Template-Free Prompting for Few-Shot Named Entity Recognition via Semantic-Enhanced Contrastive Learning

Kai He, *Member, IEEE*, Rui Mao, Yucheng Huang, Tieliang Gong,
Chen Li, *Member, IEEE*, and Erik Cambria, *Fellow, IEEE*

***Abstract*— Prompt tuning has achieved great success in various sentence-level classification tasks by using elaborated label word mappings and prompt templates. However, for solving token-level classification tasks, e.g., named entity recognition (NER), previous research, which utilizes N-gram traversal for prompting all spans with all possible entity types, is time-consuming. To this end, we propose a novel prompt-based contrastive learning method for few-shot NER without template construction and label word mappings. First, we leverage external knowledge to initialize semantic anchors for each entity type. These anchors are simply appended with input sentence embeddings as template-free prompts (TFPs). Then, the prompts and sentence embeddings are in-context optimized with our proposed semantic-enhanced contrastive loss. Our proposed loss function enables contrastive learning in few-shot scenarios without requiring a significant number of negative samples. Moreover, it effectively addresses the issue of conventional contrastive learning, where negative instances with similar semantics are erroneously pushed apart in natural language processing (NLP)-related tasks. We examine our method in label extension (LE), domain-adaption (DA), and low-resource generalization evaluation tasks with six public datasets and different settings, achieving state-of-the-art (SOTA) results in most cases.**

***Index Terms*— Contrastive learning, few-shot learning, information extraction, named entity recognition (NER), prompting.**

I. Introduction

NAMED entity recognition (NER) aims to detect entity spans from unstructured natural language and classify the entities into predefined types, such as LOCATION, PERSON, and EVENT. NER lays the foundation of many downstream tasks, such as question answering [1], recommend system [2], and knowledge graph construction [3]. Most existing NER studies [4], [5] are trained with large amounts of annotated data. However, large-scale manual annotations for supervised learning NER in a wide range of domains are cumbersome [6]. To this end, utilizing few-shot techniques in resource-constraint settings is a promising method to mitigate labor efforts and cross-domain challenges.

Recently, prompt-based research has shown great potential on few-shot learning tasks by reformulating various downstream tasks as mask language learning tasks [7], [8], [9], [10], [11]. Most prompt-based methods first construct semantic templates as prompts to obtain masked word predictions from a pretrained language model (PLM), then map these predictions into task-specific labels [12], [13]. Such a process is termed label word mappings [14]. However, manual construction of templates and label word mappings are cumbersome and subjective. The nuances in prompt templates and label word mappings may result in a huge difference in model performance [15]. Considering the above problems, there is more research focusing on generating prompts automatically and improving label word mappings [13], [16], [17]. Some studies achieved improvements by utilizing soft prompts instead of natural language-based prompts [8], [18]. These soft prompts are normally continual embeddings in embedding space, given by a PLM. However, the study [19] finds that there is no statistically significant difference in performances when using instructive or misleading prompts. The work [20] just concatenates a [MASK] special token with an input, which can achieve competitive performances with manually written prompts. This motivates us to explore whether an elaborately designed template is necessary and what really works in prompt-based methods.

Besides, prompts-based methods are intrinsically designed for sentence-level tasks [9], [21]. When prompt tuning comes to token-level NER, it needs N-gram traversal to query all the possible combinations of spans and types or use different prompts with repeatedly forwarding to obtain a single prediction [22], [23], [24]. As shown in Fig. 1, given an input "Franklin Archibald Dick is a famous lawyer in Franklin. [Span] is a [Type] entity," a typical prompt-based method needs to iteratively fill all spans in the [Span] position, such as "Franklin," "Franklin Archibald," and "Franklin Archibald

Manuscript received 24 November 2022; revised 21 April 2023; accepted 10 September 2023. Date of publication 26 September 2023; date of current version 3 December 2024. This work was supported in part by the Key Research and Development Program of Ningxia Hui Nationality Autonomous Region under Grant 2022BEG02025, in part by the Key Research and Development Program of Shaanxi Province under Grant 2021GXLH-Z-095, in part by the Industry Alignment Fund â Industry Collaboration Projects (IAF-ICP) Funding Initiative under Grant RIE2020, in part by the Innovative Research Group of the National Natural Science Foundation of China under Grant 61721002, and in part by the Innovation Research Team of the Ministry of Education through the Project of China Knowledge Centre for Engineering Science and Technology under Grant IRT_17R86. *(Corresponding author: Erik Cambria.)*

Kai He, Yucheng Huang, Tieliang Gong, and Chen Li are with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: hk52025804; huangyucheng@stu.xjtu.edu.cn; gongtl@xjtu.edu.cn; cli@xjtu.edu.cn).

Rui Mao and Erik Cambria are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: rui.mao@ntu.edu.sg; cambria@ntu.edu.sg).

Digital Object Identifier 10.1109/TNNLS.2023.3314807

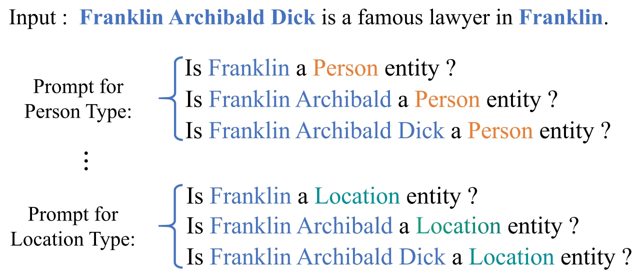

Input : Franklin Archibald Dick is a famous lawyer in Franklin.
Prompt for Person Type:
Is Franklin a Person entity ?
Is Franklin Archibald a Person entity ?
Is Franklin Archibald Dick a Person entity ?
⋮
Prompt for Location Type:
Is Franklin a Location entity ?
Is Franklin Archibald a Location entity ?
Is Franklin Archibald Dick a Location entity ?

Fig. 1. Example of redundancy problem when applying prompt tuning for sequence labeling-based NER task.

Dick." Meanwhile, all predefined types in a label set need to iteratively fill in the [Type] position for each span, such as CITY and PERSON, to differentiate "Franklin Archibald Dick" and "Franklin." Obviously, such a method suffers catastrophic time costs when sentence length or entity types increase.

To tackle these modeling issues, we propose template free-prompting (TFP) for few-shot NER via semantic-enhanced contrastive learning. TFP employs prior knowledge to initialize semantic anchors for each entity type in the vector space. The prior knowledge is obtained from Wikipedia[1] to represent the definition of labels in natural language. Such prompts are understandable for humans compared with soft prompts. Then, the semantic anchors are simply appended with the embeddings of an original sentence as prompts without template construction and label word mappings. Finally, semantic anchors are in-context-encoded together with the input sentence to form the prototypes of entity types. Noticeably, these prototypes are context-dependent, because different inputs have different original sentences for the in-context encoding. By the comparison between each token in an input sentence with these in-context-encoded prototypes, TFP allocates a label for each token and parses the results as normal IO-based NER (namely binary classification for each token), avoiding the issues of N-gram traversal and appending different prompts for the same sentence.

Inherently, such a comparison can be achieved by contrastive learning [25]. However, traditional contrastive learning cannot be used in few-shot learning, because it needs a large volume of negative samples [26] that cannot be supported in few-shot settings. Furthermore, when previous contrastive learning [27], [28], [29] developed a negative sample set, the negative instances were naively considered as nonpositive instances without comparing the semantic similarity between positive and negative instances. This results in an issue in that many negative instances share similar semantics to a positive instance, whereas the negative ones are undesirably pushed away to the positive one in vector space.

To overcome the learning issues, a hybrid granularity contrastive loss is developed in our TFP. The loss aims to optimize the distances between tokens with calculated semantic prototypes, instead of typical token-wise distances. Meanwhile, the loss also optimizes the distances between different prototypes. Since the above prototypes are initialized with semantic anchors, they can alleviate the bias from randomly sampled data and mean-based prototyping under the few-shot setting [30]. By contrastive learning presentations of introduced semantic information and input tokens, our loss can be used in few-shot settings without using many negative samples.

We demonstrated that the proposed TFP is robust and generalizable by evaluating its abilities in label extension (LE), domain-adaption (DA), and no-adapting (NA) under few-shot settings. Specifically, TFP is tested with 26 subtasks, six employed datasets, and three different few-shot NER setups, achieving better performance on 19 tasks. For example, compared with the strongest baseline, the proposed TFP raises an averaged F1 measure of 11.37% and 8.95% in 1- and 5-shot I2B2 under DA settings. Also, various analysis experiments are carried out to demonstrate its effectiveness. Our contributions can be summarized as follows[2]:

1) We propose an effective template-free prompt (TFP)-based method for few-shot NER. The method aims to address the cumbersome template construction and N-gram traversal-based inference when prompt learning is employed in token-level labeling tasks.
2) We propose a novel semantic-enhanced contrastive learning loss. The loss can achieve contrastive learning in a few-shot context, yielding more effective and distinguishable representations for positive and negative samples by their semantics.
3) We conduct detailed comparisons and analysis to explore what really works in prompt-based methods and find that in-context encoding plays a more important role than elaborately designed prompts.
4) We conduct three few-shot learning evaluation tasks to evaluate the capacity of our model in LE, domain adaption, and low-resource generalization. Our proposed method achieves 19/26 state-of-the-art (SOTA) results in these few-shot NER evaluation tasks.

II. Related Work

A. Few-Shot NER

Numerous practical challenges still persist in NER tasks, such as multimodel NER [31], discrete NER [32], and few-shot NER [33]. The primary emphasis of this article is on addressing the challenges associated with few-shot NER. Many advanced natural language processing (NLP) applications and specific scenes need such technology, such as dialogue systems [34], [35], personalized recommendations [36], [37], and handling long tail data distributions [38], [39]. The study [40] represents an early effort that concentrates on the few-shot NER task. The researchers have put forth an end-to-end trainable memory network, which has the ability to identify and differentiate named entities in an online fashion. The network is capable of performing one-shot learning and can cope with a limited number of sparse supervisions. According to METABDRY [41], presently available NER methods are encountering difficulties in dealing with sparse boundary tags. In addition, when the source domains differ from the target domains, existing methods require more training data to adapt to the new domains. To address these challenges, METABDRY employs adversarial learning to encourage the development of domain-invariant representations. Furthermore, they utilize meta-learning to explicitly

[1] https://dumps.wikimedia.org/

[2] Code and data will be released after review.

simulate domain shifts during training, thereby enabling effective aggregation of meta-knowledge from multiple resource domains. The work presented in [42] utilizes synthetic data augmentation to simultaneously tackle few-shot and incremental learning for NER. PCBERT [43] proposes a novel Parent and Child BERT method for Chinese few-shot NER, where an annotating model is first trained on high-resource datasets to discover implicit labels on low-resource datasets. SDNet [44] proposes a self-describing mechanism for few-shot NER, which can leverage illustrative instances and precisely transfer knowledge from external resources by describing both entity types and mentions using a universal concept set. In contrast to the aforementioned methods, our proposed TFP method focuses on a simple yet efficient prompt-based approach that can unlock the true potential of large language models (LMs) without requiring complex changes to the model structure.

B. Prompt Learning

The early studies [9], [21] explore manually constructing prompts for sentence-level text classifications, which reformulate downstream tasks as cloze questions with a PLM. Considering manual prompts are troublesome and subjective, some studies propose automated methods for prompt creation. P-tuning [15] proposes soft prompts, which employ continual embeddings as prompts rather than natural language. They first employed trained parameters as continuous prompts and further used long short-term memory (LSTM) to fuse contextual information. Also, this study found that inserting anchor words can effectively improve the performance of automatically generated prompts. This method achieves significant improvement over the traditional fine-tuning method in the knowledge detection task. The idea of Prefix-Tuning [8] is similar to P-tuning, where the model only optimizes a small number of parameters in the process of training. The difference is that Prefix-Tuning adds a small number of parameters to each layer of the LM, which do not need to correspond to any specific word. Prompt tuning with rules (PTRs) [12] applied logic rules to construct auto-generated prompts. AutoPrompt [13] utilizes gradient-guided search to automatically generate prompts for diverse tasks. The study [45] further investigated the performance of prompt tuning on various LMs. The study pointed out that a key advantage of prompt tuning is that it can freeze the entire PLM and accomplish a given predictive classification task by only tuning a small number of parameters. Therefore, this method can be of great practical value in the application of large-scale PLMs. At the same time, the study concludes experimentally that this method can only perform on par with the fine-tuning method when using very large-scale PLMs (10B parameters or more). On the contrary, our proposed TFP shows strong prediction ability with small LMs.

The above prompt-based methods are designed for sentence-level tasks. For token-level tasks, such as NER, prompting each token with all potential classes is challenging. The work [22] proposes a template-based method for prompting NER, which enumerates all possible spans of input sentences combined with all entity types to predict labels. This method suffers serious redundancy when sentence length or entity types increase. COPNER [46] introduced class-specific words into prompt tuning, following the idea of distance metric learning to compare each token with manually selected class-specific words. Although this method avoided enumeration of all possible spans, manual selection for class-specific words is still labor-intensive and the method is sensitive to selected class-specific words. The work [47] tries to explore a prompt-free method for few-shot NER. This study proposes entity-oriented LM fine-tuning to directly decode input tokens to corresponding label words and then maps these label words to related labels. However, this method heavily depends on the label word mapping. Compared with the above studies, our TFP needs neither template construction nor label word mapping, which is more effective and high-performing.

C. Contrastive Learning

The goal of typical contrastive learning [48] is constructing a representations space where instances from the same input are pulled closer and instances from different inputs are pushed apart, regardless of their semantic information. Contrastive learning is widely utilized in the field of computer vision [26], [49], [50], where we can easily construct an augmentation for an image by flipping, rotating, and cropping to form positive pairs. Contrastive clustering [51] and twin contrastive learning (TCL) [52] combine an instance- and cluster-level contrastive learning with clustering methods, achieving significant improvements on CIFAR [53] and ImageNet [54] datasets. Our hierarchical contrastive loss shares similarities with instance- and cluster-level contrastive learning. However, for image-related tasks, there is no requirement to consider semantic consistency. The partially view-aligned problem is addressed in partially view-aligned problem (PVP) [55] using a noise-robust contrastive loss, which focuses on alleviating the influence of the false negative pairs. In contrast, our loss is designed to handle true negative pairs that have negative effects on specific tasks.

In NLP tasks, randomly inserting, deleting, or switching tokens are not perfect methods [56] for data argumentation, because these processes may cause incoherence or even incoherence meaning. SimCSE [27] proposed a novel method for sentence argumentation by repeatedly forwarding a sentence with different dropout results, achieving strong contrastive learning on textual similarity tasks. CADAN [57] has introduced a contrastive approach that involves dividing the feature extractor into two contrastive branches. One branch is responsible for capturing the class-dependence in the latent space, while the other focuses on achieving domain-invariance. To fulfill these contrasting objectives, CADAN shares the first and last hidden layers but maintains decoupled branches in the middle hidden layers. CoLA [58] explores contrastive learning in anomaly detection tasks with a graph neural network, which exploits the local information by sampling a novel type of contrastive instance pair. CLEAR [59] proposed a sentence-level contrastive learning method, which utilized random-words-deletion, spans-deletion, synonym-substitution, and reordering as augmentation strategies to learn a noise-invariant representation. DeCLUTR [60] focused on how to

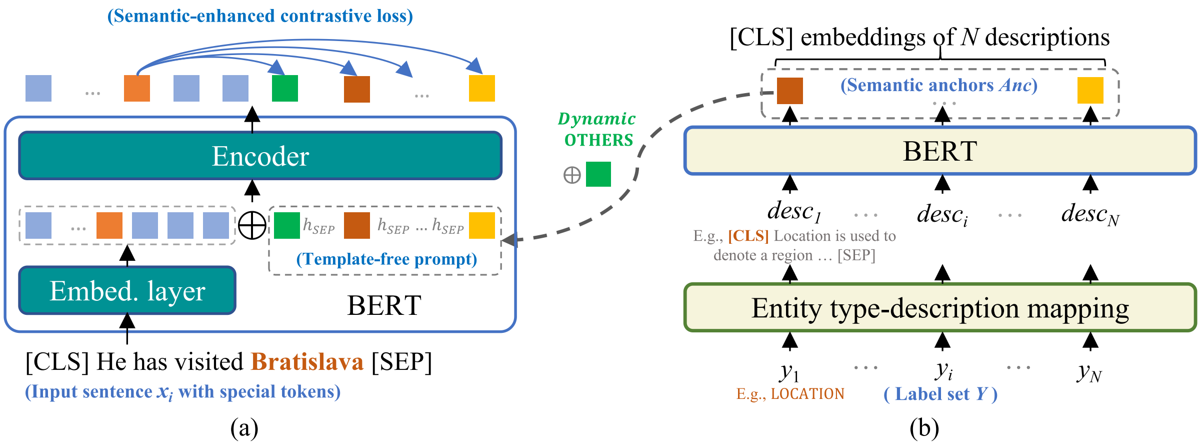

(Semantic-enhanced contrastive loss)
...
Encoder
...
h_{SEP} h_{SEP} ... h_{SEP}
(Template-free prompt)
Embed. layer
BERT
[CLS] He has visited Bratislava [SEP]
(Input sentence x_i with special tokens)
(a)
Dynamic OTHERS
[CLS] embeddings of N descriptions
(Semantic anchors Anc)
BERT
$desc_1$ ... $desc_i$ ... $desc_N$
E.g., [CLS] Location is used to denote a region … [SEP]
Entity type-description mapping
y_1 ... y_i ... y_N
E.g., LOCATION
(Label set Y)
(b)

Fig. 2. TFP framework. (a) Contrastive in-context learning with semantic anchor guarding. (b) Initialization of semantic anchors with prior knowledge. BERT in (a) and (b) share the same parameter set. y_i and desc_i are label and its description as (1). Anc is the set of semantic anchors as (2). h_{SEP} is the representation of a special marker [SEP] in BERT. A TFP consists of the representations of Anc, Dynamic OTHERS, and h_{SEP} as (4).

learn better sentence representations from large amounts of unlabeled data with contrastive learning. This method assumed that if two text fragments (span) are from the same document, then their semantic representations should be relatively close to each other, otherwise, they are far away. Furthermore, when two text fragments are both from the same document if they are located closer together in the document, their semantics indicate proximity, otherwise far away. Both the above research are exploring unsupervised contrastive learning, they cannot take advantage of semantic information within labels.

III. Methodology

First, we propose TFP tuning. TFP collects external descriptions for all label classes in a used dataset and encodes these descriptions as semantic anchors [Fig. 2(b)]. These anchors are used to compose TFPs, and concatenated with the embedded original input sentences, feeding into a PLM encoder [Fig. 2(a)]. Second, we introduce semantic-enhanced contrastive learning that achieves effective latent type prototypes and token representations. TFP does not introduce extra parameters for classification, which is an advantage in few-shot tasks.

A. TFP Tuning for NER

An input of TFP consists of two parts. The first part [Fig. 2(a)] consists of tokens from an original input sentence, and special tokens [CLS] and [SEP] at the beginning and the end of the original sentence ($X = [x_1, x_2, \ldots, x_t]$). The second part [Fig. 2(b)] is a label set $Y = [y_1, y_2, \ldots, y_N]$, where N is the number of predefined entity types for predictions in the current episodes. TFP obtains the representations (H) of X from the embedding layer of an employed PLM, i.e., BERT-base-uncased[3] [61], and the initialized semantic anchors $\text{Anc} = \{\text{anc}_1, \text{anc}_2, \ldots, \text{anc}_N\}$ (the representations of Y with prior knowledge) in vector space. For obtaining Anc, we collect the description set Desc of Y, where each entity type $y_i \in Y$ can find a definition sentence ($\text{desc}_i \in \text{Desc}$) given by the first sentence of the related Wikipedia page. We define such a process as a mapping function

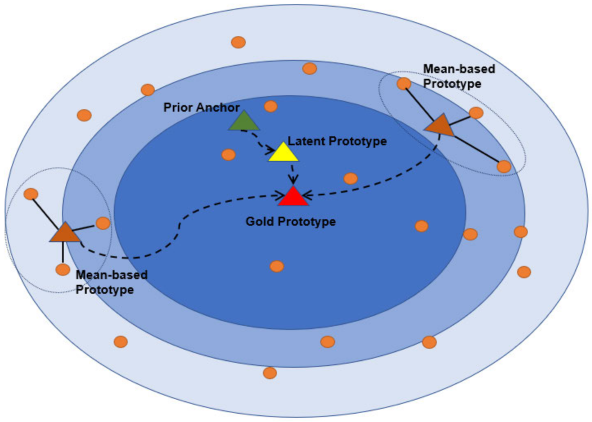


Fig. 3. Comparison of prototype representations between our prior anchor-based method and a traditional mean-based method. The orange circles denote data distribution.

$$\text{desc}_i = \mathcal{M}(y_i). \tag{1}$$

For example, the description of an entity type LOCATION (desc_{loc}) is “*location or place are used to denote a region (point, line, and area).*” This description contains the definition of LOCATION and important entity features, such as “region” and “place.” TFP encodes this description to obtain a semantic anchor (Anc_{loc}) with prior knowledge as the initialization of the prototype of LOCATION. For each u batches (u is a hyper-parameter), TFP takes $\{\text{desc}\}_{i=1}^{N}$ as inputs (namely N description sentences as an extra batch) to obtain updated Anc for the construction of prompts

$$\text{Anc} = \text{BERT}\big(\{\text{desc}\}_{i=1}^{N}\big). \tag{2}$$

Different from using mean-based representations of randomly sampled data as prototypes [62], [63], Anc are embedded by external prior knowledge. Hence, a prior anchor (the green triangle in Fig. 3) is more stable than sampling from sparse data in different training episodes (the orange triangles in Fig. 3), because different sampled data can yield very different prototypes in few-shot learning.

Besides N semantic anchors of N target labels, TFP also needs an extra semantic anchor for the entity type

[3] BERT in following equations shares the same parameter set.

of OTHER. Previous works normally defined OTHER with a unique representation [22], [46], [64]. However, we believe that OTHER should have different representations, because it is the label for the tokens that do not belong to any target types. For example, for a three-way sampled data {PERSON, LOCATION, ORGANIZATION}, a more reasonable OTHER type representation should represent "nonperson, nonlocation, and nonorganization" types. For another episode with different labels, OTHER should have a different representation. To this end, TFP takes advantage of dynamic OTHER representations in different episodes. For Anc (Anc $\in \mathbb{R}^{N\times 768}$), TFP randomly initializes a matrix Tmp with the same size of Anc and applies orthogonal triangle decomposition to obtain a dynamic OTHER representation by

$$O_{\text{dyn}} = \text{Anc} - \left\langle \text{Anc}, \frac{\text{Tmp}}{\|\text{Tmp}\|_F} \right\rangle \frac{\text{Tmp}}{\|\text{Tmp}\|_F} \tag{3}$$

where $\langle\rangle$ denotes dot product and F denotes F-norm. The intuition of using orthogonal triangle decomposition is to obtain an embedding that is distant from existing N anchors in the current episode. Then, our template-free prompt is given by

$$\text{prompt} = \left[O_{\text{dyn}}, h_{\text{SEP}}, \text{anc}_1, h_{\text{SEP}}, \ldots, \text{anc}_N\right] \tag{4}$$

where h_{SEP} is the representation of a special marker [SEP] in BERT, which is used to separate different components. These special markers are used to provide information to the employed PLM about which part is an input sentence and which parts are elements in a prompt.

We concatenate ($\oplus$) prompt and the token representations $\{h_i\}_{i=1}^{t}$ of sequence X, where $\{h_i\}$ is obtained from the BERT embedding layer (BERT$_{\text{emb}}$). With such a concatenation, we do not have to design any natural language-based prompt templates, e.g., "[Span] is a [Type] entity," or label word mappings, e.g., "map(place, area) = LOCATION." Next, the input instance (inst) of TFP is given by

$$\text{inst} = \{h_i\}_{i=1}^{t} \oplus \left\{h_j\right\}_{j=t+1}^{l} \tag{5}$$

where $h_j \in$ prompt, l is the length of inst. inst is fed into BERT encoder (enc) to obtain in-context representations by

$$\left[h'_1, h'_2, \ldots, h'_t, h'_{t+1}, \ldots, h'_l\right] = \text{BERT}_{\text{enc}}(\text{inst}) \tag{6}$$

where BERT$_{\text{enc}}$ means using the encoder of BERT without embedding layer.

TFP compares token representations $\{h'_i\}_{i=1}^{t}$ with prototypes $\{h'_j\}_{j=t+1}^{l}$ to predict probabilities as (7), from the normalized cosine-similarity that is denoted as $d(\cdot)$. We only compute the elements of Anc and O_{dyn} in prompt (the index set is denoted as J), where h_{SEP} in (4) are masked

$$P\left(\hat{y}_i\right) = \frac{\exp\left(-d\left(h'_i, h'_j\right)\right)}{\sum_{j\in J} \exp\left(-d\left(h'_i, h'_j\right)\right)}. \tag{7}$$

The final predicated label for a token is given by

$$\hat{y} = \arg\max_{i\in\{1,\ldots,t\}} P\left(\hat{y}_i\right). \tag{8}$$

B. Semantic-Enhanced Contrastive Learning

We propose a hybrid granularity contrastive loss guided by semantic information. TFP takes advantage of using stable semantic anchors to optimize distances between prototypes with tokens, as well as prototypes with other prototypes. Our semantic anchors utilize external descriptions to parameterize prompts, so they are more stable than prototypes averaged from random samples in typical prototype networks [62].

A typical contrastive loss InfoNCE [65] as

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_i \log \frac{\exp\left(v_i \cdot v'_i/\tau\right)}{\sum_j \exp\left(v_i \cdot v_j/\tau\right)} \tag{9}$$

where v_i is the embedding of an input instance; v'_i is a related positive embedding; v_j is a positive embedding plus negative embeddings from other instances; and τ is a temperature hyper-parameter. The idea of InfoNCE is to pull an instance's embedding close to its augmentations and far away from other input instances. This loss optimizes representations with many negative samples, rather than directly predicting labels.

In this article, we modify InfoNCE for few-shot learning with semantic guiding. We assume that there are $N+1$ latent variables Proto $=$ $\{o_j\}_{j\in J}$ for all entity types, including OTHER. First, the input of TFP inst contains elements $\{h_i\}_1^t$ and $\{h_j\}_{t+1}^{t'}$, where $\{h_i\}_1^t$ are the representations of input tokens X and $\{h_j\}_{t+1}^{t'}$ are the representations of prompts. Our objective is to optimize the network parameters θ that maximize the log-likelihood function of an inst as follows:

$$\theta^* = \text{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} \log p\left(\left[h_i; h_j\right]; \theta\right). \tag{10}$$

By assuming the input representations $[h_i; h_j]$ are related to $N+1$ latent variable Proto $= [o_1, o_2, \ldots, o_{N+1}]$ for each entity type, (10) can be rewrite as follows:

$$\theta^* = \text{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} \log p\left(\left[h_i; h_j\right], o_j; \theta\right). \tag{11}$$

We introduce the latent distribution $T(o_j)$ ($\sum_x T = 1$) over each prototype o_j as follows:

$$\begin{aligned}\theta^* &= \text{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} \log T\left(o_j\right) \frac{p\left(\left[h_i; h_j\right], o_j; \theta\right)}{T\left(o_j\right)} \\ &\geq \text{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} T\left(o_j\right) \log \frac{p\left(\left[h_i; h_j\right], o_j; \theta\right)}{T\left(o_j\right)} \\ &= \text{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} \log T\left(o_j\right) * p\left(\left[h_i; h_j\right], o_j; \theta\right) \\ &\quad - T\left(o_j\right) * \log T\left(o_j\right)\end{aligned} \tag{12}$$

where

$$\begin{aligned}T(o_j) &= \frac{p(\left[h_i; h_j\right], o_j; \theta)}{\sum_{j=t+1}^{t'} p(\left[h_i; h_j\right], o_j; \theta)} \\ &= \frac{p(\left[h_i; h_j\right], o_j; \theta)}{p(\left[h_i; h_j\right]; \theta)} \\ &= p(o_j; \left[h_i; h_j\right], \theta).\end{aligned} \tag{13}$$

By ignoring the constant

$$\sum_{i=1}^{t} \sum_{j=t+1}^{t'} -T\left(o_j\right) * \log T\left(o_j\right) \tag{14}$$

our object is equal to

$$\begin{aligned} \theta^* &= \operatorname{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} T\left(o_j\right) \log p\left(\left[h_i; h_j\right], o_j; \theta\right) \\ &= \operatorname{argmax}_\theta \sum_{i=1}^{t} \sum_{j=t+1}^{t'} p\left(o_j; \left[h_i; h_j\right]\theta\right) \\ &\quad * \log p\left(\left[h_i; h_j\right], o_j; \theta\right) \end{aligned} \tag{15}$$

where

$$\begin{aligned} T\left(o_j\right) &= \frac{p\left(\left[h_i; h_j\right], o_j; \theta\right)}{\sum_{j=t+1}^{t'} p\left(\left[h_i; h_j\right], o_j; \theta\right)} \\ &= \frac{p\left(\left[h_i; h_j\right], o_j; \theta\right)}{p\left(\left[h_i; h_j\right]; \theta\right)} \\ &= p\left(o_j; \left[h_i; h_j\right], \theta\right). \end{aligned} \tag{16}$$

With the assumption that there is a uniform prior over cluster centers, and the prior probability $p(c_i; \theta)$ for each o_j is $1/r$

$$\begin{aligned} p\left(\left[h_i; h_j\right], o_j; \theta\right) &= p\left(\left[h_i; h_j\right]; o_j, \theta\right) p\left(o_j, \theta\right) \\ &= 1/r * p\left(\left[h_i; h_j\right]; o_j, \theta\right). \end{aligned} \tag{17}$$

Furthermore, by assuming that the distribution around each cluster center (prototype o_j) is an isotropic Gaussian, we have

$$\begin{aligned} &p\left(\left[h_i; h_j\right]; o_j, \theta\right) \\ &\quad = \exp\left(\frac{-\left(h_i - o'_j\right)^2}{2\sigma^2}\right) \Bigg/ \sum_{j=1}^{t'} \left(\frac{-\left(h_i - o_j\right)^2}{2\sigma^2}\right) \end{aligned} \tag{18}$$

where $j' \neq j$. Then, we compute $p(o_j; [h_i; h_j], \theta)$ in (15), where $p(o_j; [h_i; h_j], \theta) = 1$ if h_i is related to o_j, otherwise $p(o_j; [h_i; h_j], \theta) = 0$. In such condition, combining (15) with (17) and (18), and calculating distances between h_i and o_j with function $d(\cdot)$, the log-likelihood function (10) can be rewritten as

$$\theta^* = \operatorname{argmin}_\theta \sum_{i=1}^{t} \log \frac{\exp\left(-d\left(h_i, o'_j\right)^2 \Big/ \tau\right)}{\sum_{j=1}^{t'} \left(-d\left(h_i, o_j\right)^2 / \tau\right)} \tag{19}$$

where τ is a constant. Namely, TPF updates parameters by minimizing the loss function with a form of InfoNCE

$$\mathcal{L}_{t2o} = -\sum_{i=1}^{t} \log \frac{\exp\left(-d\left(h_i, o'_j\right)^2 \Big/ \tau\right)}{\sum_{j=1}^{t'} \left(-d\left(h_i, o_j\right)^2 \Big/ \tau\right)}. \tag{20}$$

TFP employs in-context encoded representations $\{h'_j\}_{j=t+1}^{l}$ in (6) as the estimations for o_j. The core difference between InfoNCE loss and our loss in (20) is that TFP constructs a positive pair as (token, related prototype) and a negative pair as (token, unrelated prototype). These prototypes are semantic-enhanced label representations. Thus, (20) can optimize token-wise representations and also distinguish different classes in few-shot settings. It pulls token embeddings closer to their related prototypes and pushes them away from unrelated ones. TFP optimizes prototypes by training after the prototypes are initialized with external prior knowledge.

TFP desires prototypes can keep certain distances from each other (this will be verified in Fig. 4 later). To this end, we propose an auxiliary component, given by

$$\mathcal{L}_{o2o} = \frac{N^2/\tau_2}{\sum d\left(\left\{o_j\right\}_1^{N+1}\left(\left\{o_{j'}\right\}_1^{N+1}\right)^T\right)} \tag{21}$$

where τ_2 is a temperature hyper-parameter for scaling loss values; $j \neq j'$. Such an auxiliary loss can avoid a representation collision issue that was argued by the work [66]. The overall loss ($\mathcal{L}$) is

$$\mathcal{L} = \mathcal{L}_{o2o} + \mathcal{L}_{t2o}. \tag{22}$$

In summary, typical unsupervised InfoNCE loss is regarded as a class-agnostic auxiliary loss to update token-wised representations. Thus, they have to employ an extra class-specific loss combined with a linear layer to predict labels. Different from the above method, our semantic-enhanced contrastive loss optimizes FTP by clustering the nodes with semantic centers, i.e., latent prototypes. There is no additional parameter introduced in our model, which is an advantage in few-shot tasks.

IV. Task Formulation

NER is defined as a token-level sequence labeling task. Given an input sentence with t tokens, $X = \{x_1, x_2, \ldots, x_t\}$, NER assigns a label $y_i \in Y$ to each token x_i, where Y is a predefined label set. Y usually contains entity types such as ORGANIZATION, PERSON, and LOCATION. If a token does not belong to these classes, it is labeled as OTHER. Models can only learn from limited label-specific data in few-shot NER. Some existing few-shot NER work under various settings [47], [64], [67]. With a comprehensive survey, we conducted our experiments with three different settings, including LE, DA, and NA few-shot NER. These three settings focus on different challenges in few-shot NER, which can systematically evaluate the capacity of proposed TFP in aspects of LE, DA, and low-resource generalization. LE and DA follow typical N-way-K-shot settings[4] while NA is a stricter few-shot setting.

A. Label Extension

LE setting aims to evaluate the LE ability of a model. This evaluation is motivated by the fact that new types of entities often appear in certain domains in real-world applications. There are eight subtasks in this setting, including the combinations of 5-way and 10-way by 1–2 and 5–10 shots in both FEW-NERD INTER and FEW-NERD INTRA datasets [69]. The FEW-NERD dataset is designed with a hierarchical label scheme, which contains 66 fine-grained entity types that

[4] N-way-K-shot details refer to the work [68].

are clustered by eight coarse-grained types. In INTER and INTRA datasets, there is no overlapped fine-grained entity type between the training and validation/test sets. However, INTER can share coarse-grained entity types. If a type, e.g., LOCATION-ISLAND is in the training set, the test sets of INTRA and INTRA do not contain this fine-grained type, whereas the type LOCATION-MOUNTAIN can be in the INTER test set. FTP is fine-tuned by randomly sampling 1–2 shots each time for each type class. After training, FTP is adapted[5] in the support set of the test/validation set and then predict corresponding fine-grained types in the query set of the test/validation set. Accurate results in INTER/INTRA show that the model can recognize new types of entities with/without parts of class information sharing.

B. Domain-Adaption

DA setting evaluates the domain transferability of a model. In this task, training and test data are from different domains. This setting includes six subtasks. There is a common training dataset OntoNotes 5.0 [70] and three test datasets, i.e., CoNLL 03 [71], WNUT 17 [72], and I2B2 [73]. OntoNotes 5.0 data are from a general domain. CoNLL 03, WNUT 17, and I2B2 data are from newswire, social, and medical domains, respectively. TFP is evaluated in 1- and 5-shot subtasks with the later three test sets. First, OntoNotes 5.0 is employed as training data to fine-tune a model. Then, for the test data from CoNLL 03, WNUT 17, and I2B2, the model adapts with their support sets and predicts related instances in query sets. The reported results for CoNLL 03, WNUT 17, and I2B2 are averaged F-1 measures of the query set when models are adapted with the five sampled few-shot support sets. The used five sampled support sets come from the work [67].

C. No-Adapting

NA setting has the same predefined label set for training and testing. However, NA does not contain a source-rich training set to sample episodes for fine-tuning. Thus, NA strictly tests the low-resource generalization ability of a model. For example, when performing a 5-shot task with four entity types, all available training data are 4 × 5 instances in this setting. After training, a model is directly evaluated by test data without adaptation steps. TFP employs the training data from work [47] in this setting, which samples three limited support sets from the whole data of CONLL 03, MIT-Movie [74], and OntoNotes 5.0. The final results are reported on the original test sets of these three datasets. This setting focuses on evaluating models' few-shot ability in the strictest way.

V. Experiments

A. Datasets

We report the results of 26 subtasks within six employed datasets under three different few-shot settings (LE, DA, and NA) for evaluating the few-shot learning ability of TFP.

TABLE I
Statistics of Data Used by LE Setting. # Denotes Counting Numbers

LE	FEW-NERD INTER			FEW-NERD INTRA		
	train	dev	test	train	dev	test
# Class	36	13	17	35	14	17
# Sent	130,111	18,816	14,006	99,518	19,357	44,058
# Token	3,455,927	425,998	312,762	2,677,915	503,784	1,012,940
# Entity token	582,280	67,002	62,114	404,209	82,462	212,898
Average length	26	22	22	26	26	22

The statistics of used data are shown in Tables I–III, which shows the challenges of few-shot NER with different setups. The related settings are summarized in Table IV.

B. Compared Baselines

A total of nine recent baselines are compared with the proposed TFP under the settings of LE, DA, and NA. All these baselines and TFP take BERT-base-uncased as the employed PLM.

1) **CONTaiNER** [64] uses NER contrastive learning to optimize Gaussian-distributed token-wise distances.
2) **DML** [75] proposes a model-agnostic meta-learning method to initialize parameters for fast adaptations.
3) **COPNER** [46] proposes a prompt-based method that uses class-specific words as metric referents and supervision signals to achieve few-shot NER.
4) **ESD** [76] studies sequence labeling tasks as a span-level pipeline, including enhanced span representations, prototype aggregations, and span conflict resolutions.
5) **NNShot and StructShot** [67] use a nearest neighbor classifier to differentiate each token. StructShot adds a Viterbi decoding algorithm upon NNShot.
6) **ProtoBERT** [69] combines a classical prototypical network [62] with a BERT encoder to classify entity types.
7) **Tagger** [47] is a simple but strong baseline. The method uses a linear classifier on top of BERT, following a full supervision setting with cross-entropy.
8) **TemNER** [22] is a prompt-based method that treats few-shot NER as an LM ranking task for a full use of knowledge transfer in model parameters.
9) **EntLM** [47] defines NER as an entity-oriented LM task to address N-gram traversal. This method is seq2seq-based; it generates entities in special positions and maps them to manually defined label words.

C. Result

TFP performance on LE, DA, and NA tasks is shown in Tables V–VII, respectively. In Table V, TFP achieves averaged SOTA results, compared with strong baselines. A definite trend is that TFP performs better with fewer data. Given 1–2 shots of 5- and 10-way, TFP yields gains of 2.06%, 1.44%, 3.45%, and 2.79% on FEW-NERD INTER and INTRA, compared to the strongest baseline (DML). It shows the LE ability of TFP under a few-shot setting.

Table VI shows that TFP achieves SOTA results in all subtasks. On average, TFP exceeds COPNER by 2.15%, 4.38%,

[5] Adaptation is defined as training with support sets of a test set [46]. This process is taken under a high-source scenario. The adapted support sets have the same label space as its test set but do not overlap with train data.

TABLE II
STATISTICS OF DATA USED BY DA SETTING

DA	OntoNotes	CoNLL			I2B2			WNUT		
	Train	1 shot support set	5shot support set	Test	1 shot support set	5shot support set	Test	1 shot support set	5shot support set	Test
# Class	18	4	4	4	18	18	18	6	6	6
# Sent	59,924	2.6	8.6	683	13	58	7,527	5	25	1,287
# Token	1,088,503	48.2	192.2	46,665	119	479	120,982	53	382	23,394
# Entity token	149,374	9.6	36.2	8,112	43	188	14,652	15	54	1,740
Average length	18	21.75	22.71	12.67	9.15	8.26	16.07	10.60	15.28	18.18

TABLE III
STATISTICS OF DATA USED BY NA SETTING

NA	CoNLL					MIT-Movie					OntoNote				
	5shot	10shot	20Shot	50shot	Test	5shot	10shot	20Shot	50Shot	Test	5shot	10shot	20Shot	50Shot	Test
# Class	4	4	4	4	4	12	12	12	12	12	18	18	18	18	18
# Sent	8	18	34	79	3,683	34	72	138	311	2,443	66	111	240	584	8262
# Token	248	430	837	1968	46,665	434	872	1649	3,833	24,686	1536	2760	5,689	13,629	152,728
# Entity token	23	60	120	306	8,112	135	289	574	1,330	9,757	191	364	781	2,078	20913
Average length	31.00	23.80	24.61	24.91	12.67	12.76	12.11	11.95	12.32	10.10	23.27	24.86	23.70	23.33	18.49

TABLE IV
ILLUSTRATION OF DATASETS AND TASK SETTINGS

Setting	Corpus	Domain	N-way-K-shot	High-source	Fine-tuning data	Valid data	Test data
LE	FEW-NERDINTER FEW-NERDINTRA	General	5-1,5-5, 10-1, 10-5	Yes	1 common training set	1 support set 1 query set	1 support set 1 query set

DA	OntoNotes (train) CoNLL(test)
	WNUT(test)
	I2B2(test)
		News
		Social
		Medical
			4-1,4-5
			6-1,6-5
			18-1,18-5
				Yes	1 common training set	No	5 support sets
							1 query set

NA	CoNLL MIT-Movie OntoNotes	News Review General	4-5,4-10, 4-20, 4-50 12-5,12-10, 12-20, 12-50 18-5,18-10, 18-20, 18-50	No	3 different train sets	No	1 test set

TABLE V
F1 SCORES (%) IN THE LE SETTING

Model	INTER					INTRA				
	5-way		10-way		Avg.	5-way		10-way		Avg.
	1-2 shot	5-10 shot	1-2 shot	5-10 shot		1-2 shot	5-10 shot	1-2 shot	5-10 shot	
ProtoBERT†	$38.83_{1.49}$	$58.79_{0.44}$	$32.45_{0.79}$	$52.92_{0.37}$	45.75	$20.76_{0.84}$	$42.54_{0.94}$	$15.05_{0.44}$	$35.40_{0.13}$	28.44
ProtoBERT‡	44.44/	58.80/	39.09/	53.97/	49.08	23.45/	41.93/	19.76/	34.61/	29.94
NNShot†	$47.24_{1.00}$	$55.64_{0.63}$	$38.87_{0.21}$	$49.57_{2.73}$	47.83	$25.78_{0.91}$	$36.18_{0.79}$	$18.27_{0.41}$	$27.38_{0.53}$	26.90
NNShot‡	54.29/	50.56/	46.98/	50.00/	50.46	31.01/	35.74/	21.88/	27.67/	29.08
StructShot†	$51.88_{0.69}$	$57.32_{0.63}$	$43.34_{0.10}$	$49.57_{3.08}$	50.53	$30.21_{0.90}$	$38.00_{1.29}$	$21.03_{1.13}$	$26.42_{0.60}$	28.92
StructShot‡	57.33/	57.16/	49.46/	49.39/	53.34	35.92/	38.83/	25.38/	26.39/	31.63
CONTaiNER$^{\mp}$	$59.20_{1.34}$	$64.23_{0.65}$	$50.22_{1.64}$	$58.97_{1.42}$	58.16	$44.11_{1.01}$	$57.68_{0.81}$	$34.85_{1.20}$	$50.89_{0.42}$	46.88
CONTaiNER‡	56.10/	61.90/	48.36/	57.13/	55.87	40.40/	53.71/	33.82/	47.51/	43.86
COPNER$^{\mp}$	$66.13_{1.12}$	$67.33_{1.32}$	$59.76_{0.72}$	$63.53_{0.69}$	64.18	$53.12_{1.48}$	$57.99_{1.05}$	$45.88_{1.10}$	$51.94_{1.03}$	52.23
COPNER	65.98/	67.70/	59.56/	62.37/	63.90	54.26/	58.84/	44.26/	51.18/	52.14
ESD	$66.46_{0.49}$	**74.14**$_{0.80}$	$59.95_{0.69}$	$67.91_{1.41}$	67.12	$41.44_{1.16}$	$50.68_{0.94}$	$32.29_{1.10}$	$42.92_{0.75}$	41.83
DML	$68.77_{0.24}$	$71.62_{0.16}$	$63.26_{0.40}$	**68.32**$_{0.10}$	67.99	$52.04_{0.44}$	$63.23_{0.45}$	$43.50_{0.59}$	**56.84**$_{0.14}$	53.90
Ours	**70.83**$_{0.62}$	$72.14_{0.40}$	**64.70**$_{0.72}$	$67.65_{0.15}$	**68.83**	**55.49**$_{0.67}$	**63.31**$_{0.77}$	**46.29**$_{0.74}$	$54.01_{0.60}$	**54.78**

[*] The original baseline results† with standard deviations are cited from the work [69] and the updated baseline results‡ without standard deviations are cited from the work [64]. Considering that standard deviation is an important measure for few-shot tasks, we replicate the results$^{\mp}$ for a fair comparison. Noticeably, original CONTaiNER‡ uses incorrect data samples. Our replication of CONTaiNER$^{\mp}$ uses the revised samples published by the authors of CONTaiNER‡ later. We report our five-times averaged results, using the official data splits from the work [69]. The best results are in **bold**.

and 10.17% F1. Compared with CoNLL sourced from news, WNUT and I2B2 are more challenging. WNUT aims to extract entities from noisy text where sentences are ungrammatical. I2B2 contains many numerical entity types, which are hard

TABLE VI
F1 Scores (%) in the DA Setting

Model	CoNLL			WNUT			I2B2			Avg.
	1 shot	5 shot	Avg.	1 shot	5 shot	Avg.	1 shot	5 shot	Avg.	
ProtoBERT†	$53.00_{7.2}$	$65.90_{1.6}$	59.45	$14.80_{4.9}$	$19.80_{5.0}$	17.30	$7.60_{3.5}$	$10.30_{0.4}$	8.95	28.57
ProtoBERT+†	$56.00_{7.3}$	$67.10_{1.6}$	61.55	$18.80_{5.3}$	$23.80_{3.9}$	21.30	$7.90_{3.2}$	$10.10_{0.9}$	9.00	30.62
NNShot†	$61.30_{11.5}$	$74.30_{2.4}$	67.80	$21.70_{6.3}$	$23.90_{5.0}$	22.80	$16.60_{2.1}$	$23.70_{1.3}$	20.15	36.92
StructShot†	$62.30_{11.4}$	$75.20_{2.3}$	68.75	$25.30_{5.3}$	$27.20_{6.7}$	26.25	$22.10_{3.0}$	$31.80_{1.8}$	26.95	40.65
CONTaiNER‡	$61.20_{10.7}$	$75.80_{2.7}$	68.50	$27.50_{1.9}$	$32.50_{3.8}$	30.00	$21.50_{1.7}$	$36.70_{2.1}$	29.10	42.53
COPNER	$66.50_{2.1}$	$74.60_{3.1}$	70.55	$34.90_{1.8}$	$34.20_{2.6}$	34.55	$35.80_{1.3}$	$43.70_{1.5}$	39.75	48.28
Ours	**67.43**$_{2.2}$	**77.97**$_{1.4}$	**72.70**	**38.86**$_{0.80}$	**38.99**$_{2.6}$	**38.93**	**47.17**$_{3.5}$	**52.66**$_{1.3}$	**49.92**	**53.85**

* We report averaged F-1 with standard deviations on five different support sets, and run each support set three times. The results with † and ‡ are from the work [64], [67].

TABLE VII
F1 Scores (%) in the NA Setting

	Model	5 shot	10 shot	20 shot	50 shot	Avg.
CoNLL†	Tagger	$41.87_{12.1}$	$59.91_{10.7}$	$68.66_{5.1}$	$73.20_{3.1}$	60.91
	NNShot	$42.31_{8.9}$	$59.24_{11.7}$	$66.89_{6.1}$	$72.63_{3.4}$	60.27
	StructShot	$45.82_{10.3}$	$62.37_{11.0}$	$69.51_{6.5}$	$74.73_{3.1}$	63.11
	TemNER	$43.04_{6.2}$	$57.86_{5.7}$	$66.38_{6.1}$	$72.71_{2.1}$	60.00
	EntLM	$51.32_{7.7}$	$66.86_{3.0}$	$71.23_{3.9}$	$74.80_{1.9}$	66.05
	COPNER	$54.20_{7.9}$	$66.20_{2.9}$	**71.80**$_{1.8}$	$77.00_{1.4}$	67.30
	Ours	**61.93**$_{2.1}$	**69.46**$_{2.0}$	$71.76_{1.3}$	**77.66**$_{1.9}$	**70.20**
MIT-Movie†	Tagger	$39.57_{6.4}$	$50.60_{7.3}$	$59.34_{3.7}$	$71.33_{3.0}$	55.21
	NNShot	$38.97_{5.5}$	$50.47_{6.1}$	$58.94_{3.5}$	$71.17_{2.9}$	54.89
	StructShot	$41.60_{9.0}$	$53.19_{5.5}$	$61.42_{3.0}$	$72.07_{6.4}$	57.07
	TemNER	$45.97_{3.9}$	$49.30_{3.4}$	$59.09_{0.4}$	$65.13_{0.2}$	54.87
	EntLM	$49.15_{8.9}$	$59.21_{4.0}$	$63.85_{3.7}$	$72.99_{1.8}$	61.30
	COPNER	$50.10_{3.6}$	$61.90_{1.4}$	$68.90_{2.4}$	$74.60_{0.3}$	63.88
	Ours	**59.25**$_{4.4}$	**65.82**$_{1.0}$	**70.87**$_{1.7}$	**75.42**$_{0.4}$	**67.84**
OntoNotes‡	Tagger	$21.01_{1.7}$	$31.71_{1.6}$	$36.23_{1.4}$	$46.18_{1.2}$	33.78
	NNShot	$38.62_{3.3}$	$42.91_{4.0}$	$48.77_{1.0}$	$50.95_{0.5}$	45.31
	StructShot	$38.91_{3.4}$	$43.02_{5.1}$	$49.00_{2.6}$	$51.28_{1.2}$	45.55
	TemNER	$39.06_{3.1}$	$50.82_{1.9}$	$59.28_{1.0}$	$67.94_{0.8}$	54.28
	EntLM	$36.41_{3.5}$	**53.20**$_{1.8}$	**61.22**$_{2.3}$	**68.92**$_{1.6}$	54.94
	COPNER	$38.72_{6.4}$	$50.61_{7.3}$	$59.35_{3.7}$	$64.21_{3.0}$	54.05
	Ours	**40.96**$_{2.5}$	$51.14_{2.2}$	$59.50_{2.5}$	$68.59_{3.1}$	**55.05**

* The results with † are from three different support sets sampled by the work [47]. Each support set repeats three times. The results with ‡ are reported from our sampled three support sets, because the work [47] exclude seven entity types from the original OntoNotes. To keep the same OntoNotes with our DA setting, we include these types in the NA setting.

to distinguish, e.g., a Medical Record entity "471-90-84-7" and an ID Number entity "GL735LM." Meanwhile, training sentences from OntoNotes are sourced from a general domain with formal formats. TFP yields large gains in such a context, showing its strong domain transferability.

In Table VII, TFP shows its few-shot generalization ability under NA setting. TFP achieves 61.93%, 59.25%, and 40.96% F1 on CoNLL, MIT-Movie, and OntoNotes, respectively, by only training with five annotated sentences in each class. For 5-shot of CoNLL and MIT-Movie, TFP outperforms the strongest baselines by 7.73% and 9.15% average F1.

TABLE VIII
Ablation Study Measured by F1 (%) in Few-Nerd Inter 5-Way-1-Shot (LE), CoNLL 1-Shot (DA), and CoNLL 5-Shot (NA)

Prompt Form	LE	DA	NA	Avg.
EP (w/o. in-context)	$48.82_{4.2}$	$49.90_{3.6}$	$56.65_{2.8}$	51.79
WP (w/. in-context)	$66.13_{1.1}$	$66.50_{2.1}$	$54.20_{7.9}$	62.28
SP (w/. in-context)	$67.49_{2.3}$	$66.38_{3.1}$	$56.21_{3.8}$	63.36
SP^1 (w/o. in-context)	$46.12_{3.1}$	$47.01_{4.2}$	$55.01_{3.1}$	49.37
CP (w/. in-context)	$68.02_{2.8}$	$63.42_{9.0}$	$58.26_{3.5}$	63.23
FTP^1 (w/o. in-context)	$52.96_{2.4}$	$38.83_{8.7}$	$43.83_{3.0}$	45.21
FTP^2 (w/o. shuffle)	$69.54_{0.9}$	$68.26_{4.2}$	$50.24_{1.1}$	62.68
FTP^3 (w/o. [SEP])	$67.16_{1.4}$	$65.01_{3.0}$	$60.01_{3.3}$	64.06
FTP^4 (w/o. dynamic O)	$68.28_{1.6}$	$66.01_{2.9}$	$59.67_{6.3}$	64.65
FTP	**70.83**$_{0.6}$	**67.43**$_{2.2}$	**61.93**$_{1.2}$	**65.84**

* w/. and w/o. denote with and without.

VI. Analysis

A. Prompt Ablation Analysis

We compare prompts in different forms and perform ablation analysis, which finds that in-context learning plays a core function, rather than various construction formats of prompts that were studied by recent research [22], [23], [24], [47]. Besides, semantic-type representations also work.

Table VIII shows the results of ten compared methods. external prompt (EP) uses fixed label names with manual label word mappings as prompts, which are separately inputted into a model with original sentences, without in-context encoding. Namely, we first input the first part $\{h_i\}_{i=1}^{t}$ of input *inst* in (5) to BERT, aiming to get the representations of each token $\{h'_1, h'_2, \ldots, h'_t\}$ in (6). Then, the label set $Y = [y_1, y_2, \ldots y_N]$ as a prompt is separately fed into BERT to obtain the representation of each class to replace the part $\{h_j\}_{j=t+1}^{l}$ in (6). By such a method, we exclude the effects from in-context encoding a sentence with a prompt. The following "without in-context encoding" means the same method to exclude the effects from in-context encoding.

Words prompt (WP) is from the work [46], which uses the same prompts with EP but with in-context encoding. Namely, EP uses a label name to replace a description sentence (replace (1) into $\text{desc}_i = y_i$).

TABLE IX

SEMANTIC-ENHANCED CONTRASTIVE LOSS ANALYSIS

	LE	DA	NA	Avg.
Random semantics	$26.29_{3.3}$	$30.80_{9.6}$	$49.19_{6.4}$	35.43
Token-wise contrastive	$31.11_{1.1}$	$42.63_{2.4}$	$35.55_{5.1}$	36.43
Mean-based prototype	$56.2_{2.3}$	$58.2_{3.0}$	$57.90_{4.5}$	57.43
FTP	$\mathbf{70.83}_{0.6}$	$\mathbf{67.43}_{2.2}$	$\mathbf{61.93}_{1.2}$	**65.84**

* The used data and measure keep the same with Table VIII.

Synonyms prompt (SP) utilizes averaged embeddings of three synonymous label names from PLM as prompts. SP[1] refers to SP without in-context encoding.

Continual prompts (CPs) use randomly initialized embeddings plus a special prompt encoder for further encoding, which follows the work [77].

Our FTP uses prior semantic anchors for initialization and performs in-context encoding with input sentences. FTP[1] denotes that we separately input the prompts and original sentences into a model, without in-context encoding. FTP[2] uses prompts in which all elements are not shuffled. FTP[3] denotes that no specific marker [SEP] is used to separate input anchors in prompts [see (4)]. FTP[4] uses the fixed representation of OTHER instead of dynamic OTHER described in (3).

By comparing EP with WP, we find that in-context-learning can significantly improve the results by 10.49%. Similar results are also observed when comparing SP[1] with SP and FTP[1] with FTP, where in-context learning achieves 13.99% and 20.63% averaged F1 gains. By comparing FTP with WP, SP, and CP, it is apparent that using our proposed semantic anchors to construct prompts is better than using label names with manual mappings, label name synonyms, and continual embeddings, as F1 improvements by 3.56%, 2.48%, and 2.61% showing in Table VIII. The F1 gains of FTP over FTP[2,3,4] show the effectiveness of prompt element shuffle, using [SEP] markers, and introducing dynamic OTHER learning.

B. Semantic-Enhanced Contrastive Loss Analysis

In Table IX, we analyze the utility of our semantic-enhanced contrastive learning. Random semantics means we replace our semantic anchors $\{\text{anc}_i\}_{i=1}^{N}$ in (4) with random vectors. Token-wise contrastive means we adopt typical contrastive learning without semantic-enhancement. This method uses InfoNCE for representation optimization and a linear layer combined with cross-entropy loss for predictions. Mean-based prototype means we randomly sample some embeddings from label-specific tokens and take mean representations instead of the semantic anchors. In Table IX, TFP surpasses random semantics and is token-wise contrastive with large margins (30.41% and 29.41% in F1). Most existing contrastive learning is token-wise [27], [57], which will wrongly push away the presentations of negative instances that share similar semantics. Notably, this is particularly important in NLP tasks, where it is necessary to maintain consistent and proper semantic information for input tokens, even when they are negative pairs. Besides, the improvements of TFP over the mean-based prototype indicate that our method alleviates the bias of random sampling in few-shot NER.

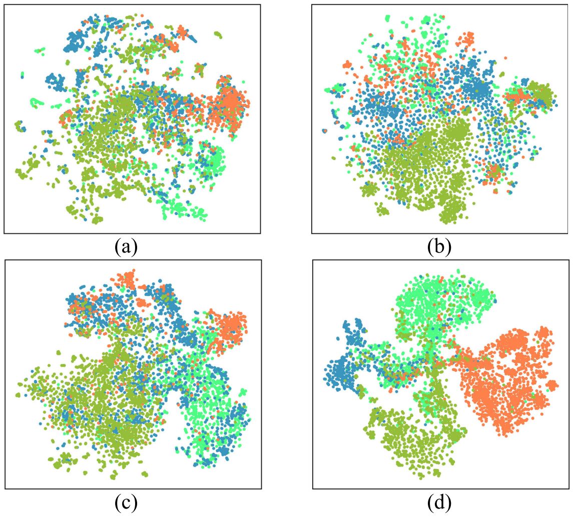

(a) (b) (c) (d)

Fig. 4. t-SNE visualization for the test set of CoNLL in the NA setting. Four colors represent four classes in CoNLL. (a) Tagger. (b) StructShot. (c) COPNER. (d) TFP.

TABLE X

AVERAGED SEMANTIC SIMILARITY OF POSITIVE AND NEGATIVE PAIRS ON DA-BASED CoNLL 1-/5-SHOT

Semantic similarity	Pos	Neg	$SimNeg$
Initial similarity	13.28	12.88	10.58
Random semantics	76.47/81.57	77.17/78.99	77.67/78.66
Token-wise contrastive	37.63/37.25	31.93/31.35	18.72/18.24
Mean-based prototype	42.66/42.64	27.77/27.12	19.37/18.30
FTP	84.15/84.46	66.39/65.03	66.28/64.52

* A higher value denotes more similar. Pos and Neg means the average distance of instances to all positive and negative pairs, respectively. $MinNeg$ denotes the minimum Neg distance. Pos means the average distance of all positive pairs and Neg for negative pairs. Min neg stands for the minimum distance of negative pairs with most similar semantic. All distances are calculated with the same scaling parameter.

Fig. 4 shows the effects of our semantic-enhanced contrastive loss in the NA-based CoNLL test set. Compared with external baselines, TFP can generate the most distinguishable representations optimized by our loss. The distribution of token embeddings $[h'_{1:t}$ in (6)] shows four separated clusters via t-SNE. The nodes from four classes are pulled to four directions by TFP. This finding is statistically supported by Table X, which shows the averaged semantic (cosine) similarity between instances and positive samples (Pos), negative samples (Neg), and semantically similar negative samples (Sim Neg). The semantically similar negative samples are given by original BERT hidden states and cosine similarity. We use $h'_{1:t}$ in (6) to compute cosine similarity. The values are based on a DA-based CoNLL test set (1- and 5-shot). In Table X, initial similarity shows that the representations of instances and the representations of positive and negative samples are not well distinguished, because Pos, Neg, and Sim Neg are small and close. After training, the representations of the random semantics-based method are still indistinguishable in vector space, because the values are near. However, FTP shows a large gap between Pos and Neg, which means the positive and negative pairs are well distinguished. More importantly, the semantically similar negative examples are not further pushed away from the instances, because its Sim Neg is similar to its Neg. In contrast, the benchmarking methods, e.g., token-wise contrastive and mean-based prototypes push those semantically similar negative samples further (their Sim Neg is smaller than their Neg). Thus, it proves that our semantic-enhanced contrastive loss can distinguish positive and negative samples by inputting instances in vector space and also can prevent the distance of semantically similar negative samples from being pushed too far.

VII. Conclusion

In this article, we have introduced the TFP framework, which utilizes prompt tuning to improve token-level NER tasks without the need for template construction or label word mapping. Our prompt-based approach is straightforward to implement and achieves significant performance gains without requiring any complex modifications to the neural architecture. By incorporating the proposed hybrid granularity loss, TFP achieves semantic-guided contrastive learning in few-shot tasks. We demonstrate that our proposed semantic guided loss can effectively address the problem of wrongly pushing away the presentations of negative instances that share similar semantics in typical contrastive learning. Through comprehensive evaluations, we show that our model exhibits strong performance in LE, domain adaptation, and low-resource generalization, achieving 19 out of 26 SOTA results on few-shot NER tasks. Moreover, we find that in-context encoding plays a more critical role than elaborately designed prompts, which is the primary reason why prompt tuning works effectively.

References

[1] N. Jinjie, P. Vlad, Y. Tom, Z. Haicang, and C. Erik, "HiTKG: Towards goal-oriented conversations via multi-hierarchy learning," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 1–9.

[2] D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A survey on conversational recommender systems," *ACM Comput. Surveys*, vol. 54, no. 5, pp. 1–36, 2021.

[3] K. He, L. Yao, J. Zhang, Y. Li, and C. Li, "Construction of genealogical knowledge graphs from obituaries: Multitask neural network extraction system," *J. Med. Internet Res.*, vol. 23, no. 8, Aug. 2021, Art. no. e25670.

[4] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Comput. Surveys*, vol. 54, no. 1, pp. 1–39, 2021.

[5] Y. Wang, H. Tong, Z. Zhu, and Y. Li, "Nested named entity recognition: A survey," *ACM Trans. Knowl. Discovery from Data*, vol. 16, no. 6, pp. 1–29, Dec. 2022.

[6] Z. Liu et al., "CrossNER: Evaluating cross-domain named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, pp. 13452–13460, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17587

[7] R. Mao, C. Lin, and F. Guerin, "Word embedding and WordNet based metaphor identification and interpretation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1222–1231.

[8] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Aug. 2021, pp. 4582–4597.

[9] 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.*, Jun. 2021, pp. 2339–2352.

[10] R. Mao, X. Li, M. Ge, and E. Cambria, "MetaPro: A computational metaphor processing model for text pre-processing," *Inf. Fusion*, vols. 86–87, pp. 30–43, Oct. 2022.

[11] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Trans. Affect. Comput.*, early access, Sep. 8, 2023, doi: 10.1109/TAFFC.2022.3204972.

[12] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt tuning with rules for text classification," 2021, *arXiv:2105.11259*.

[13] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 4222–4235. [Online]. Available: https://aclanthology.org/2020.emnlp-main.346

[14] K. He, Y. Huang, R. Mao, T. Gong, C. Li, and E. Cambria, "Virtual prompt pre-training for prototype-based few-shot relation extraction," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118927. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422019455

[15] X. Liu et al., "GPT understands, too," 2021, *arXiv:2103.10385*.

[16] T. Vu, B. Lester, N. Constant, R. Al-Rfou, and D. Cer, "SPoT: Better frozen model adaptation through soft prompt transfer," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5039–5059. [Online]. Available: https://aclanthology.org/2022.acl-long.346

[17] S. Hu et al., "Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2225–2240. [Online]. Available: https://aclanthology.org/2022.acl-long.158

[18] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2021, pp. 5203–5212. [Online]. Available: https://aclanthology.org/2021.naacl-main.410

[19] A. Webson and E. Pavlick, "Do prompt-based models really understand the meaning of their prompts?" in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2022, pp. 2300–2344.

[20] R. Logan IV, I. Balazevic, E. Wallace, F. Petroni, S. Singh, and S. Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2824–2835. [Online]. Available: https://aclanthology.org/2022.findings-acl.222

[21] T. Schick, H. Schmid, and H. Schütze, "Automatically identifying words that can serve as labels for few-shot text classification," in *Proc. 28th Int. Conf. Comput. Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, Dec. 2020, pp. 5569–5578. [Online]. Available: https://aclanthology.org/2020.coling-main.488

[22] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using bart," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 1835–1845.

[23] X. Chen et al., "LightNER: A lightweight generative framework with prompt-guided attention for low-resource NER," 2021, *arxiv:2109.00720*.

[24] A. T. Liu, W. Xiao, H. Zhu, D. Zhang, S.-W. Li, and A. Arnold, "QaNER: Prompting question answering models for few-shot named entity recognition," 2022, *arXiv:2203.01543*.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[26] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[27] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552

[28] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[29] D. Wang, N. Ding, P. Li, and H. Zheng, "CLINE: Contrastive learning with semantic negative examples for natural language understanding," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 2332–2342. [Online]. Available: https://aclanthology.org/2021.acl-long.181

[30] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," in *Computer Vision—ECCV 2020*. Berlin, Germany: Springer-Verlag, 2020, pp. 741–756, doi: 10.1007/978-3-030-58452-8_43.

[31] S. Zhao, M. Hu, Z. Cai, and F. Liu, "Dynamic modeling cross-modal interactions in two-phase prediction for entity-relation extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1122–1131, Mar. 2023, doi: 10.1109/TNNLS.2021.3104971.

[32] X. Dai, S. Karimi, B. Hachey, and C. Paris, "An effective transition-based model for discontinuous NER," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 5860–5870. [Online]. Available: https://aclanthology.org/2020.acl-main.520

[33] T. Liang, W. Wang, and F. Lv, "Weakly supervised domain adaptation for aspect extraction via multilevel interaction transfer," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5818–5829, Oct. 2022, doi: 10.1109/TNNLS.2021.3071474.

[34] T. Shi and Y. Song, "A novel two-stage generation framework for promoting the persona-consistency and diversity of responses in neural dialog systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1552–1562, Mar. 2023, doi: 10.1109/TNNLS.2021.3105584.

[35] F. Cui, Q. Cui, and Y. Song, "A survey on learning-based approaches for modeling and classification of human–machine dialog systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1418–1432, Apr. 2021, doi: 10.1109/TNNLS.2020.2985588.

[36] J. Wu et al., "Leveraging multiple types of domain knowledge for safe and effective drug recommendation," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 2169–2178.

[37] J. Wu, Y. Dong, Z. Gao, T. Gong, and C. Li, "Dual attention and patient similarity network for drug recommendation," *Bioinformatics*, vol. 39, no. 1, Jan. 2023, Art. no. btad003.

[38] K. He, R. Mao, T. Gong, C. Li, and E. Cambria, "Meta-based self-training and re-weighting for aspect-based sentiment analysis," *IEEE Trans. Affect. Comput.*, early access, Aug. 30, 2022, doi: 10.1109/TAFFC.2022.3202831.

[39] H. Bao et al., "Bert-based meta-learning approach with looking back for sentiment analysis of literary book reviews," in *Natural Language Processing and Chinese Computing*. Qingdao, China: Springer, Oct. 2021, pp. 235–247.

[40] M. Maggini, G. Marra, S. Melacci, and A. Zugarini, "Learning in text streams: Discovery and disambiguation of entity and relation instances," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4475–4486, Nov. 2020, doi: 10.1109/TNNLS.2019.2955597.

[41] J. Li, S. Shang, and L. Chen, "Domain generalization for named entity boundary detection via metalearning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3819–3830, Sep. 2021, doi: 10.1109/TNNLS.2020.3015912.

[42] R. Wang et al., "Few-shot class-incremental learning for named entity recognition," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 571–582.

[43] P. Lai et al., "PCBERT: Parent and child BERT for Chinese few-shot NER," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2199–2209.

[44] J. Chen, Q. Liu, H. Lin, X. Han, and L. Sun, "Few-shot named entity recognition with self-describing networks," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5711–5722. [Online]. Available: https://aclanthology.org/2022.acl-long.392

[45] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3045–3059.

[46] Y. Huang et al., "COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition," in *Proc. 29th Int. Conf. Comput. Linguistics (COLING)*, Gyeongju, South Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2515–2527. [Online]. Available: https://aclanthology.org/2022.coling-1.222

[47] R. Ma et al., "Template-free prompt tuning for few-shot NER," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Seattle, WA, USA: Association for Computational Linguistics, Jul. 2022, pp. 5721–5732. [Online]. Available: https://aclanthology.org/2022.naacl-main.420

[48] Z. Gao et al., "Unsupervised representation learning for tissue segmentation in histopathological images: From global to local contrast," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3611–3623, Dec. 2022.

[49] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9640–9649.

[50] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2020, pp. 9729–9738.

[51] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 10, pp. 8547–8555, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17037

[52] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng, "Twin contrastive learning for online clustering," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2205–2221, Sep. 2022.

[53] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 20001706980, 2009.

[54] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 547–558.

[55] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New York, NY, USA: Computer Vision Foundation, 2021, pp. 1134–1143. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Yang_Partially_View-Aligned_Representation_Learning_With_Noise-Robust_Contrastive_Loss_CVPR_2021_paper.html

[56] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 6382–6388.

[57] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 2236–2245, May 2022, doi: 10.1109/TNNLS.2020.3044215.

[58] Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou, and G. Karypis, "Anomaly detection on attributed networks via contrastive self-supervised learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2378–2392, Jun. 2022, doi: 10.1109/TNNLS.2021.3068344.

[59] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma, "CLEAR: Contrastive learning for sentence representation," 2020, *arXiv:2012.15466*.

[60] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 879–895. [Online]. Available: https://aclanthology.org/2021.acl-long.72

[61] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171–4186, doi: 10.18653/V1/N19-1423.

[62] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 4080–4090.

[63] W. Wen, Y. Liu, C. Ouyang, Q. Lin, and T. Chung, "Enhanced prototypical network for few-shot relation extraction," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102596.

[64] S. S. S. Das, A. Katiyar, R. Passonneau, and R. Zhang, "CONTaiNER: Few-shot named entity recognition via contrastive learning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6338–6353. [Online]. Available: https://aclanthology.org/2022.acl-long.439

[65] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[66] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5628–5637.

[67] Y. Yang and A. Katiyar, "Simple and effective few-shot named entity recognition with structured nearest neighbor learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 6365–6375. [Online]. Available: https://aclanthology.org/2020.emnlp-main.516

[68] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," 2019, *arXiv:1904.04232*.

[69] N. Ding et al., "Few-NERD: A few-shot named entity recognition dataset," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 3198–3213. [Online]. Available: https://aclanthology.org/2021.acl-long.248

[70] R. Weischedel et al., "Ontonotes release 5.0 LDC2013T19," in *Proc. Linguistic Data Consortium*, Philadelphia, PA, USA, vol. 23, 2013, pp. 23–170.

[71] E. F. T. K. Sang and F. De Meulder, "Introduction to the CONLL-2003 shared task: Language-independent named entity recognition," 2003, *arXiv:cs/0306050*.

[72] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proc. 3rd Workshop Noisy User-Generated Text*, 2017, pp. 140–147.

[73] A. Stubbs and Ö. Uzuner, "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus," *J. Biomed. Informat.*, vol. 58, pp. S20–S29, Dec. 2015.

[74] J. Liu, P. Pasupat, Y. Wang, S. Cyphers, and J. Glass, "Query understanding enhanced by hierarchical parsing structures," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2013, pp. 72–77.

[75] T. Ma, H. Jiang, Q. Wu, T. Zhao, and C.-Y. Lin, "Decomposed meta-learning for few-shot named entity recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1584–1596. [Online]. Available: https://aclanthology.org/2022.findings-acl.124

[76] P. Wang et al., "An enhanced span-based decomposition method for few-shot sequence labeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2022, pp. 1–9.

[77] C. Li et al., "Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis," 2021, *arXiv:2109.08306*.

Kai He (Member, IEEE) is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, under the supervision of Chen Li.

He is on an academic visit to the School of Computer Science and Engineering, Nanyang Technological University, Singapore, under the supervision of Erik Cambria. His research interests include information extraction and sentiment analysis in the natural language processing (NLP) field.

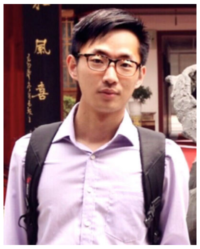
Rui Mao received the Ph.D. degree in computing science from the University of Aberdeen, Aberdeen, U.K., in 2020.

Currently, he is a Research Fellow with Nanyang Technological University, Singapore, studying conversational AI. His research interests include computational metaphor processing, token-level and sequence-level semantic understanding, and affective computing. He and his founded company have developed the first neural network search engine (www.wensousou.com) for searching ancient Chinese poems by using modern language, and a system (metapro.ruimao.tech) for linguistic and conceptual metaphor understanding. He has published several papers as the first author in top-tier international conferences and journals, e.g., Annual Meeting of the Association for Computational Linguistics (ACL), Association for the Advancement of Artificial Intelligence (AAAI), and *Information Fusion*.

Dr. Mao served as the Area Chair for International Conference on Computational Linguistics (COLING) and Conference on Empirical Methods in Natural Language Processing (EMNLP), and a reviewer for *Knowledge-Based Systems (KBS)*, *COGN COMPUT*, and *INFFUS*.

Yucheng Huang received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently pursuing the M.E. degree in computer science with Xi'an Jiaotong University, Xi'an, under the supervision of Chen Li.

His research interests include information extraction, noise learning, and low-resource learning in natural language processing.

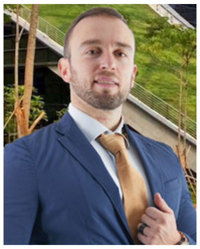


Tieliang Gong graduated from Xi'an Jiaotong University, Xi'an, China, in 2018.

He conducted postdoctoral research at the School of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada, from October 2018 to September 2020. His research interests include statistical learning theory, robust statistical inference, and machine learning. His research results are mainly published in NeurIPS, Association for the Advancement of Artificial Intelligence (AAAI), *IEEE Transactions on Cybernetics*, *Neural Computing and Applications*, and other top international conferences and journals, and participated in two general projects of the National Natural Science Foundation of China.

Chen Li (Member, IEEE) is a Doctor at the University of Cambridge, Cambridge, U.K., and a Postdoctoral Fellow with the Massachusetts Institute of Technology, Cambridge, MA, USA. At present, he works as a Professor with Xi'an Jiaotong University, Xi'an, China. He has been engaged in the research and development of biomedical text mining in EMBL-EBI. The biomodels data standard system developed under his leadership has been rated as the most important resource in the field of systems biology, with more than 200 journals in several top international academic publishing institutions supporting biomodels and recommending contributors to store data on the platform.

Mr. Li was an Overseas Fellowship Winner at Cambridge University and a Fellowship Winner at the European Molecular Biology Laboratory. He is a reviewer and project reviewer of many international journals and conferences, the project reviewer of BBSRC, and the Organizing Committee Member of Conference on Empirical Methods in Natural Language Processing (EMNLP) and Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).



Erik Cambria (Fellow, IEEE) received the joint Ph.D. degree from the University of Stirling, Stirling, U.K., and the MIT Media Laboratory, Cambridge, MA, USA, in 2012, through a joint program.

He is the Founder of SenticNet, Singapore, a Singapore-based company offering B2B sentiment analysis services, and a Professor at NTU, Singapore, where he also holds the appointment of Provost Chair in computer science and engineering. Prior to joining NTU, he worked at Microsoft Research Asia, Beijing, China, and HP Labs India, Bengaluru, India. His research interests include neurosymbolic AI for explainable natural language processing in domains, such as sentiment analysis, dialog systems, and financial forecasting.

Dr. Cambria was a recipient of several awards, such as the IEEE Outstanding Career Award. He was listed among the AI's 10 to watch. He was featured in Forbes as one of the five people building our AI future. He is involved in various international conferences as the program chair, SPC member, and invited speaker. He is an Associate Editor of many top-tier AI journals, such as *INFFUS* and IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.