Which target is Biden, it is [MASK]. Obama publicly supports Biden. Which target is Biden, it is great. Trump beat Biden. Which target is Biden, it is terrible.
Positive Set
S:When you hearing Trump win the debate. T: Trump
[MASK] win the debate. T: Trump
[MASK] win the debate. Which target is Trump, it is [MASK]. Trump used to win the poll. Which target is Trump, it is great. Trump was caught in a scandal. Which target is Trump, it is terrible.
E
$-E_x \left[\log \frac{\exp 🟩}{\exp 🟩 + \exp 🟥} \right]$
$\mathcal{L}_s$
$\mathcal{L}_p$
$\mathcal{L}_c$
great (Label : Favor)
terrible (Label : Against) √
Label mapping M(Y)
[CLS] When your hearing [MASK] the [MASK]. [SEP] which target is donald trump , it was [MASK] . [SEP]
Demo1
Demo2
Template sentence
Template target
Template

Figure 1: An illustration of (a) masked language model (MLM) pre-trianing, (b) fine-tuning, and (c) our proposed knowledge stimulation prompt fine-tuning approach.

cial form of contrastive learning that clusters two augmented batches at the class level in the feature space. Thereby, SupCon generates more negative pairs, which is often more efficient in practice.

3 Task Definition

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ is a dataset of instance, $x_i = (s_i, t_i)$ is the instance, where s_i, t_i terms on document sequence and target respectively, $y_i \in \{favor, against, neutral\}$ is the label, where n is the number of examples. The task is to learn a model $p_\theta(y|x)$ without any oracles (e.g., a sentence in Wikipedia) indicating the collation of a sentence, target and the related knowledge.

4 Method

Heading for learning an effective knowledge stimulation model for the task of knowledge-grounded stance detection, we need to deal with several challenges: (1) how to stimulate knowledge without inject external data in stance detection; (2) how to stimulate knowledge with low-resource data.

This section first formulates our proposed Knowledge Stimulated Contrastive Prompting method (KSCP) for low-resource stance detection task. We then introduce the three important components in KSCP, including i) prompt-based learning in pre-trained language models; ii) word cloze task and iii) contrastive learning. Figure 1 shows the overall of KSCP.

4.1 Low-Resource Learning Framework

Figure 2 gives the graphical model of our approach. The model depicts dependency among three variables: dataset D_T, sentence context S, target T, label Y, latent knowledge Z_K^b stimulated by word cloze task, and latent knowledge Z_K^t stimulated by prompt-based learning, where Z_K^b and Z_K^t bridges S and T respectively, Z_K^b and Z_K^t stimulated each other at the same time. prompt-based learning use prompt to stimulate task related knowledge Z_K^b, word cloze task bridges the gap between pretrain and prompt, stimulated more enrichment knowledge from PLMs. Furthermore, Z_K^b is an unsupervised knowledge, don't need labeled data and have a good extensibility. However, Z_K^t is a supervisied knowledge, need labeled data. Finally, Z_K^b apply it's more enhanced knowledge inject to the Z_K^t to make detection of downstream task. Hence, variables endow us with the flexibility of methods, and we can model the objectives of different knowledge levels with respect to target (e.g., from an absent target in the context to an informative statement that delivers necessary content for continuing the detection) in a unified framework. More advantages credited to joint learning include (1) the model and contrast learning protected by each task are more robust to the noise in Z_K^t inferred in the training process; and (2) in terms of prediction, the model can automatically control the expression of knowledge, so it can easily adapt to different scenarios without too much extra effort. The overall objective of learning can be formulated as

$$\mathcal{L}_\theta = \mathbb{E}_{(S,T)\sim D_T}[\log_{\mathcal{M}_\theta}(S,T)] \tag{1}$$

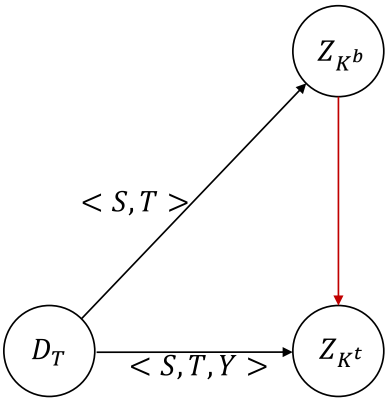


Figure 2: Abstract Logic of the proposed approach.Solid lines mean that there exists stimulation links in both the probabilistic graph and the neural graph, different colors means different strength of stimulation degree.

4.2 Prompt-based Learning

Fine-tuning is a general way to adapt PLM to specific downstream tasks (Devlin et al., 2019). However, for low resource data augmentation, we expect the stimulated synthetic knowledge $\mathcal{K}_{\mathcal{LM}}$ to be different from $\mathcal{K}$, and provide new information for SDT model learning. Fine-tuning PLM may not be an optimal solution, when biased towards a small number of training examples.

Based on prompt learning, start from the zero shot instructions in GPT3 (Brown et al., 2020), keep the whole PLM parameters frozen, and add discrete natural language task instructions (e.g. "translate into English") only before task input. Freezing PLM parameters may help with generalization during training. However, to find a suitable discrete task introduction can not be easily optimized in an end-to-end manner, which requires additional manpower. Compared with the previous method (Brown et al., 2020; Gao et al., 2020b) of generating templates by manual or neural network, we find that the target T in SDT naturally provides hints for prompt construction. Specifically, we design two template mapping based on several heuristic rules: $\mathcal{G}_t$ represents the mapping between the document sequence and the target, $\mathcal{G}_p$ represents the mapping of stance detection.

Let $G_t = \mathcal{G}_t(x)$ denote a target template, and $G_p = \mathcal{G}_p(x)$ denote a instance (contain sentence and target) detection prompt. Then we get input:

$$x_{input} = \text{[CLS]}\ S\ \text{[C]}\ G_t\ \text{[C]}\ G_p\ \text{[SEP]} \tag{2}$$

where [C] is a special separate token. Take input of Figure 1 as example: "I do support Biden. which target is Biden, it was [MASK]", G_t is "which target is Biden", G_p is "it was [MASK]". If a sentence does not conform to these rules, multiple [MASK] tokens will be added directly to the end of the sentence. The number of [MASK] tokens in the prompt is treated as a predefined hyper-parameter l_{mask}. We use demonstrations of label words to construct our input: $x_d = s_0, t_0(\text{[MASK]}), s_i, t_i(word_i)$, where $word_i$ is the label word for sentence s_i towards target t_i, and s_i is sampled from the traing set. During training, we update the parameters by MLM loss:

$$\mathcal{L}_p = \text{MLM}(x_{input}, y) \tag{3}$$

where y is the label word corresponding to x_{input}.

4.3 Word Cloze Task

Prompt-based learning approaches conventionally only use one mask token to predict the label of whole sentence, however, language model doesn't see the prompt template in the pretraining stage, this will lead a gap to stimulate knowledge from PLMs. To tackle the issue, we introduce word cloze task to bridge the gap between pre train, prompt, and fine-tune in SDT.

The word cloze task has a significant impact on knowledge stimulation, especially in low-resource stance detection task. Kawintiranon and Singh (2021) proposes further pre-train the full PLMs parameters through external tweets to enhance knowledge capability. However, this strategy (i.e., full PLM pre-training) introduces huge data collection costs and significant computing overhead. On the contrary, we propose directly training parameters through word cloze task without any external training data. Suppose that knowledge stimulation updates parameters based on partial information (such as keywords) through MLM model, We propose Significant Keywords to Sentence cloze task. Given a piece of text, we use the unsupervised keyword extraction algorithm TF-IDF (Aizawa, 2003) to extract keywords. Given these synonym keywords, the Cloze Sequence is trained to reconstruct original text blocks. When Cloze Task is applied to knowledge stimulation, we only need to fine-tune the Cloze Sequence under the condition of unsupervised learning. This training process is conducted with an prompt-based learning process jointly. We only use the few-shot training data.

Formally, Significant Keywords to Sentence cloze task creates a corrupted version x_c for an input x_{input}:

$$x_c = \text{[CLS]}\ S^{'}\ \text{[C]}\ G_t^{'}\ \text{[C]}\ \text{[SEP]} \tag{4}$$

After constructing this corrupted version of the sequence, MLM aims to predict the masked tokens to recover the original tokens. Then the word cloze task loss becomes:

$$\mathcal{L}_{\mathcal{C}} = \text{MLM}(x_c, x_{input}) \tag{5}$$

Algorithm 1 Optimization Algorithm

```
Require: $s$:number of training iterations
 1: $\mathcal{D}$ : few-shot labeled dataset
 2: $M$ : model
 3: $N \leftarrow 1$
 4: $\mathcal{H}_I \leftarrow \text{GEN}(\mathcal{D}, I)$                ▷ input view
 5: $\mathcal{H}_O \leftarrow \text{GEN}(\mathcal{H}_I, O)$               ▷ output view
 6: $\hat{\mathcal{H}}_{LM} \leftarrow \mathcal{H}_I \cup \mathcal{H}_O$
 7: while $N \neq s$ do
 8:     $\mathcal{L}_S = \text{SupCon}(M^N, \text{H}_{LM})$
 9:     $\mathcal{L}_{\mathcal{P}} = \text{CE}(M^N, \hat{\mathcal{H}}_{LM})$
10:     $\mathcal{L}_{\mathcal{C}} = \text{CE}(M^N, \hat{\mathcal{H}}_{LM})$
11:     $\mathcal{L} = \mathcal{L}_{\mathcal{P}} + \gamma\mathcal{L}_{\mathcal{C}} + \beta\mathcal{L}_{\mathcal{S}}$
12:     $M^N \leftarrow \text{TRAIN}(M, \hat{\mathcal{H}}_{LM})$
13:     $N \leftarrow N + 1$
14: end while
15: return $M$
```

4.4 Multi-View Constrastive Learning

Previous works often restrict the encoder inputs to demonstration or view in a random strategy, such as random demonstration (Gao et al., 2020b) and random view (Jian et al., 2022). The relatively random sample could mislead model with cross target or event result in group together in the latent space. To enrich the positive pairs construction, we propose Multi-View to generate a positive pairs from the input view (conditional on keywords in the input statement) and the output view (conditional on labels).

Figure 1 shows examples of these two views. As shown in Algorithm 1 (line 4 to 5), after fine-tuning the Word Close task in PLMs, KSCP first generates $\mathcal{H}_I$ and $\mathcal{H}_O$ from the input view and output view respectively. KSCP then extracts labels from $\mathcal{H}_I$ and [MASK] statement from $\mathcal{H}_O$. We choose sentences with the same [MASK] tokens and the same label as positive instances, negative instances conversely. In order to reduce the inconsistency caused by masking between training and evaluation, we keep the probability δ mentioned by a phrase unchanged in a direct alignment. In this way, the resulting output text should maintain a higher level of distinguishability and diversity in latent space and stimulate more task/target agnostic novel knowledge.

We use SupCon (Khosla et al., 2020) to compute the contrastive learning loss. To apply SupCon on multiple views of input text, we need to first obtain two views of text:

$x_1 = s_0, G_{t0}, G_{p0}, s_i, G_{ti}, G_{p_i}$
$x_2 = s_0, G_{tk}, G_{p_j}, s_j, G_{tk}, G_{p_j}$

We create candidate demonstration for each input instance according to different G_p and G_t. Let $\tilde{x}_{2b-1}$, $\tilde{x}_{2b}$ be two augmented views of input batch x_b, and r_{2b} and r_{2b-1} are the features of $\tilde{x}_{2b-1}$, $\tilde{x}_{2b}$, then we can compute SupCon loss as:

$$\mathcal{L}_{\mathcal{S}} = \text{SupCon}(r_{2b-1}, r_{2b}, y_b) \tag{6}$$

where y_b is the label for x_b.

4.5 Joint Learning

In order to make the word cloze task and prompt-based learning activated knowledge work together, we propose a joint training method:

$$\mathcal{L} = \mathcal{L}_{\mathcal{P}} + \gamma\mathcal{L}_{\mathcal{C}} + \beta\mathcal{L}_{\mathcal{S}} \tag{7}$$

where γ is loss balance weight of word cloze task, and $\gamma, \beta \in (0.0, 1.0)$. It is worth noting that $\gamma > 0.0$ is to ensure that the parameters of the word cloze task can be optimized through back propagation. $\gamma < 1.0$ is to prevent cloze task loss and reduce the performance of prompt based learning (Zhang et al., 2019). β is the loss balance weight of contrastive learning.

5 Experiments

This section first introduces the experimental settings in Sec 5.1, and then presents the main experimental results in Sec 5.2. Ablation studies were conducted in section 5.3. In section 5.4, we compare KSCP and unlabeled data, and propose diversity analysis.

Model	Trump	Biden	Sanders	Avg.
Fully-supervised Result				
TAN	77.1	77.6	71.6	75.1
BiCE	77.2	77.7	71.2	75.4
PGCNN	76.9	76.6	72.1	75.2
GCAE	79.0	78.0	71.8	76.3
BERT	78.3	78.7	72.5	76.5
BERTweet	82.5	81.0	78.1	80.5
BERTweet♣	85.2	82.5	78.5	82.1
WS-BERT-Dual♣	85.8	83.5	79.0	82.8
Low-resource Result				
BERTweet♠	55.1	35.9	50.3	47.1
WS-BERT-Dual♠	46.2	45.4	52.8	48.1
Sup-Con♠	61.7	68.8	64.3	64.9
KSCP	69.2	73.9	68.4	70.5
KSCP‡	**86.8**	**86.9**	**80.2**	**84.6**

Table 1: Experiment Results of the target-specific stance detection on P-Stance (K = 16). ♣ results taken from (He et al., 2022). ♠ we run (He et al., 2022; Jian et al., 2022)'s source code. ‡refers K = 2048.

Model	Fauci	School	Home	Mask
Fully-supervised Result				
TAN	0.547	0.534	0.536	0.546
ATRGU	0.612	0.527	0.521	0.599
GCAE	0.640	0.490	0.645	0.633
CT-BERT	0.818	0.755	0.800	0.803
CT-BERT-NS	0.821	0.753	0.784	0.833
CT-BERT-DAN	0.832	0.717	0.787	0.825
CT-BERT♣	0.830	0.817	0.836	0.838
WS-BERT-Dual♣	0.836	0.822	0.850	0.866
Low-resource Result				
CT-BERT♠	0.384	0.579	0.517	0.386
WS-BERT-Dual♠	0.364	0.426	0.590	0.407
Sup-Con♠	0.500	0.589	0.490	0.566
KSCP	0.525	0.658	0.504	0.574
KSCP‡	**0.838**	**0.835**	**0.851**	**0.876**

Table 2: Macro-average F1 scores of target-specific stance detection on COVID-19-Stance (K = 16). ♣ results taken from (He et al., 2022). ♠ we run (He et al., 2022; Jian et al., 2022)'s source code. ‡refers K = 2048.

5.1 Experimental Setup

We conduct experiments on PStance (Li et al., 2021) and COVID-19-Stance (Glandt et al., 2021). For each benchmark, we conduct shot-16, 32, 64, 128 experiment following (Gao et al., 2020b). We repeated the experiment 5 times and averaged the macro-F1 according to the previous works (Mohammad et al., 2016, 2017). The **Baseline** model is *BERT-BASE* model, which only uses a few-shot training data $\mathcal{D}$ for training. We use the same hyper-parameters set to train the same *BERT-BASE* model. In P-Stance stance detection tasks, we compare to the baselines TAN (Du et al., 2017), BiCE (Augenstein et al., 2016), PGNN (Huang and Carley, 2018), BERT, and BERTweet. We use WS-BERT-Dual, a state-of-the-art stance detection method, which requires additional external data, and SupCon (Jian et al., 2022), a state-of-the-art prompt learning method. For COVID-19-Stance tasks, TAN, ATGRU (Zhou et al., 2017), GCAE (Xue and Li, 2018), COVID-Twitter-BERT, COVID-Twitter-BERT-NS (Xie et al., 2020), and COVID-Twitter-BERT-DAN (Xu et al., 2020) methods are used. We use the state-of-the-art stance detection method WS-BERT-Dual (He et al., 2022) as a knowledge enhanced method for all tasks.

Implementation Details KSCP is built on the top of the RoBERTa-base (Liu et al., 2019). We use Adam optimizer with learning rate 1e-5, warm-up rate of 0.1 and weight decay 1e-3 to train our model. The number of [MASK] tokens in word cloze task is l_{mask} = 2. We set 16 as batch size. The training of KSCP is conducted on 8 Nvidia Tesla V100 32G GPU cards. Each step, we first training contrastive learning, then jointly train word cloze task and prompt-based learning. The γ in Eq.7 is 0.1. Early stopping on validation is adopted as a regularization strategy. All the hyper parameters are determined by grid search. On target-specific stance detection, we follow the standard train/validation/test splits of the datasets. On cross-target stance detection, the model is trained on the train set of the source target, evaluated on the validation set of the source target, and tested on the combination of train, validation, and test set of the destination target, following the setup in P-Stance.

5.2 Main Results

Target-specific Stance Detection For target-specific stance detection on P-Stance and COVID-19-Stance, we train a model for each target and test it on the same target. Table 1 and Table 2 summarizes the experiment results in shot-16. In both settings, the performance of KSCP are boosted up by a large margin (i.e., 28.5% and 23% for Biden and Trump, respectively). KSCP also outperforms fully fine-tuned PLM methods and wiki-enhanced PLM methods. In general, PLM-based methods can stimulate knowledge better than fine-tuned does. Surprisingly, KSCP achieves better performance than WS-BERT Dual, which uses external in-domain data. This shows that KSCP can potentially reduce the additional manpower required to collect unlabeled in-domain data for low-resource NLU tasks. Figure 3 shows

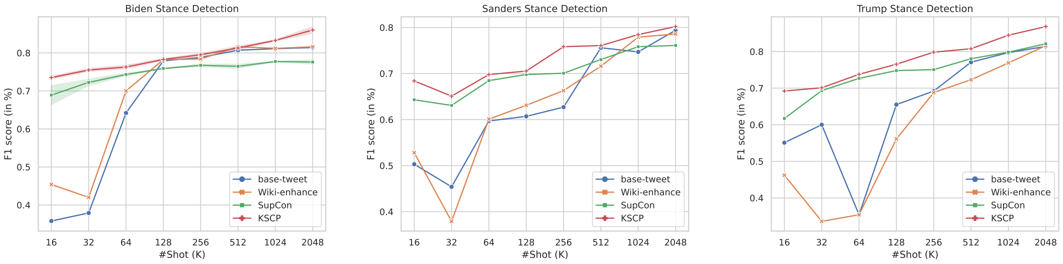


Figure 3: Results of sample efficiency analysis. We compare KSCP with strong baselines with different shot K over Pstance detection task, base-tweet refers to BERTweet♠, Wiki-enhance represents WS-BERT-Dual♠.

the performance in shot-16, 32, 64, 128, 256, 512, 1024, 2048 settings. KSCP is always superior to other systems in all settings. Compared with Biden, Sanders' improvement is smaller. This may be due to the relatively high performance of Sanders baseline.

Cross-target Stance Detection We use P-Stance for cross-target stance detection, where the model is trained on one target, and tested on another target. **Baselines.** We use DualBERTweet and WS-BERT-Dual as strong baselines, which is reported in (Li et al., 2021) and (He et al., 2022). Table 3 shows the experimental results in shot-16. Similar to the results of target-specific stance detection tasks, KSCP significantly improved the performance of the model. KSCP is also superior to various competitive methods, including BERTweet, WS-BERT-D and fully supervised BERTweet. Although WS-BERT-D has high flexibility and introduces external background knowledge data, its performance is lower than BERTweet in low-resource setting. This may be due to the over-fitting problem when using smaller training data for fine tuning. Prompt empowered KSCP successfully avoids this problem and stimulate more knowledge to support model training.

Discussion The performance of WS-BERT-Dual is always inferior to that of KSCP. This is because fully fine-tuned PLMs can easily memorize limited labeled training data. In contrast, the prompt-based learning allows KSCP to maintain high generalization ability and provides new training signals for the model. Compared with the baseline model, the results of KSCP were statistically significant (paired student t-test, $p<0.05$). In the cross-target attitude detection task. We see that WS-BERT Dual is lower than BERTweet in few-shot setting. Depict that infusing Wikipedia knowledge will harm the knowledge capacity of PLMs, however, KSCP is superior to other systems. It is worth noting that the transfer performance of "Trump" → "Biden" is improved more than that of "Trump" → "Sanders". We believe that this is because in the pre-training corpus, the number of co-occurrences of "Trump" and "Biden" is greater than that of "Trump" and "Sanders", so the knowledge about the relationship between "Trump" → "Biden" stored in PLMs is greater than that of "Trump" → "Sanders". In this case, the knowledge stimulus to "Biden" is greater than the information gain brought by the knowledge stimulus to "Sanders", which leads to the performance difference of different target migration. In addition, compared with the performance gain of target-specific stance detection, the performance gain of cross-target task is more significant, which means that when the test target exceeds the training set, knowledge stimulation is more important.

Target	BERTw†	BERTw	WS-BERT-D	KSCP
Trump->Biden	58.9	47.2	37.4	**71.7**
Trump->Sanders	56.5	45.4	39.6	**66.2**
Biden->Trump	63.6	63.8	35.6	**64.6**
Biden->Sanders	**67.0**	56.7	35.3	61.7
Sanders->Trump	58.7	57.7	40.4	**62.2**
Sanders->Biden	73.0	55.4	51.0	**73.2**

Table 3: Macro-average F1 scores of cross-target stance detection on P-Stance. Trump→Biden indicates that the model is trained on "Donald Trump" and tested on "Joe Biden". †refers use Full data.

5.3 Ablation Study

We conduct ablation study on the componential word cloze task and multi-view contrastive learning of Biden, Trump and Sanders' datasets under the shot-16 setting.

Model	Trump	Biden	Sanders
Few-shot Baseline	46.2	45.4	52.8
KSCP	**69.2**	**73.9**	**68.4**
w/o. word cloze task	66.3	70.1	66.5
$AblationforMulti-View\ Contrastive\ Learning$			
output	67.1	70.3	67.9
input	68.1	71.6	66.2
w/o. Multi-View Contrastive Learning	67.0	69.7	66.3

Table 4: The ablation F1 scores over PStance of KSCP for few-shot learning setting. w/o. denotes that we only remove one component from KSCP.

Word Cloze Task As shown in Table 4, word cloze task outperforms the baselines by up to 2%. The results verify the correctness of our motivation and the effectiveness of the word knowledge stimulation. This is probably because the word knowledge pushes the model to focus on the interactions between sentence semantic and its items, and this kind of connection between global information and local information can promote the stance detection task.

Multi-View Contrastive Learning Next, we show the effect of Multi-View Contrastive Learning in KSCP. Input and output generate positive and negative pairs data via the input view and output view, respectively. As shown in table 4, the data pairs in these two single view models have successfully improved the model performance. However, the performance of their corresponding models is worse than that supported by KSCP. This shows that data from different views provide meaningful different training signals for the SDT. Interestingly, models trained in the "output" view perform better than those trained in the "input" view, which indicates that the output pair provides more beneficial positive and negative examples for the task, and can guide the model to better conduct training procedure of contrastive learning.

5.4 Discussion

KSCP with Roberta-Large We verify that KSCP can work with different pre-trained language models. We replace Roberta-base model with Roberta-large model. The new KSCP can also greatly improve the baseline model of few-shot setting. In the Pstance shot-16 setting, the average F1 score of this model is improved from 70.5 to 72.3, which is also better than other models in Table 1.

KSCP in the high-resouce setting In order to show the advantages of KSCP in the high-resource setting, we replace the few-shot training data with complete training data. We found that KSCP can still improve the performance of the baseline model. In Pstance, after knowledge stimulation, the model performance is improved from 70.5 to 72.6 F1 score in average.

γ	0.01	0.02	0.1	0.2	0.3	0.4	0.5	0.6
KSCP	69.9	71.3	**73.9**	72.2	71.3	69.6	69.5	68.3

Table 5: Impacts of loss balance weights on P-Stance.

Joint Learning We examine the effect of Joint Learning in KSCP. In Table 5, we show the model performance with different loss balance weights. It can be observed that, generally, better performance can be achieved with a lower weight loss balance weight in most cases. More specifically, Biden, for example, is less sensitive to seeds because of more training samples and longer sequence length. As the loss balance weight decreases, the advantages of word cloze task, prompt-based learning and KSCP gradually increase. In Eq.7, the loss balance weight of γ equal to 0.1 is always better than other weights in Biden dataset. This is probably because the weight of the word cloze task should not be a huge value, so as to avoid inappropriate influence on prompt-based learning tasks.

Improvement Margin Difference As shown in Table 1, Table 2 and Table 3, the improvement range of cross-target stance detection task is usually greater than that of target-specific stance detection task. This may be because i) the cross-target stance detection task is more fine-grained and knowledge intensive than target-specific stance detection task; ii) The stimulation knowledge of target-specific stance detection task includes more detailed entity background knowledge and boundary. Compared with cross-target stance detection task, PLMs stimulation is more challenging, especially in low-resource settings.

6 Conclusion

In this paper, we present the first prompt-based model KSCP for low-resource stance detection. Experiments on three benchmarks show the effectiveness of our proposed KSCP method. In the future, we plan to expand KSCP to other NLP tasks and other settings, including question answering, machine reading comprehension and text generation tasks.

Limitations

Since KSCP is currently applicable to SDT, more extensive applications need to be considered. Also, KSCP constructs different positive and negative sample pairs based on templates, demonstrations, and labels, it uses SupCon during training, which requires a large GPU memory.

Acknowledgements

We thank for the work of the authors who published the dataset, and anonymous reviewers for their insightful suggestions to improve this paper.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

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

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020a. Making pre-trained language models better few-shot learners.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020b. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *ACL*.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection.

Binxuan Huang and Kathleen Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096, Brussels, Belgium. Association for Computational Linguistics.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. *arXiv preprint arXiv:2205.01308*.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. twt–wt: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *FINDINGS*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, et al. 2021. On transferability of prompt tuning for natural language understanding. *arXiv preprint arXiv:2111.06719*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020a. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020b. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

Chang Xu, Cécile Paris, Surya Nepal, Ross Sparks, Chong Long, and Yafang Wang. 2020. Dan: Dual-view representation learning for adapting stance classifiers to new domains. *arXiv preprint arXiv:2003.06514*.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. 2022. Ontology-enhanced prompt-tuning for few-shot learning. In *Proceedings of the ACM Web Conference 2022*, pages 778–787.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, pages 18–32. Springer.