EPNER: An External Prototype Augmented Framework for Few-Shot Named Entity Recognition

Weixin Zuo
Faculty of Artificial Intelligence
Shanghai University of Electric Power
Shanghai, China
zuoweixin@mail.shiep.edu.cn

Xilong Wang*
Faculty of Artificial Intelligence
Shanghai University of Electric Power
Shanghai, China
sallywang@shiep.edu.cn

Nuoxian Liu
Faculty of Artificial Intelligence
Shanghai University of Electric Power
Shanghai, China
758160601@mail.shiep.edu.cn

***Abstract*—Few-Shot Named Entity Recognition (NER) aims to identify novel entities relying on scarce annotations. However, the traditional two-stage span-based model is still constrained by two major bottlenecks: the increase of false-positive spans and the instability of prototypes derived from sparse support sets, which exacerbates the indistinguishability of fine-grained categories. We propose EPNER to tackle these issues by augmenting prototypes with external knowledge. In the detection stage, we introduce an Entity-Non-Entity Contrastive Learning mechanism to enhance representational distinctiveness, complemented by a learnable width bias to suppress spurious candidate spans. In the classification stage, we construct a high-quality external knowledge prototype library by large language models (LLMs) and frequent entities from external corpora. They are used as stable prior knowledge. These priors are integrated with support prototypes via a Dynamic Attention Fusion mechanism. Furthermore, we have also designed a prototype separation loss function to widen the inter-class distance between different prototype categories. A large number of experiments have shown that our proposed method significantly outperforms state-of-the-art approaches.**

***Index Terms*—Few-Shot Learning, Named Entity Recognition, Prototypical Networks, Contrastive Learning.**

I. Introduction

Named Entity Recognition (NER) is a fundamental task in information extraction and natural language processing (NLP), aiming to locate and classify entities (e.g., persons, locations and organizations) within unstructured text [1]. Although traditional deep learning-based NER approaches have achieved remarkable success, they rely heavily on large-scale, high-quality annotated data. In many specialized domains, such as medicine, industry, and law, acquiring such datasets is expensive. Consequently, Few-Shot NER has garnered significant attention in recent years [2]. Its few-shot learning capability renders it substantial potential in nascent domains, including medicine, finance, and industry.

Recently, the two-stage method has emerged as the dominant paradigm for Few-Shot NER [3]. This paradigm typically employs a class-agnostic Span Detector to extract all potential entity candidate spans, followed by a classifier to assign these spans to N target categories. This approach reduces interference from non-entity tokens 'O' in traditional methods and better handles entity boundaries [4].

*Corresponding author.

Two-stage methods face two critical challenges. The Span Detector in the detection stage is class-agnostic. It tends to recall "plausible spans" without selection. It leads to an increase of false positives, which significantly increases the computational burden and introduces substantial noise into the subsequent classification stage. Due to the extreme scarcity of samples in the classification stage, contextual prototypes derived from only K support samples are highly susceptible to sample bias. It may lead to create inaccurate and unstable representations. Fine-grained categories with similar semantics often exhibit severe overlap in the embedding space, hindering discrimination. The prototypes for "person-actor" and "person-director" may become indistinguishable in the latent space. It is difficult for the classifier to separate them effectively.

We have proposed a novel two-stage collaborative framework called "External Prototype Augmented Named Entity Recognition" (EPNER) to address these issues. Our main contributions are summarized as follows:

- We propose "Entity-Non-Entity Contrastive Learning" to optimize latent feature discrimination and reduce false positives, complemented by a learnable Width Bias that leverages length inductive bias to filter spurious spans.
- We propose a Knowledge-Augmented Fusion Strategy in order to construct stable and distinct prototypes. This approach combines stable external priors—derived from LLMs and external corpora—with context-aware prototypes via a Dynamic Attention Fusion mechanism. Furthermore, a Prototype Separation Module is designed to maximize inter-class semantic distances, effectively alleviating confusion among fine-grained categories.
- Extensive experiments on two standard benchmarks demonstrate the superiority of our proposed framework over state-of-the-art baselines.

II. Related Work

A. Few-Shot NER

Few-Shot NER [5] aims to identify/classify novel entity types with minimal annotated samples under the N-way K-shot meta-learning paradigm. Unlike traditional NER relying on large labeled data, it faces classification bias and entity boundary ambiguity. The mainstream "span detection

followed by classification" mode eliminates non-entity ("O") interference but still has two bottlenecks: false alarms in span detection, and classification confusion from limited K-sample prototypes.

B. Prototypical Network

The prototype network is the main paradigm in this field. Its core idea is that samples of the same category cluster around the prototype in the embedding space. The query sample completes classification by measuring the distance between itself and various prototypes. [6] proposed constructing class prototypes by computing the mean vector of samples in the support set. It classifies query samples based on a nearest-neighbor criterion. In the realm of NER, [7] adapted this architecture for few-shot settings, calculating class prototypes by averaging the embeddings of all tokens belonging to each specific entity type. However, the performance of such methods is heavily contingent on the quality of the support set. When the sample size is extremely small or exhibits sampling bias, the variance of prototype estimation increases significantly. This results in a drift of decision boundaries, thereby impeding accurate classification.

C. Prompt Learning

Prompt-based methods guide entity recognition via pre-designed natural language templates, leveraging rich priors from Pre-trained Language Models (PLMs). Representative works include the Self-Describing Network [8] for enhanced entity type representations, and a BART-based framework reformulating NER as text infilling [9]. While they enable architecture-agnostic domain knowledge injection, their performance is limited by template design quality and pre-training objective alignment, with unsolved challenges in open-domain adaptation and entity boundary dependency modeling.

Despite progress in Few-Shot NER, reliance on sparse support sets remains a critical classification bottleneck. Targeting these bottlenecks, we propose an external knowledge-augmented robust framework. It strives to achieve stable and precise entity recognition under extremely low-resource conditions.

III. Task Definition

Formally, we adopt the standard meta-learning paradigm for Few-Shot NER. The entire dataset $\mathcal{D}$ is partitioned into three disjoint subsets: a source domain $\mathcal{D}_{train}$, a development domain $\mathcal{D}_{eval}$, and a target domain $\mathcal{D}_{test}$.

Adopting the standard N-way K-shot episodic paradigm, we sample episodes $\mathcal{E} = (\mathcal{S}, \mathcal{Q}, \mathcal{T})$ from $\mathcal{D}$. Here, the support set $\mathcal{S}$ contains K labeled examples for each of the N types in $\mathcal{T}$, providing supervision for the query set $\mathcal{Q}$. A 2-way 1-shot example is depicted in Fig. 1.

We formulate the task as a span-based classification problem. For a given input sentence $X = \{x_1, x_2, \ldots, x_L\}$ comprising L tokens, the objective is to extract a set of mention spans $\mathcal{G} = \{(s_j, e_j)\}_{j=1}^{M}$ and assign a corresponding label $y_j \in \mathcal{T}$ to each span, where s_j and e_j represent the start and end indices, respectively.

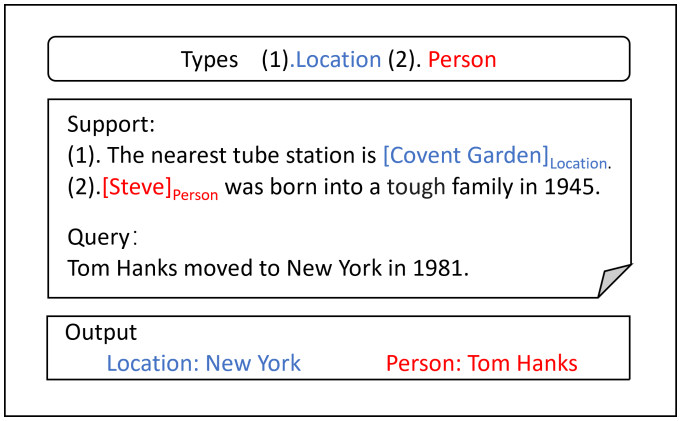

Fig. 1. An example of the 2-way 1-shot setting on NER task.

IV. Methodology

As illustrated in Fig. 2, the proposed framework consists of two stages: **Span Detection** and **Entity Classification**.

A. Span Detection Stage

The objective of this stage is to efficiently and accurately recall all candidate entity spans while maximally suppressing false positives.

1) Base Detector: We employ Global Pointer [10] as the backbone for span detection. The core premise of this module is to reformulate the recognition of a span (s, e) as a token-pair interaction scoring problem between the start index s and the end index e.

Given an input sentence, we first utilize a BERT encoder to extract the contextualized representation $\mathbf{h}_i \in \mathbb{R}^d$ for each token i. Subsequently, two independent learnable linear transformations project each token representation $\mathbf{h}_i$ into a query vector $\mathbf{q}_i$ and a key vector $\mathbf{k}_i$. Formally, we have:

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{h}_i + \mathbf{b}_q, \quad \mathbf{k}_i = \mathbf{W}_k \mathbf{h}_i + \mathbf{b}_k \tag{1}$$

where $\mathbf{q}_i, \mathbf{k}_i \in \mathbb{R}^d$ denote the query and key embeddings, respectively; $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d \times d}$ are learnable weight matrices; and $\mathbf{b}_q, \mathbf{b}_k \in \mathbb{R}^d$ represent the bias vectors.

Drawing inspiration from the self-attention mechanism [11], we calculate the score for a potential span starting at i and ending at j as follows:

$$S(i, j) = \mathbf{q}_i^T \mathbf{k}_j + \mathbf{W}_v (\mathbf{h}_i + \mathbf{h}_j) \tag{2}$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times d}$ is a trainable parameter matrix.

Let $S(s, e)$ denote the final score assigned to the span (s, e). We obtain the probability via the Sigmoid function. Following the decoding strategy in [12], we recall all candidate spans whose scores exceed a predefined confidence threshold. The training objective is defined as:

$$\mathcal{L}_{span} = \log \left(1 + \sum_{1 \le s \le e \le L} \exp \left((-1)^{\Omega(s,e)} S(s, e) \right) \right) \tag{3}$$

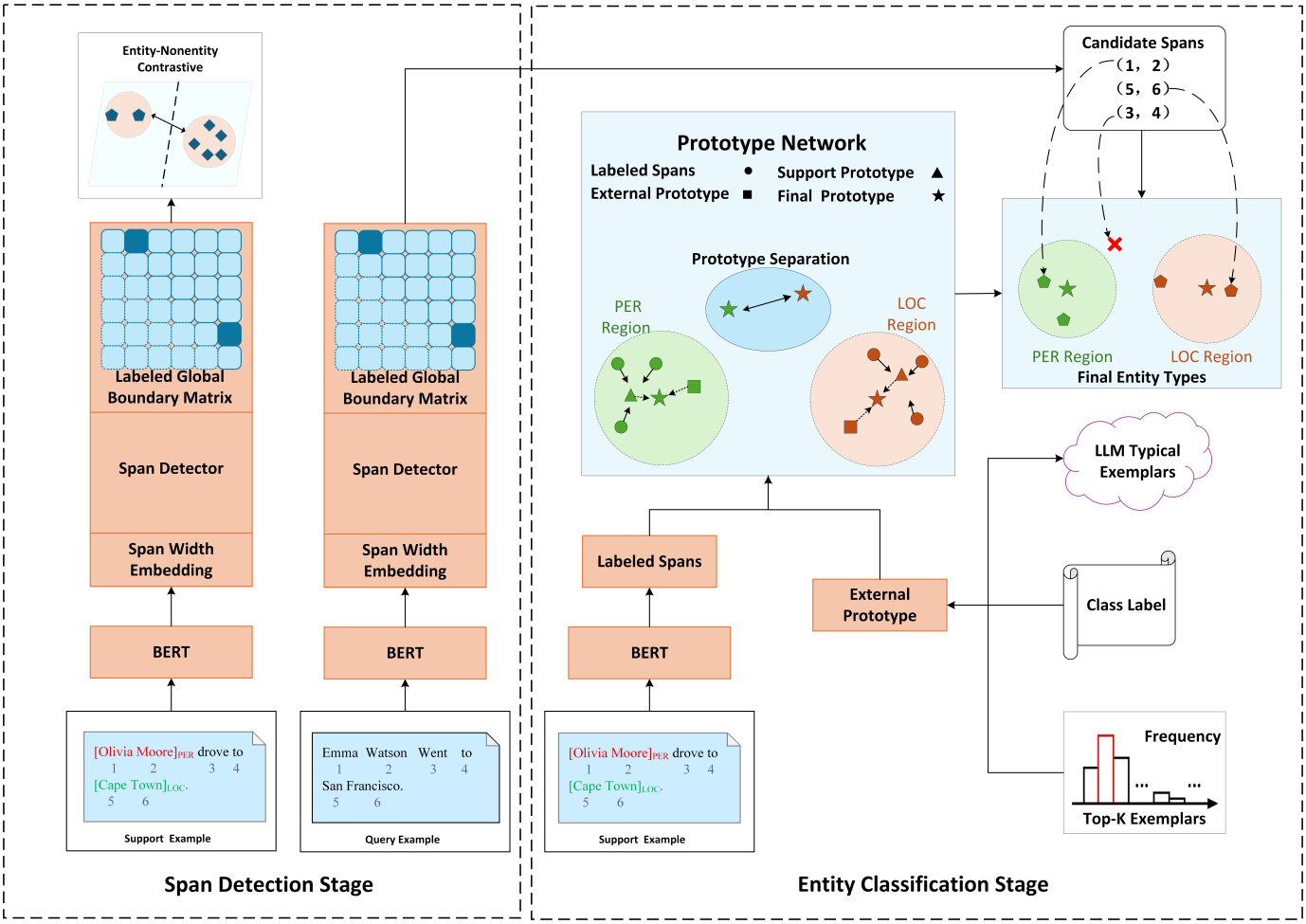

Fig. 2. Overall architecture of EPNER: a two-stage framework with contrastive span detection and external prototype augmented entity classification.

where $\Omega(s, e)$ represents the global boundary matrix, taking a value of 1 if the span (s, e) corresponds to a ground-truth entity (positive sample), and 0 otherwise.

2) Width Bias Mechanism: Named entities typically exhibit characteristic prior distributions regarding their lengths. We incorporate a learnable embedding matrix $\mathbf{E}_{\text{width}} \in \mathbb{R}^{(L_{\max}+1)\times 1}$, where $L_{\max}$ denotes the maximum allowable span width. For an arbitrary span (s, e), we first compute its width $w = e - s$. Subsequently, the corresponding width bias b_w is retrieved from the embedding matrix:

$$b_w = \mathbf{E}_{\text{width}}[w] \tag{4}$$

The final detection score is formulated as:

$$S'(s, e) = S(s, e) + b_w \tag{5}$$

Note: For brevity, we denote this final score as $S(s, e)$ hereafter.

This mechanism empowers the model to implicitly capture the prior probabilities associated with varying span lengths, thereby effectively suppressing false-positive spans that exhibit implausible lengths.

3) Entity-Non-Entity Contrastive Learning: We incorporate contrastive learning to enhance the model's ability to distinguish between entities and non-entities at the semantic level. Adhering to the findings in SimCLR [13], we first project the contextual representation $\mathbf{h}$ yielded by the BERT encoder into a lower-dimensional latent feature $\mathbf{z}$ via a non-linear projection head. For an arbitrary span (s, e), its specific representation is derived by aggregating the projected embeddings of its start and end boundaries: $\mathbf{z}_{(s,e)} = \mathbf{z}_s + \mathbf{z}_e$.

For each positive anchor span (s_i, e_i), the contrastive loss is formulated as follows. Let $\mathcal{B} = \mathcal{P}_{\text{pos}} \cup \mathcal{N}_{\text{neg}}$ denote the union set of spans in the current contrastive batch:

$$L^{(i)}_{\text{contrast}} = -\log \frac{\sum_{(s_p,e_p)\in\mathcal{P}_{\text{pos}}\setminus(s_i,e_i)} \exp\left(\frac{\text{sim}(\mathbf{z}(s_i,e_i),\mathbf{z}(s_p,e_p))}{\tau}\right)}{\sum_{(s_k,e_k)\in\mathcal{B}\setminus(s_i,e_i)} \exp\left(\frac{\text{sim}(\mathbf{z}(s_i,e_i),\mathbf{z}(s_k,e_k))}{\tau}\right)} \tag{6}$$

where $\mathcal{P}_{\text{pos}}$ and $\mathcal{N}_{\text{neg}}$ represent the sets of positive (entity) and negative (non-entity) spans, respectively. The function $\text{sim}(\cdot, \cdot)$ computes the cosine similarity between two vectors, with τ being the temperature hyperparameter. In the summation,

p iterates over positive samples with the same label as the anchor, and k iterates over all spans in the union set $\mathcal{B}$.

B. Entity Classification Stage

The objective of this stage is to accurately classify the candidate spans recalled from the preceding stage.

1) External Prototype: We constructed a high-quality prototype library of external knowledge before conducting the model training. This injects stable and real-world relevant prior knowledge into the model. This process comprises the following steps:

LLMs-Based Canonical Instance Generation: We leverage the robust world knowledge and comprehension capabilities of Large Language Models (LLMs) to generate "canonical" instances for each entity category. To this end, we design a refined Prompt Template, as illustrated in Fig. 3.

High-Frequency Instance Mining: Inspired by [14], we mine high-frequency entities from Wikipedia to complement LLM priors and align prototypes with real-world distributions. We first align our label system with Wikidata for category consistency, then conduct entity linking, frequency statistics and filtering to select the top 3 representative entities per category (excluding overlaps with LLM-generated instances). These mined instances and LLM-generated ones together form the external instance set for prototype construction.

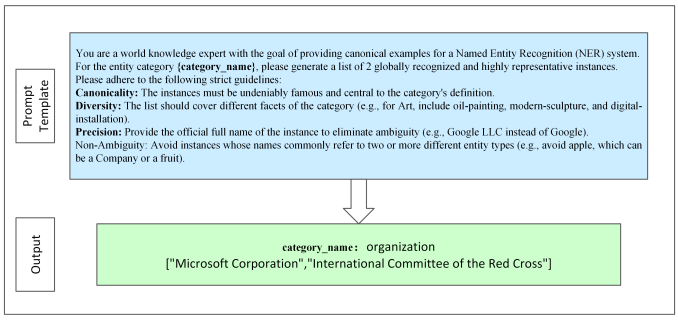

Fig. 3. An example of the prompt template and the corresponding output generated by the LLMs for the "organization" category.

We concatenate the collected instances with the category name to synthesize a Prototypical Sentence using the template: *"[Instance 1], ..., [Instance N] are all [Category Name]."*

We take the *organization* category as an illustration, after integrating the LLMs-generated and high-frequency instances, the resulting prototypical sentence is formulated as: *"Microsoft Corporation, International Committee of the Red Cross, United Nations and Google LLC are all organizations."*

Finally, leveraging the principle of Prototypical Networks, we apply average pooling over the contextual embeddings of all constituent entity spans and the category label token within the sentence to derive the External Knowledge Prototype $P_{external}$ for each category. This procedure yields a stable and information-rich External Knowledge Prototype Library.

2) Dynamic Attention Fusion: We use a dynamic attention fusion mechanism to integrate external knowledge with the contextual information. For a given category c, the contextual prototype $P^{(c)}_{contextual}$, derived from the support set, serves as the semantic anchor (Query), while the external knowledge prototype $P^{(c)}_{external}$ functions as both the Key and Value. Crucially, this design inherently mitigates the impact of potential noise or hallucinations from LLMs. Since the attention weights are calculated based on the semantic similarity between the verified support examples (Query) and the external priors (Key), irrelevant or conflicting information in $P^{(c)}_{external}$ is assigned negligible weights. This ensures that the model selectively fuses only the knowledge that is semantically consistent with the current task definition. This mechanism injects external priors into the contextual representation by using a residual connection,effectively alleviating the inherent problem of information scarcity in scenarios with limited samples while maintaining robustness against external noise. Consequently, this process yields a robust final category prototype $P^{(c)}_{final}$. The calculation is formulated as follows:

$$P^{(c)}_{final} = P^{(c)}_{contextual} + \text{Attention}(Q = P^{(c)}_{contextual}, K = P^{(c)}_{external}, V = P^{(c)}_{external}) \tag{7}$$

3) Prototype Separation: We use a prototype Separation Loss to enhance the discriminability of fine-grained categories by optimizing the distribution of the feature space. This objective explicitly minimizes the semantic similarity between prototypes of distinct categories. The loss function is defined as follows:

$$L_{\text{sep}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1 j \neq i}^{N} \max\left(0, (P^{(i)}_{final})^T P^{(j)}_{final}\right) \tag{8}$$

4) Classification: We introduce a Margin-based Loss, denoted as $\mathcal{L}_{mrg}$. This objective is designed to repel spurious candidate spans—those not belonging to any target category—away from all known category prototypes by at least a pre-defined margin r. This constraint empowers the model to robustly distinguish and reject non-target entities. The formal definition is as follows:

$$\mathcal{L}_{mrg} = \frac{1}{|\mathcal{M}_{fp}| \cdot N} \sum_{\mathbf{u}_k \in \mathcal{M}_{fp}} \sum_{c=1}^{N} \max\left(0, r - d(\mathbf{u}_k, P^{(c)}_{final})\right) \tag{9}$$

where $\mathcal{M}_{fp}$ denotes the set of false positive spans, $\mathbf{u}_k$ represents the embedding of a false positive span k, $P^{(c)}_{final}$ indicates the final prototype for category c, and $d(\cdot, \cdot)$ denotes the Euclidean distance.

For a candidate span (s_j, e_j) in the query set, let $\mathbf{q}_j$ denote its embedding representation. We calculate the classification probability by computing the Softmax over the negative squared Euclidean distances between $\mathbf{q}_j$ and all final prototypes $P^{(c)}_{final}$:

$$p(y = c|(s_j, e_j)) = \frac{\exp\left(-d(\mathbf{q}_j, P^{(c)}_{final})^2\right)}{\sum_{c'=1}^{N} \exp\left(-d(\mathbf{q}_j, P^{(c')}_{final})^2\right)} \tag{10}$$

TABLE I

F1 SCORES WITH STANDARD DEVIATIONS ON FEW-NERD DATASET. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINED INDICATES THE SECOND-BEST RESULT. THE RESULTS OF BASELINE MODELS ARE CITED FROM THEIR ORIGINAL PAPERS.

Models	Intra					Inter				
	1∼2-shot		5∼10-shot		Avg.	1∼2-shot		5∼10-shot		Avg.
	5 way	10 way	5 way	10 way		5 way	10 way	5 way	10 way	
ProtoBERT	20.76±0.84	15.05±0.44	42.54±0.94	35.40±0.13	28.44	38.83±1.49	32.45±0.79	58.79±0.44	52.92±0.37	45.75
NNShot	25.78±0.91	18.27±0.41	36.18±0.79	27.38±0.53	26.90	47.24±1.00	38.87±0.21	47.24±1.00	38.87±0.21	43.06
StructShot	30.21±0.90	21.03±1.13	38.00±1.29	26.42±0.60	28.92	51.88±0.69	43.34±0.10	57.32±0.63	49.57±3.08	50.53
DecomposedMetaNER	49.48±0.85	42.84±0.46	62.92±0.57	57.31±0.25	53.14	68.77±0.24	63.26±0.40	71.62±0.16	68.32±0.10	67.99
CEPTNER	**50.43±0.77**	44.18±0.27	65.70±0.16	59.85±0.16	55.04	64.92±0.23	59.79±0.35	72.24±0.16	69.52±0.05	66.62
MetaR	45.27±0.42	37.67±0.65	63.54±0.25	55.33±0.46	50.45	62.49±0.33	56.70±0.14	71.22±0.42	66.83±0.26	64.31
EPNER (Ours)	47.86±0.62	**45.78±0.25**	**66.38±0.18**	**61.83±0.15**	**55.46**	**70.23±0.21**	**63.82±0.32**	**76.14±0.14**	**71.10±0.11**	**70.32**

TABLE II

F1 SCORES WITH STANDARD DEVIATIONS ON CROSS-DATASET. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINED INDICATES THE SECOND-BEST RESULT. THE RESULTS OF BASELINE MODELS ARE CITED FROM THEIR ORIGINAL PAPERS.

Models	1-shot					5-shot				
	CoNLL	GUM	WNUT	Ontonotes	Avg.	CoNLL	GUM	WNUT	Ontonotes	Avg.
ProtoBERT	32.49±2.01	3.89±0.24	10.68±1.40	6.67±0.46	13.43	50.06±1.57	9.54±0.44	17.26±2.65	13.59±1.61	22.61
DecomposedMetaNER	46.09±0.44	17.54±0.98	25.14±0.24	34.13±0.92	30.73	58.18±0.87	31.36±0.91	31.02±1.28	45.55±0.90	41.53
CEPTNER	47.06±0.52	18.43±0.77	25.86±0.27	34.75±0.65	31.53	59.64±0.90	**33.58±0.79**	32.60±0.73	46.28±0.93	43.03
MetaR	49.64±0.59	17.73±0.38	28.46±0.45	35.33±0.86	32.79	59.09±0.45	29.90±0.75	32.96±0.51	46.83±0.69	42.20
EPNER (Ours)	**50.73±0.48**	**19.41±0.55**	**29.52±0.31**	**36.95±0.58**	**34.15**	**61.29±0.42**	32.86±0.68	**33.52±0.49**	**48.63±0.55**	**44.08**

The classification loss is defined as:

$$\mathcal{L}_{proto} = -\frac{1}{|\mathcal{Q}|} \sum_{(s_j, e_j) \in \mathcal{Q}} \log p\left(y = y_j \mid (s_j, e_j)\right) \tag{11}$$

During the inference phase, the final predicted category is determined by the class with the highest probability:

$$\hat{y} = \underset{c \in \{1, \ldots, N\}}{\arg\max}\ p(y = c | (s_j, e_j)) \tag{12}$$

Consistent with training, we enforce the margin constraint during inference. Specifically, spans with distances to their predicted prototype exceeding r are classified as non-entities.

V. EXPERIMENTS

A. Datasets and Baselines

We evaluate our framework on two benchmarks under the N-way K-shot setting: Few-NERD [15] and Cross-Dataset [16]. Few-NERD features a hierarchy of 8 coarse-grained and 66 fine-grained types with two evaluation modes: **Intra**, where coarse-grained types are mutually exclusive across splits, and **Inter**, where fine-grained types are strictly disjoint despite sharing coarse categories. Cross-Dataset encompasses four diverse domains: CoNLL-03(News) [17], GUM(Wiki) [18], WNUT-17(Social) [19], and OntoNotes(Mixed) [20]. Two datasets of them are randomly selected to constitute the training set, while the remaining two serve as the development and test sets, respectively. For fair comparison, we utilize the standard pre-processed episode data provided in [15] and [16].

We benchmark our method against competitive baselines, including ProtoBERT [15], NNShot [2], StructShot [2], DecomposedMetaNER [3], CEPTNER [21], and MetaR [22].

B. Implementation Details

Following prior works, we utilize BERT-base-uncased [23] as the encoder, with the maximum sequence length set to 128. We employ the AdamW optimizer with a warmup rate of 0.1. We set the classification margin r to 6. To ensure reproducibility, all experiments are conducted over five random seeds $\{12, 21, 42, 87, 100\}$, and we report the averaged results along with their standard deviations. This paper utilizes large language models (LLMs) with Gemini 2.5 Pro to build external knowledge. The temperature of the LLMs is set to 0.7. We take 3 LLMs-generated canonical examples and top-3 high-frequency entities as our external knowledge.

C. Results

Consistent with standard NER evaluation protocols, we adopt the F1 score as the primary metric. Table I and Table II summarize the experimental results on the Few-NERD and Cross-Dataset benchmarks, respectively. As observed, compared with state-of-the-art (SOTA) baselines, our method achieves average improvements of 0.42% and 2.33% on the Intra and Inter settings of Few-NERD. Notably, under the 5-way 5∼10-shot setting of the Inter scenario, our model yields a significant gain of 3.9%. Similarly, on Cross-Dataset, the results improve by 1.36% and 1.05% under the 1-shot and 5-shot settings. The lower performance on the Few-NERD Intra scenario compared to Inter underscores its difficulty, stemming from the disjoint coarse-grained types across data splits. Nevertheless, EPNER achieves competitive performance. This validates our model's robust adaptation capability to novel domains characterized by unseen coarse and fine-grained hierarchies.

TABLE III
AVERAGE F1 SCORES FROM FEW-NERD ABLATION STUDY.

Model	Intra	Inter
EPNER (Ours)	**55.46**	**70.32**
w/o Entity-Non-Entity Contrastive	54.31	68.53
w/o Width Bia	55.03	69.18
w/o External Prototype	53.83	67.18
w/o Prototype Separation	54.62	67.49

D. Ablation Study

Ablation studies on Few-NERD (Table III) compare full EPNER with four variants: **w/o Entity-Non-Entity Contrastive**, **w/o Width Bias**, **w/o External Prototype**, and **w/o Prototype Separation**. Removing any component causes consistent F1 drops on both splits, verifying their indispensability. The severest degradation comes from removing external prototypes (Intra: -1.63, Inter: -3.14), highlighting its role in stabilizing prototypes. The other two modules reduce false positives and fine-grained confusion respectively, while the lightweight Width Bias effectively filters implausible spans. All components work synergistically to boost model performance.

VI. CONCLUSION

In this paper, we propose a two-stage cooperative framework that integrates external knowledge augmentation with robust training mechanisms. We introduce Entity-Non-Entity Contrastive Learning and a Width Bias mechanism to effectively reduce false positives in the span detection stage. For the classification stage, we construct an External Knowledge Prototype Library and employ Dynamic Attention Fusion alongside a Prototype Separation Loss. They enable the generation of stable and highly discriminative prototypes. Extensive experiments demonstrate that our framework consistently outperforms strong baselines. In future work, we plan to extend our framework to zero-shot scenarios and refine it for industrial and other specialized domains.

REFERENCES

[1] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[2] Y. Yang and A. Katiyar, "Simple and effective few-shot named entity recognition with structured nearest neighbor learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6365–6375.

[3] 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

[4] P. Wang, R. Xu, T. Liu, Q. Zhou, Y. Cao, B. Chang, and Z. Sui, "An enhanced span-based decomposition method for few-shot sequence labeling," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5012–5024.

[5] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] A. Fritzler, V. Logacheva, and M. Kretov, "Few-shot classification in named entity recognition task," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019, pp. 993–1000.

[8] J. Chen, Q. Liu, H. Lin, X. Han, and L. Sun, "Few-shot named entity recognition with self-describing networks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5711–5722.

[9] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using BART," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1835–1845.

[10] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, W. Huang, B. Wen, and Y. Liu, "Global Pointer: Novel efficient span-based approach for named entity recognition," *arXiv preprint arXiv:2208.03054*, 2022. [Online]. Available: https://arxiv.org/abs/2208.03054

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6470–6476.

[13] D. Jiang, W. Li, M. Cao, W. Zou, and X. Li, "Speech SimCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning," in *Interspeech*, 2021.

[14] R. Ma, X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, and X.-J. Huang, "Template-free prompt tuning for few-shot NER," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5721–5732.

[15] N. Ding, G. Xu, Y. Chen, X. Wang, X. Han, P. Xie, H. Zheng, and Z. Liu, "Few-NERD: A few-shot named entity recognition dataset," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3198–3213.

[16] Y. Hou, W. Che, Y. Lai, Z. Zhou, Y. Liu, H. Liu, and T. Liu, "Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1381–1393.

[17] E. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[18] A. Zeldes, "The GUM corpus: Creating multilayer resources in the classroom," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 581–612, 2017.

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

[20] S. Pradhan, A. Moschitti, N. Xue, H. T. Ng, A. Björkelund, O. Uryupina, Y. Zhang, and Z. Zhong, "Towards robust linguistic analysis using OntoNotes," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 2013, pp. 143–152.

[21] E. Zha, D. Zeng, M. Lin, and Y. Shen, "Ceptner: Contrastive learning enhanced prototypical network for two-stage few-shot named entity recognition," *Knowledge-Based Systems*, vol. 295, p. 111730, 2024.

[22] B. Sun, K. Gong, W. Li, and X. Song, "MetaR: Few-shot named entity recognition with meta-learning and relation network," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 4171–4186.