A Two Stage Adaptation Framework for Frame Detection via Prompt Learning

Xinyi Mou[1], **Zhongyu Wei**[1,2]*, **Changjian Jiang**[3], **Jiajie Peng**[4]
[1]School of Data Science, Fudan University, China
[2]Research Institute of Intelligent and Complex Systems, Fudan University, China
[3]School of International Relations & Public Affairs, Fudan University, China
[4]School of Computer Science, Northwestern Polytechnical University, China
{xymou20,zywei,Changjian}@fudan.edu.cn
jiajiepeng@nwpu.edu.cn

Abstract

Framing is a communication strategy to bias discussion by selecting and emphasizing. Frame detection aims to automatically analyze framing strategy. Previous works on frame detection mainly focus on a single scenario or issue, ignoring the special characteristics of frame detection that new events emerge continuously and policy agenda changes dynamically. To better deal with various context and frame typologies across different issues, we propose a two-stage adaptation framework. In the framing domain adaptation from pre-training stage, we design two tasks based on pivots and prompts to learn a transferable encoder, verbalizer, and prompts. In the downstream scenario generalization stage, the transferable components are applied to new issues and label sets. Experiment results demonstrate the effectiveness of our framework in different scenarios. Also, it shows superiority both in full-resource and low-resource conditions.

Issue	Source	# label	Example	Frame
abortion	tweets	14	Women have the **freedom** to make their own **choices** about their reproductive health care. #AbortionLaw protects the **safety** of women.	Fairness & Equality Health & Safety
gun	news	9	Florida shooter a **troubled loner** with **white supremacist ties**.	Mental Health Race/Ethnicity
gun	debates	5	The **2nd amend** does not say people need guns for self defense !	Legality, Constitutionality and Jurisprudence

Figure 1: A brief landscape of frame detection. These three samples are taken from **twitter** (Johnson et al., 2017), **gvfc** (Liu et al., 2019a), **fora** (Hartmann et al., 2019) respectively.

1 Introduction

Framing is a communication strategy, used to bias the discussion toward a specific stance, by selecting particular aspects of reality and making them more salient (Entman, 1993; Liu et al., 2019a). It is widely adopted by politicians, the media, and the voting public to seek support, express opinions, and advance political agendas (Levendusky, 2013), thus having important implications for public opinion understanding and policy decision-making (Mendelsohn et al., 2021). Boydstun et al. propose 15 generic frame dimensions based on policy agenda, including *economic*, *morality*, and so on, paving the way for frame analysis. In order to analyze the framing strategy automatically, researchers explore the task of frame detection.

Given a piece of statement related to a topic, frame detection aims to recognize which dimensions of frames are employed. Formally, it's a multi-label classification task with a pre-defined label set. Existing researches for automatic frame detection explore different supervised methods including feature-based machine learning, deep neural networks, and fine-tuning pre-trained models (Card et al., 2016; Naderi and Hirst, 2017; Liu et al., 2019b). Despite the success and contribution made by previous studies, these methods ignore the special characteristics of frame detection that events develop quickly and political focus of agendas changes frequently.

We demonstrate the landscape of frame detection in Figure 1 with examples picked from three existing datasets and three challenges stand out. (1) Dynamic nature of languages: framing is used in complex scenarios of quite different styles of language expressions, i.e., coming from different sources and discussing various issues. Previous methods didn't consider learning from existing issues, thus once new issues emerge, data annotation and model training must be repeated. (2) Diverse categories of frame dimensions: frames are defined from different perspectives that can be general like *fairness & equality* or issue-specific like *mental health*. (3) Variance of frames across issues: la-

*Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2968–2978
October 12–17, 2022.

bel set of frames changes for different issues. The complex label system means frame detection can not be easily transferred, unlike traditional tasks including sentiment analysis and stance detection that shares a common label set. In order to address these three challenges and mitigate the limitations of existing research, we propose to develop a generalized framework for frame detection that obtains robust language modeling capability, possesses background knowledge of framing strategy, and can be adapted to new issues easily. Recently, progress in prompt learning provides a mechanism to exploit pre-trained models by task-specific cloze or prefix prompts. Previous research shows its effectiveness in few-shot and zero-shot learning (Schick and Schütze, 2020b; Liu et al., 2021a). Motivated by the success of prompt learning on low-resource sentence classification, we aim to design a prompting framework for frame detection.

In this paper, we propose a two-stage adaptation framework for frame detection based on prompt learning that optimizes prompts and a pre-trained language model for the general domain of framing and applies it to new scenarios. In the framing domain adaptation from pre-training stage, we train a pivot-based encoder on top of the pre-trained language model based on a generic corpus with a general framing label set. Then, we design a shared prompt and several issue-specific prompts and learn their parameters by prompt learning and adversarial training to acquire transferable prompts and a verbalizer. In the downstream scenario generalization stage, we adopt the transferable components to new issues and scenarios. Our contributions are mainly three-fold:

- We propose a generalized prompting framework for frame detection that can deal with different scenarios, issues, and typologies, bridging the gap of cross-issue generalization missing in previous work.
- Different from previous transfer learning, our framework does not train the target data with the source data together, but only reuses the parameters, ensuring data security in the real environment. It is more flexible since data related to politics can be confidential and sensitive in some countries.
- We present the largest study of frame detection, covering 5 datasets from 3 different scenarios. Experiment results show the effectiveness of our framework in both full-source and low-resource situations.

2 Framing Domain Adaptation from Pre-training

Given a sentence x and a set of frame labels $\mathcal{F}$, we aim to detect which frames are used. It is a multi-label classification problem. Since we hope the model can deal with various issues and label sets, it's necessary to urge the model to learn general and transferable knowledge related to frames. In this stage, namely framing domain adaptation from pre-training, we design two tasks to achieve this goal. A generic corpus $\mathcal{C}$ is used as an anchor corpus for the training in this stage.

2.1 Transferable Pivot-based Encoder

Primarily, we design a masked-pivot prediction task (Task1 in Figure 2) to learn a transferable encoder, to capture issue-unrelated features of frames. The process is divided into two steps.

Frame Pivots Generation According to Field et al., some indicators are informative for frame detection, e.g., *cost*, *wage*, *economy* for *economic* frame. We believe these indicators are similar to transferable sentiment words in sentiment classification, which can serve as pivots. We generate frame pivots by mutual information (Church and Hanks, 1990). For a given frame F in $\mathcal{C}$, we calculate point-wise mutual information (PMI) for each word by:

$$I(F, w) = \log \frac{P(F, w)}{P(F)P(w)} = \log \frac{P(w \mid F)}{P(w)} \quad (1)$$

where $P(w \mid F)$ is estimated as $\frac{Count(w)}{Count(allwords)}$ by taking all texts annotated with F, while $P(w)$ is similarly computed using the entire corpus. Words that occur in fewer than 0.5% or more than 98% of documents will be discarded. Finally, we reserve top K words with highest PMI score for each frame, where K is a hyper-parameter.

MLM Training We optimize an encoder on top of a pre-trained language model based on modeling the relationship between pivots and non-pivots following (Li et al., 2020; Ben-David et al., 2020). Concretely, we employ a pre-trained BERT (Devlin et al., 2019), optimized by an MLM objective. As shown in Task1 in Figure 2, given a document, we randomly mask pivots with probability p_v and mask non-pivots with probability p_n, instead of

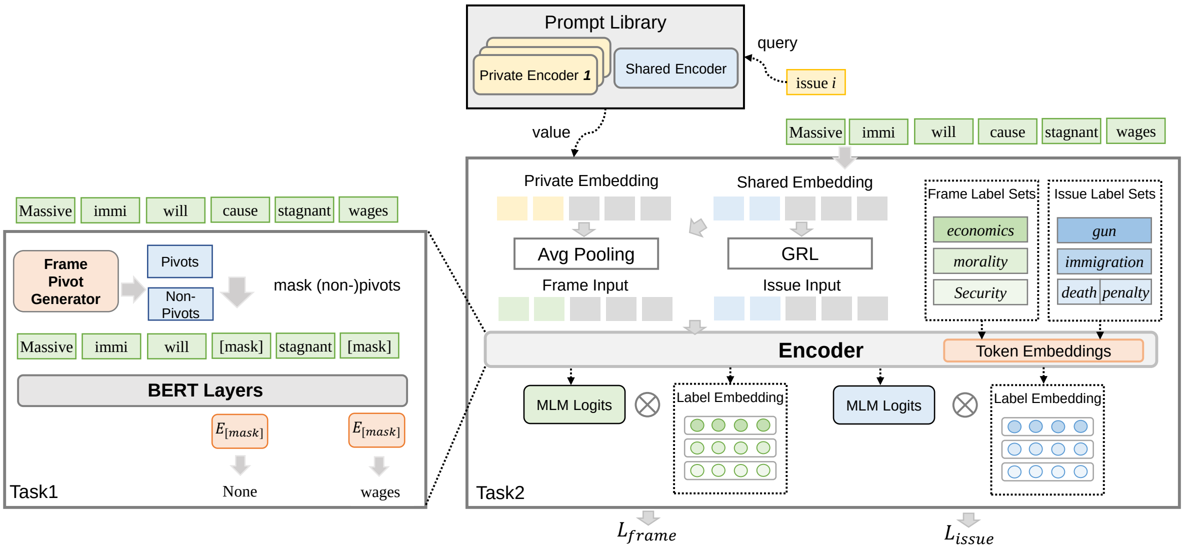

Figure 2: The framework architecture in framing domain adaptation from pre-training stage. In masked-pivot prediction task(left), we train a transferable encoder. In prompt learning-based frame detection task(right), we further optimize the prompts, encoder and verbalizer.

masking each token with the same probability. Like what is done in BERT, for the chosen (non-)pivot words, most of the time they are changed to the mask token, with a small probability of being unchanged. Besides, we don't predict the token on the entire vocabulary size, but only focus on whether the masked token is a pivot or not and which pivot it is. Therefore, it's a classification task with $p+1$ classes, where p is the number of unique pivots. Then, masked cross-entropy is used to calculate the loss for masked token prediction.

$$\mathcal{L}_{MLM} = \frac{1}{\sum m_i} \sum_{i=1}^{n} m_i \cdot L\left(\hat{\mathbf{y}}_i^p, \mathbf{y}_i^p\right) \qquad (2)$$

where $m_i \in \{0, 1\}$ indicates whether the token is masked, and $L\left(\hat{\mathbf{y}}_i^p, \mathbf{y}_i^p\right)$ denotes the cross entropy loss for pivot prediction.

2.2 Transferable Prompts and Verbalizer

To pre-train transferable prompts and verbalizer, we adopt the second task, prompt learning-based frame detection (Task2 in Figure 2). Each component will be described in detail as follows.

Prompt Template To retrieve the knowledge encoded in the pre-trained language model, we design a template to wrap the original input text into a prompt, by discrete natural language or continuous parameters (Liu et al., 2021a,c). To avoid troublesome prompt engineering, we use continuous prompt embeddings. Inspired by Wang et al., for prompt configuration, we build a private prompt template for each issue, and a shared prompt template for all the issues, constituting the prompt library in Figure 2. For each input text x of issue i, the issue-specific template and the shared template will respectively convert the input as:

$$t^{(i)}(x) = w_1^{(i)}, ..., w_m^{(i)}, x, w_{m+1}^{(i)}, ..., w_M^{(i)}, w_{[mask]} \qquad (3)$$

$$t^s(x) = w_1^s, ..., w_m^s, x, w_{m+1}^s, ..., w_M^s, w_{[mask]} \qquad (4)$$

where w_m is prompt pseudo token (Liu et al., 2021c), M is the number of prompt pseudo tokens. A bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with multi-layer perceptrons are applied as prompt encoders. We can get prompt embedding by:

$$PE^{(i)}(x) = MLP(BiLSTM(t^{(i)}(x))) \qquad (5)$$

and the shared template will get:

$$PE^s(x) = MLP(BiLSTM(t^s(x))) \qquad (6)$$

Then, the average of both prompts will result in the final prompt embedding, which is a sequence as the input of the encoder for frame detection, as shown in Figure 2.

Verbalizer Learning for Frames In prompt learning, a mapping from words to labels, e.g.,

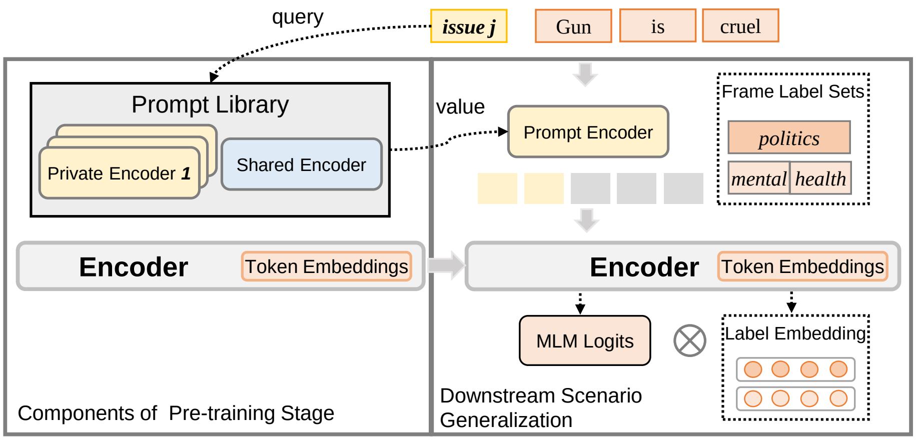

Figure 3: Downstream Scenario Generalization Stage: reuse the transferable components acquired in the framing domain adaptation from pre-training stage.

good, *great* for label *positive*, which is also called verbalizer, is needed for final prediction. However, manual label word selection needs domain-expert knowledge and is defective in dealing with answers of different lengths. Moreover, it's difficult to express the meaning of a frame label by single words, e.g., neither word *school* nor *safety* is a proper answer to a cloze problem about *school safety*. To overcome this problem, we attempt to use the average of the token embeddings to get label embedding. One of the biggest advantages of doing this is the implicit flexibility and transferability of label embedding. For example, once we learned token embedding for label *health* in the pre-training stage, it's beneficial to better understand label *mental health* in a new dataset, since they share the common token *health*. Specifically, we tokenize the original label, extract corresponding token embeddings and take the average as the final label representation:

$$E_l = \frac{1}{n_l} \sum_{i=1}^{n_l} TE(token_i) \tag{7}$$

where n_l is the number of tokens of label l, and $TE(token_i)$ represents token embedding of $token_i$. Then, we calculate the dot product of MLM output with each label embedding. Finally, a binary-cross entropy loss (BCE) is applied for each label. Thus, loss for frame detection is:

$$\mathcal{L}_{frame} = -\frac{1}{N} \sum_{j}^{N} \sum_{k}^{F} y_j^f(k) \log \hat{y}_j^f(k) \tag{8}$$

where $y_j^f(k) \in \{0, 1\}$, N is the number of samples, and F is the number of unique frame labels.

Adversarial Training for Prompts In order to learn a shared prompt that can be easily transferred to new issues, we hope the shared prompt pays attention to cross-issue features related to frames, rather than any specific issue or topic content. To achieve this goal, we design an issue adversarial task to make the shared prompt cannot distinguish issues. Concretely, as shown in Figure 2, we use the shared template only to wrap each input text, and we input this issue prompt to the encoder to predict the corresponding issue. Here, a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) is involved, before the input is sent to the encoder. During the backpropagation, it reverses the gradient by multiplying a negative scalar. For this single-label prediction task, the target is to minimize the cross-entropy loss:

$$\mathcal{L}_{issue} = -\frac{1}{N} \sum_{j}^{N} \sum_{k}^{I} y_j^i(k) \log \hat{y}_j^i(k) \tag{9}$$

where where $y_j^i(k) \in \{0, 1\}$, N is the number of samples, and I is the number of unique issues in the generic corpus $\mathcal{C}$.

In this case, we train the model using following loss:

$$\mathcal{L} = \mathcal{L}_{frame} + \lambda \mathcal{L}_{issue} \tag{10}$$

where λ is a hyper-parameter controlling the weight of different losses.

3 Downstream Scenario Generalization

After the pre-training stage, we get a set of issue-specific prompt encoders, a shared prompt encoder, and an encoder that can be adapted to downstream scenarios. In the generalization stage, we reuse these transferable components for frame detection in new scenarios. As shown in Figure 3, for issues that have been "seen" in the generic corpus $\mathcal{C}$, we initialize the prompt encoder by the corresponding private prompt encoder in the prompt library, and for those unseen issues, the shared prompt encoder is applied for generalization. And encoder along with its token embeddings (used for verbalizer) in the previous stage is also adopted here.

4 Datasets

Based on review of previous work on frame detection, we get datasets available mainly from three resources: (1) news media (articles), (2) social media, (3) debates and statements. Five datasets are listed.

mfc News articles of 15 general frame annotation, constructed by Card et al.. The last version covers 6 issues-*climate change*, *death penalty*, *gun control*, *immigration*, *same-sex marriage* and *tobacco*, with 5,199 articles for each issue on average.

gvfc (Liu et al., 2019a) 1,300 headlines of news articles on *gun violence*, annotated with 9 issue-specific frames.

twitter (Johnson et al., 2017) 2,050 tweets posted by 40 politicians, annotated with 17 general frames (among which 14 are the same with **mfc**, and 3 designed specifically for tweets). 5 issues are included-*abortion*, *aca*, *immigration*, *isis* and *lgbt*. We also treat each as an independent dataset, to verify effectiveness of our framework on unseen new issues like *abortion*.

immi (Mendelsohn et al., 2021) tweets about *immigration*, published by the public, among which 2,325 are annotated with general frames and 1,375 are annotated with issue-specific frames.

fora (Hartmann et al., 2019) 868 arguments on online discussion fora, annotated with 5 most frequent general frames, covering 4 topics.

Following Johnson et al., we drop *other* category in **mfc** and **immi**, and the overview of the datasets is shown in Table 1. Considering that **mfc** is the largest existing corpus with the generic setting of frames, it is used in the framing domain adaptation from pre-training stage. Note that, any other corpus can also be used as the anchor in our framework.

Datasets	Typology	#Label	Size
mfc	general	14	31,960
gvfc	issue-specific	9	1,300
twitter	general	17	2,050
immi	general	14	2,325
	issue-specific	11	1,375
fora	general	5	868

Table 1: Statistics of listed datasets.

5 Experiment and Results

5.1 Experiment Setup

Models for Comparison. We compare our model with some state-of-art methods for frame detection as well as some prompt learning and transfer learning methods.

- Bi-LSTM (Naderi and Hirst, 2017) with initialization by GloVe word embeddings (Pennington et al., 2014).
- Bi-GRU (Naderi and Hirst, 2017) with initialization by GloVe word embeddings.
- FT, fine-tune a pre-trained language model (PLM). (Mendelsohn et al., 2021)
- MP, use a manually-designed prompt and verbalizer to implement prompt learning on specific datasets. We use manual prompt $T(X) = X.\ It\ emphasizes\ [MASK]$., according to definition of framing.
- P-tuning, use vanilla ptuning (Liu et al., 2021c) prompt to implement prompt learning on specific datasets and use label embedding in our framework as verbalizer.
- P-tuning v2 (Liu et al., 2021b), adds prefix into each layer of the PLM and use a fully-connected layer for classification.
- FT-mtl (Sun et al., 2019), fine-tunes PLM by multi-task learning, where frame detection on **mfc** dataset serves as the auxiliary task.
- FT-adv (Hu et al., 2019), on the basis of FT-mtl, it fine-tunes PLM with an additional adversarial domain loss, where domains are issues in **mfc** dataset.

Method	gvfc		twitter		immi-general		immi-specific		fora	
	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF
Bi-LSTM	66.52	59.92	49.26	38.00	46.22	39.84	33.13	21.67	69.87	68.17
Bi-GRU	71.49	60.58	46.25	32.82	42.72	35.98	33.65	19.39	75.29	74.00
FT	83.09	77.29	65.37	56.98	74.38	67.51	60.19	45.71	80.94	81.65
MP	82.38	74.80	62.73	50.27	70.58	60.01	58.11	41.36	78.13	77.89
P-tuning	84.17	77.09	63.38	55.45	74.06	67.23	60.32	46.20	78.64	78.94
P-tuning v2	82.87	76.12	62.31	51.55	71.52	63.87	60.21	**51.41**	78.07	78.63
FT-mtl	84.51	73.02	62.02	52.75	71.54	60.05	62.58	44.84	78.53	79.41
FT-adv	81.50	72.20	59.91	42.36	71.89	63.89	60.94	39.01	75.82	76.22
FT-meta	81.69	75.03	62.25	51.14	68.58	61.22	61.24	46.87	76.43	75.77
Ours	**85.25**	**80.36**	**66.39**	**60.56**	**75.07**	**68.15**	**62.59**	50.66	**81.26**	**82.61**

Table 2: Results in full-resource experiments across datasets. (**Bold**: the best performance in the column)

Method	aca		abortion		isis	
	MiF	MaF	MiF	MaF	MiF	MaF
Bi-LSTM	28.33	8.01	41.60	9.52	38.85	17.46
Bi-GRU	35.55	15.92	36.97	17.89	55.01	25.41
FT	55.29	**30.06**	42.84	15.04	53.91	26.96
MP	54.88	24.74	46.57	13.72	57.16	20.15
P-tuning	54.68	28.18	47.36	17.70	60.32	33.55
P-tuning v2	56.34	26.37	44.44	14.32	**68.04**	29.21
FT-mtl	37.23	14.63	34.28	7.22	67.41	25.11
FT-adv	40.61	13.55	34.89	8.91	64.58	23.66
FT-meta	39.92	13.57	35.71	10.09	64.41	24.61
Ours	**56.72**	29.18	**53.03**	**19.74**	65.19	**36.04**

Table 3: Results in full-resource training on unseen issues in **twitter** dataset. (**Bold**: the best performance in the column)

- FT-meta (Wang et al., 2020), where typical instances of each issues in **mfc** dataset are used to train a meta learner, and the meta learner is further fine-tuned by specific new datasets.

- Ours, method proposed in this paper.

Experiment Settings (1) **Full-resource setting**: all training data are used. (2) **Zero-shot setting**: the training set and validation set are not available. Models are required to detect frames directly on the test set. Since recurrent neural networks and fine-tuning are not applicable, we only compare the prompting-based methods. (3) **Few-shot setting**: N-Way K-shot setup is applied where N is the number of classes in each dataset, and for each class, we take K samples for training and validation respectively. K is set to 2, 4, and 8 in our experiments.

Evaluation Metrics. Since each text can have more than one frame, this prediction task is a multi-label classification task. Results are reported in terms of micro- and macro-F1 score.

5.2 Implementation Details

For pre-trained language model involved in our framework and baselines, we use `bert-base-uncased` (Devlin et al., 2019). For pivot generation, we reserve top 50 frame indicators for each frame dimension. For masking probability, we set $p_v = 0.5$ and $p_n = 0.1$, following (Ben-David et al., 2020). We set training batch size = 16, learning rate of encoder=2e-5, learning rate of prompt encoder = 1e-2, λ = 1. Early stopping strategy is applied to avoid over-fitting. Empirically, we find the best threshold of multi-label classification by validation set for each model respectively. Our implementation[1] is partially based on OpenPrompt (Ding et al., 2022).

5.3 Overall Performance

Results for supervised learning across different methods are reported in Table 2. We can find: (1) Our method has shown improvement compared to all other baselines. Results across different datasets indicate that our method can handle frames of different topics and typologies in a unified framework. (2) Previous study (Vu et al., 2021) has shown that it's not easy for prompt tuning to surpass fine-tuning when using a small PLM like BERT. Benefiting from the pre-training stage, we can achieve comparable results to vanilla fine-tuning. (3) Traditional transfer methods don't perform as well as expected. It demonstrates that it's not easy to optimize multiple objectives due to the gaps between datasets. Encoding knowledge of auxiliary data

[1]Codes are publicly available at `https://github.com/xymou/Frame_Detection`

Method	Avg MiF	Avg MaF
Full Implementation	**74.11**	**68.47**
:w/o pivot-based encoder	72.64	66.69
:w/o issue adversarial training	73.18	66.09
:w/o both	72.55	65.59

Table 4: Results for ablation study. (**Bold**: the best performance in the column)

into PLM in the pre-training stage is an alternative.

Table 3 presents the results individually learned on unseen issues in **twitter** dataset, where training data is less. The results show that learned prompts can be well adapted to unseen issues.

5.4 Ablation study

We implement an ablation study to verify the effectiveness of pivot-based encoder and adversarial training for prompts. When we test without a pivot-based encoder, we use an ordinary BERT backbone. The average MiF and MaF of all datasets in ablation study are in Table 4. It is shown that both components contribute to the framework. This indicates that motivating both encoder and prompts to learn shared, transferable information across issues is crucial for frame detection.

5.5 Zero-shot and Few-shot Analysis

Method	gvfc	twitter	immi-general	immi-specific	fora
MP	43.03	34.48	59.08	19.92	67.21
P-Tuning	43.51	35.71	57.58	20.03	67.60
Ours	**44.65**	**37.44**	**60.30**	**20.28**	**68.96**

Table 5: P@1 in zero-shot experiments. We only compare prompt-based methods since neural networks and fine-tuning with random initialization is not applicable. (**Bold**: the best performance in the column)

Table 5 shows results in zero-shot setting. We didn't report the performance of neural networks, fine-tuning, and P-tuning v2 because a randomly initialized classifier may not produce reasonable results. Since validation data is not available, comparing F1 scores of different models using a random threshold may be unfair, we only illustrate P@1(precision at one). It indicates that prompting methods without fine-tuning can already induce reasonable predictions. Furthermore, with framing domain pre-training, our method shows effectiveness in mining frame-related knowledge in pre-trained language models.

Table 6 shows results in few-shot setting. Overall, prompting methods outperform fine-tuning by a large margin. Meanwhile, our method shows superiority in most datasets. Surprisingly, MP performs best in **immi**-specific. Several factors may account for this result. (1) With a manual template and verbalizer, MP has the least additional parameters and thus has the least risk of over-fitting. (2) Since the labels have some uncommon words like *humanitarian*, having less overlap with those in the pre-training stage, the label embedding verbalizer for this dataset is not well initialized in the generalization stage. Once we replace the verbalizer with a manual mapping, performance can be improved by about 3%. (3) Some data in **immi** is singly annotated without consensus-coding, so the potential noise brings more randomness to few-shot training.

6 Discussion

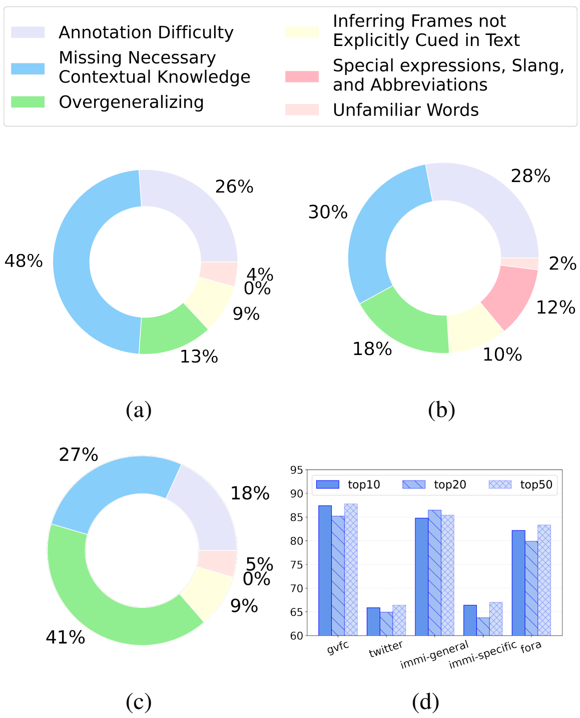

Figure 4: (a)error type proportion in **gvfc**; (b)error type proportion in **twitter** and **immi**; (c)error type proportion in **fora**; (d)P@1 across three setups with a varying number of pivots. (Best viewed in color.)

6.1 Analysis on the Number of Pivots

Here, we explore impact of different number of pivots. We reserve top 10, 20 and 50 frame indicators for each frame in Sec 2.1. Figure 4d presents our results. It is observed that the performance is stable across pivot numbers, especially for datasets having labels with more overlap to **mfc** dataset, i.e., **twitter** and **immi**-general.

Shot	Method	gvfc		twitter		immi-general		immi-specific		fora	
		MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF
2	FT	28.06	22.44	14.59	11.75	29.92	28.53	19.52	18.85	35.69	34.85
	MP	35.86	25.93	26.32	20.79	26.10	18.62	**33.63**	**25.49**	52.55	52.18
	P-tuning	42.73	27.06	29.40	26.39	**52.44**	**49.43**	21.32	15.36	64.22	64.26
	Ours	**43.66**	**29.65**	**29.45**	**26.61**	52.09	47.60	22.91	17.38	**65.92**	**64.85**
4	FT	30.27	25.39	17.12	14.17	33.16	29.13	20.54	18.36	36.54	35.82
	MP	53.05	40.82	37.06	26.99	36.89	26.55	**32.78**	**23.84**	51.09	50.31
	P-tuning	48.49	36.27	35.69	27.08	57.34	52.02	24.24	14.76	66.85	**67.04**
	Ours	**53.18**	**41.03**	**38.90**	**29.33**	**58.44**	**52.86**	26.39	19.34	**66.99**	66.20
8	FT	47.58	40.53	21.74	16.48	41.45	36.46	23.51	19.87	38.27	37.47
	MP	57.88	**52.80**	42.13	31.91	37.11	25.79	**36.44**	**28.19**	65.03	64.40
	P-tuning	56.83	48.88	39.91	33.03	60.20	54.48	31.33	21.02	68.43	67.66
	Ours	**58.46**	50.14	**42.93**	**35.56**	**60.35**	**54.99**	31.97	23.55	**68.49**	**67.97**

Table 6: Results in few-shot experiments, where we randomly sample 2,4,8 samples of each class for training. (**Bold**: the best performance in the column)

6.2 Error Analysis

We identify prediction errors by analyzing 100 random samples for each scenario. On the basis of (Mendelsohn et al., 2021), we add several additional types, and all error types included are shown in the table in Appendix A. Using the sampled instances, we also counted the proportion of each error type to all errors and get Figure 4a, 4b and 4c. We notice that *missing necessary contextual knowledge* is a common challenge, where key lexicons like names of politicians are cued, but the model lacks real-world knowledge or meta-data to make the right induction. External knowledge may be useful to deal with this issue. Also, *overgeneralizing* is a tricky problem, where informative words appearing in different contexts may mislead models. Compared to other models, ours is mainly trapped by this problem, since we have focused on the pivots, which sometimes can be misleading in highly-related frames like *legality, constitutionality and jurisdiction* and *crime and punishment*. Besides, there are scenario-specific challenges. For social media where expression is informal and diverse, slang and abbreviations like hashtags are useful but models may not make full use of them.

7 Related Work

Framing (Entman, 1993) is a communication strategy widely adopted by news media (Card et al., 2015) and politicians (Mou et al., 2021). Most researches on frames focus on news media. Boydstun et al. firstly conclude 15 frames that cross-cut issues from the policy agenda and use them to analyze news articles on 3 issues. Based on this, Card et al. construct Media Frame Corpus (mfc), one of the first large-scale datasets of frame annotations on news articles. This corpus is used to detect frames by dirichlet persona model (Card et al., 2016), deep recurrent neural networks (Naderi and Hirst, 2017) and lexicon analysis(Field et al., 2018). (Liu et al., 2019a) curated Gun Violence Frame Corpus (gvfc), which contains news headlines. They fine-tune BERT (Devlin et al., 2019) for prediction. Frames have also been studied on tweets and statements of politicians and the public. Johnson et al. annotated some tweets of politicians and use the Probabilistic Soft Logic model to detect frames. Roy et al. use a similar method to identify Morality frames in tweets. Since framing is an important communication strategy, frame detection in online fora and debates also aroused researchers' interest (Hartmann et al., 2019; Ajjour et al., 2019; Heinisch and Cimiano, 2021).

According to Boydstun et al., there are mainly two frame schemas, i.e., general and issue-specific. Although most work focuses on general frame prediction, some studies specialize in frames designed for a certain issue. Mendelsohn et al. fine-tune RoBERTa (Liu et al., 2019b) for analysis of different typologies of frames in immigration. Liu et al. conclude 9 issue-specific frames for gun violence issue.

Prompt learning aims to wrap the original input text using a template with a cloze or prefix prompt, and then the language model is used to fill the unfilled information to obtain a final string, which will be mapped into labels by a verbalizer (Schick and Schütze, 2020a; Liu et al., 2021a). This paradigm performs well on few-shot and zero-shot settings (Brown et al., 2020). Prompts can be manually designed or learned with differentiable parameters (Liu et al., 2021c). Despite progress in prompt learning, few studies explore the transferability of prompts. Recently, (Vu et al., 2021; Wang et al., 2021) verify the effectiveness of reusing prompts of similar tasks to realize task transfer.

Massive work on cross-domain applies domain adversarial approaches to learn domain-invariant features (Ganin and Lempitsky, 2015; Ganin et al., 2016). Some researches on cross-domain sentiment classification extract common shared features, called pivots, which are frequent in both source and target domains and are prominent (Ziser and Reichart, 2018; Ben-David et al., 2020). Learning pivots and non-pivots help capture features for the task, rather than a specific domain.

8 Conclusion and Future Work

In this paper, we propose a general framework for frame detection that can handle various issues and frames. With the help of domain adaptation techniques, we enable both the encoder and prompts to learn transferable knowledge related to frames, thus yielding improvement on several datasets. Taking advantage of prompt learning, the framework can also deal with low-resource scenarios. In the future, we plan to conduct experiments on other formulations of framing analysis, e.g., diffusion of frames.

Acknowledgement

This work is partially supported by Natural Science Foundation of China (No.6217020551, 71991471), National Social Science Foudation of China (No. 20ZDA060), Science and Technology Commission of Shanghai Municipality Grant (No.20dz1200600, 21QA1400600, 21511101000) and Zhejiang Lab (No. 2019KD0AD01).

References

Yamen Ajjour, Milad Alshomary, et al. 2019. Modeling frames in argumentation. In *Proc. of EMNLP*.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. Perl: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *TACL*.

Amber E Boydstun, Dallas Card, et al. 2014. Tracking the development of media frames within and across policy issues. *Technical report, University of California, Davis.*

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dallas Card, Amber Boydstun, et al. 2015. The media frames corpus: Annotations of frames across issues. In *Proc. of ACL*.

Dallas Card, Justin H Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proc. of EMNLP*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *CL*.

Jacob Devlin, Ming-Wei Chang, et al. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proc. of ACL: System Demonstrations*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*.

Anjalie Field, Doron Kliger, et al. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Proc. of EMNLP*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proc. of ICML*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, et al. 2016. Domain-adversarial training of neural networks. *JMLR*.

Mareike Hartmann, Tallulah Jansen, et al. 2019. Issue framing in online discussion fora. In *Proc. of NAACL*.

Philipp Heinisch and Philipp Cimiano. 2021. A multi-task approach to argument frame classification at variable granularity levels. *it-Information Technology*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Mengting Hu, Yike Wu, Shiwan Zhao, Honglei Guo, Renhong Cheng, and Zhong Su. 2019. Domain-invariant feature distillation for cross-domain sentiment classification. In *Proc. of EMNLP*.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter. In *Proc. of ACL*.

Matthew S Levendusky. 2013. Why do partisan media polarize viewers? *American Journal of Political Science*.

Liang Li, Weirui Ye, Mingsheng Long, et al. 2020. Simultaneous learning of pivots and representations for cross-domain sentiment classification. In *Proc. of AAAI*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Siyi Liu, Lei Guo, et al. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proc. of CoNLL*.

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

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021c. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proc. of NAACL*.

Xinyi Mou, Zhongyu Wei, Lei Chen, Shangyi Ning, Yancheng He, Changjian Jiang, and Xuanjing Huang. 2021. Align voting behavior with public statements for legislator representation learning. In *Proc. of ACL*.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. *Policy*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proc. of EMNLP*.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. In *Proc. of EACL*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. In *Proc. of NAACL*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. *arXiv preprint arXiv:2003.13003*.

Chengyu Wang, Jianing Wang, Minghui Qiu, et al. 2021. Transprompt: Towards an automatic transferable prompting framework for few-shot text classification. In *Proc. of EMNLP*.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proc. of NAACL*.

A Error Types

Table 7 shows the common error types mentioned in Error Analysis.

Code	Error Type	Description	Examples
1	Annotation Difficulty	These instances highlight the challenges of annotation: there are convincing arguments that model's predicted frames can be appropriate labels.	*Oh and the death penalty does not deter crime.* Model predicted *Policy Evaluation*.
2	Missing Necessary Contextual Knowledge	Some frames are cued by lexical items but lack real-word knowledge or meta-data to induce the frames. E.g., model doesn't know the author of tweets is politicians so that it can not recognize "my collegues" refers to Political Factor.	*Joined 210 of my colleagues in urging supremecourt to ensure equal marriage rights.* Model missed *Political Factor*.
3	Overgeneralizing	Some words are highly related. The model makes erroneous predictions when such features are used in different contexts.	*American bombs aren't yet falling on syria, but chuck hagel suggested they will.* Model erroneously predicted *Security*.
4	Inferring Frames not Explicitly Cued in Text	Model predicts frames that may capture an author's intention but without sufficient evidence from the text.	*Stop immigration.* Model erroneously predicted *Public Order*.
5	Special Expressions, Slang, and Abbreviations	Some special terms, hashtags, abbreviations that indicate certain frames but not captured by the model.	*American tax dollars must not be used to aid and abet any dictatorial regime that stands with terrorists! #noaid2egypt* Model missed *External Reputation*.
6	Unfamiliar Words	Some unfamiliar but important clues are mentioned. Since they appear infrequently during training, language models may not understand them well enough .	*Dementia complicates us gun ownership.* Model missed *Mental Health*.

Table 7: Common errors in frame detection.