Robust Few-Shot Named Entity Recognition with Boundary Discrimination and Correlation Purification

Xiaojun Xue, Chunxia Zhang*, Tianxiang Xu, Zhendong Niu

School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
{xiaojunx, cxzhang, txxu, zniu}@bit.edu.cn

Abstract

Few-shot named entity recognition (NER) aims to recognize novel named entities in low-resource domains utilizing existing knowledge. However, the present few-shot NER models assume that the labeled data are all clean without noise or outliers, and there are few works focusing on the robustness of the cross-domain transfer learning ability to textual adversarial attacks in few-shot NER. In this work, we comprehensively explore and assess the robustness of few-shot NER models under textual adversarial attack scenario, and found the vulnerability of existing few-shot NER models. Furthermore, we propose a robust two-stage few-shot NER method with Boundary Discrimination and Correlation Purification (BDCP). Specifically, in the span detection stage, the entity boundary discriminative module is introduced to provide a highly distinguishing boundary representation space to detect entity spans. In the entity typing stage, the correlations between entities and contexts are purified by minimizing the interference information and facilitating correlation generalization to alleviate the perturbations caused by textual adversarial attacks. In addition, we construct adversarial examples for few-shot NER based on public datasets Few-NERD and Cross-Dataset. Comprehensive evaluations on those two groups of few-shot NER datasets containing adversarial examples demonstrate the robustness and superiority of the proposed method.

Introduction

Few-shot named entity recognition (NER) aims to locate and classify new named entities in the target domain where there are only a few labeled examples (Lample et al. 2016; Kato et al. 2020; Li et al. 2022b; Ma et al. 2023), which is trained using the present data within the source domain. Recently, few-shot NER has attracted increasing attention, largely due to the fact that it reduces the dependence on labeled data in NER (Yang and Katiyar 2020; Li et al. 2022a). Few-shot NER can reflect the generalized learning and cross-domain knowledge transfer abilities of humans (Lake, Salakhutdinov, and Tenenbaum 2013; Lu et al. 2021), who can infer new knowledge from a few examples based on existing knowledge.

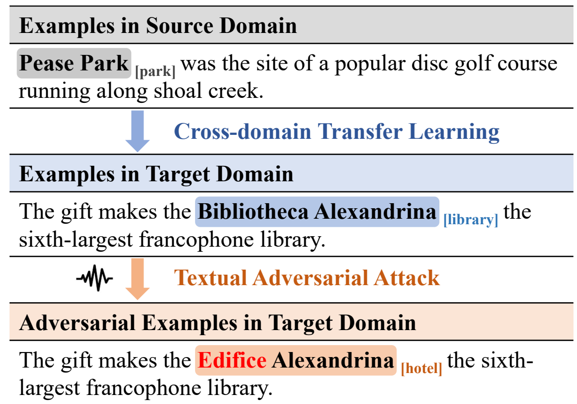


Figure 1: Vulnerabilities exhibited in few-shot NER under textual adversarial attack (i.e. synonym substitution) scenario. The subscripts indicate entity types.

Present few-shot NER methods are usually developed on token-level metric learning (Fritzler, Logacheva, and Kretov 2019; Hou et al. 2020; Yang and Katiyar 2020) and span level metric learning (Yu et al. 2021; Wang et al. 2022a). In the former family, novel entities in target domain are recognized by measuring the distance between each query token and the prototype of each entity category or each token of support examples (Snell, Swersky, and Zemel 2017). In contrast, the latter family bypasses the token-wise label dependency issue by measuring the distance between spans (Ma et al. 2022). However, existing few-shot NER methods typically assume that the examples are all clean without noise or outliers (Lu et al. 2021), which is overly idealistic in the real world. Thus, it is essential to maintain robust cross-domain transfer learning ability when handling adversarial examples with interferences for few-shot NER. In addition, at present, some works have investigated the robustness of NER task (Lin et al. 2021; Wang et al. 2022b), emphasizing the context-based reasoning. Different from NER task, few-shot NER focuses on learning the correlation transfer between entity and contexts from source domain to target domain. Our deep investigation on literature reveals that this particular topic has not gained enough attention.

Therefore, in this paper, we explore and assess the adver-

*Corresponding author.

sarial robustness of the cross-domain transfer learning ability in few-shot NER. For textual adversarial attack, synonym substitution is a widely used method, where the words in original texts are replaced by their synonyms (Li et al. 2020, 2021; Zeng et al. 2023). The substituted words are imperceptible to humans, while they can "fool" the neural networks to make wrong predictions. In this work, synonym substitution attack is performed on clean samples in few-shot NER. We have observed that synonym substitution attack misleads the existing few-shot NER models, as shown in Figure 1, and resulting in a significant drop in recognition performance for unseen entity types (detailed in Experiments section). Typically, the cross-domain transfer learning ability in few-shot NER is vulnerable when handling adversarial examples.

Motivated by the above observations, we proposed a robust two-stage few-shot NER method with Boundary Discrimination and Correlation Purification (BDCP). Overall, the span detection stage is to detect the entity spans in the input text, then the entity typing stage is to classify the detected entity spans to the corresponding unseen entity types. First, in the span detection stage, the entity boundary discriminative module is introduced to provide a highly distinguishing boundary representation space to detect entity spans, which contains multiple components assigned to all boundary classes. The span detection is regarded as a boundary classification task (e.g. BIOES (Ma et al. 2022)), and the token representations are diversely assigned to the corresponding closest components in the entity boundary discriminative module. The backbone model (i.e. encoder layer) and the boundary discriminative module are simultaneously learned by utilizing two mutually complementary losses, which can improve the adversarial robustness of span detection. Second, in the entity typing stage, the correlations between entities and contexts are purified to alleviate the perturbations caused by adversarial attacks. The correlations are purified by minimizing the interference information in correlations and facilitating correlation generalization from an information theoretic perspective, which can alleviate the perturbations caused by textual adversarial attacks.

The contributions of this paper are summarized as follows:

- We explore and assess the adversarial robustness of the cross-domain transfer learning ability in few-shot NER. A robust two-stage few-shot NER method with Boundary Discrimination and Correlation Purification (BDCP) is proposed to defend against the textual adversarial attacks. The codes are publicly available[1].
- Entity boundary discriminative module is introduced to provide a highly distinguishing boundary representation space, thereby improving the adversarial robustness of entity span detection. Two mutually complementary losses are utilized to diversely assign each token representation to corresponding component.
- Aiming at improving the adversarial robustness of entity typing, we implement correlation purification between entities and contexts by minimizing the interference information in correlations and facilitating correlation generalization. Correlation purification can alleviate the perturbations caused by textual adversarial attacks.

[1]https://github.com/ckgconstruction/bdcp

Related Work

Few-Shot Named Entity Recognition

Current works about few-shot named entity recognition (NER) mainly focus on metric learning methods at token-level (Fritzler, Logacheva, and Kretov 2019; Yang and Katiyar 2020) and span-level (Ma et al. 2022; Wang et al. 2022a). The token-level approaches recognize novel entities in query samples by measuring the distance between each query token and the prototype of each entity category or each token of support examples (Snell, Swersky, and Zemel 2017), while the span-level methods bypass the token-wise label dependency issue by measuring the distance between spans.

In the first family, Fang et al. (2023) designed an additional memory module that stored token representations of entity types to adaptively learn cross-domain entities. In the second family, Ma et al. (2022) introduced a decomposed meta-learning approach to decompose the few-shot NER task into span detection and entity typing, which enabled the few-shot NER model to learn suitable initial parameters and embedding space.

However, the existing few-shot NER methods suffer a significant drop in recognition performance for unseen entity types when handling adversarial examples. It demonstrates the vulnerability of the cross-domain transfer learning abilities of present methods under textual adversarial attack scenarios.

Adversarial Robustness on Texts

Recently, a magnitude of adversarial attacks have been introduced for texts (Zhou et al. 2019; Liu et al. 2022; Zeng et al. 2023), such as synonym substitution (Lin et al. 2021), adversarial perturbation generating (Wang et al. 2021; Zhu et al. 2020), which can maintain human understanding of sentences but "fool" the deep neural networks. Textual adversarial attacks make deep neural networks output incorrect predictions and point out the vulnerability of current models regarding textual adversarial attacks. Aiming at verifying the robustness of the NER models, Lin et al. (2021) generated entities substitutions using the other entities of the same semantics in Wikidata, and utilized pre-trained language models (Devlin et al. 2019) to replace the words in context. Zeng et al. (2023) generated a set of copies for input texts by randomly masking words to improve robustness.

Current adversarial robustness tasks on text mainly focus on learning ability in the same domain, such as text classification (Li et al. 2020; Liu et al. 2023), sentiment analysis (Li et al. 2021) and named entity recognition (Lin et al. 2021). However, there are few works about the robustness of cross-domain transfer learning ability for texts. In this work, we focus on the effect of textual adversarial attacks on the cross-domain transfer learning ability in few-shot NER task.

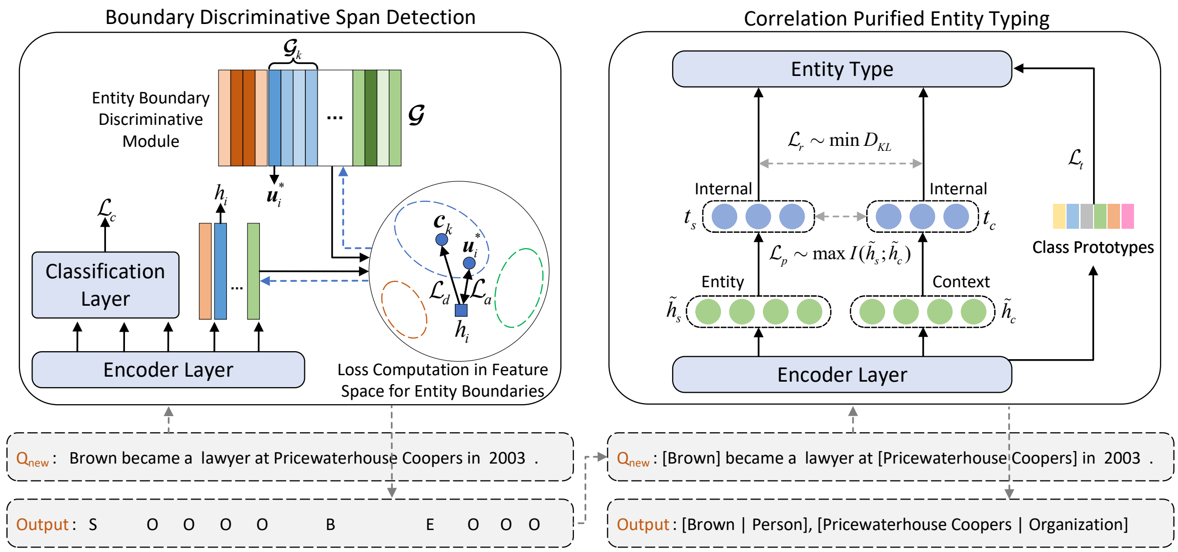


Figure 2: Overview of the robust two-stage few-shot NER method with Boundary Discrimination and Correlation Purification (BDCP). In the span detection stage, we introduce entity boundary discriminative module $\mathcal{G}$ to provide a highly distinguishing boundary representation space. Each token representation h_i is diversely assigned to the closest component $\boldsymbol{u}_i^*$ through two mutually complementary assignment loss $\mathcal{L}_a$ and diversity loss $\mathcal{L}_d$. The blue dashed lines represent backpropagation. In the entity typing stage, the correlations between entities and contexts are purified to alleviate the perturbations caused by textual adversarial attacks. Correlation purification is implemented by minimizing interference information in correlations (i.e. $\mathcal{L}_r$) and facilitating correlation generalization (i.e. $\mathcal{L}_p$).

Preliminaries

Few-Shot Named Entity Recognition. Given a sequence $\boldsymbol{x} = \{x_i\}_{i=1}^n$ containing n tokens, named entity recognition (NER) task aims to output the entity sequences $\boldsymbol{e} = \{e_j\}_{j\geq 0}$ and assign the corresponding entity type y_j to each entity sequence e_j. The few-shot NER model is trained in a source domain $\varepsilon_{source} = \{(\mathcal{S}_s, \mathcal{Q}_s, \mathcal{Y}_s)\}$, where $\mathcal{S}_s, \mathcal{Q}_s, \mathcal{Y}_s$ represent the support set, query set, and the entity types in the training data, respectively. Then the few-shot NER model is transferred to a data-scarce target domain $\varepsilon_{target} = \{(\mathcal{S}_t, \mathcal{Q}_t, \mathcal{Y}_t)\}$ with the similar data construction for testing. A few-shot NER model learned in the source domain ε_{source} is expected to leverage the support set $\mathcal{S}_t$ of the target domain to predict novel entities in the query set $\mathcal{Q}_t$. Since $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$, the few-shot NER model needs to learn cross-domain transfer knowledge to be generalized to unseen entity types in few-shot NER. In the N-way K-shot setting, there are N entity types in the target domain (i.e. $|\mathcal{Y}_t| = N$), and each entity type is associated with K examples in the support set $\mathcal{S}_t$.

Adversarial Attack on Few-Shot NER. The entities and corresponding entity types recognized by the few-shot NER model in the target domain are denoted as the set $\{C_i\}$, where $C_i = (e_p, y_p)$, e_p and y_p are the entity and corresponding entity type. The adversarial attack on few-shot NER aims to construct adversarial examples to "fool" the neural network based few-shot NER models. An adversarial example $\boldsymbol{x'}$ makes the few-shot NER model that correctly recognizes entities and corresponding entity types to predict incorrect entity-type pairs $\{C_i'\}$, i.e.

$$\{C_i\} \neq \{C_i'\}. \tag{1}$$

In this work, the adversarial example $\boldsymbol{x'}$ is constructed by replacing the original tokens with synonyms w_i' in the synonym set.

Information Bottleneck (IB) principle utilizes the idea of mutual information to analyze the training and inference of deep neural networks (Shwartz-Ziv and Tishby 2017; Tishby and Zaslavsky 2015). Given the input data X and the label Y, it attempts to learn an internal representation T that makes an information trade-off between predictive accuracy and representation compression:

$$\mathcal{L}_{IB} = -I(T;Y) + \beta * I(T;X), \tag{2}$$

where I stands for mutual information (MI), aiming to measure the interdependence between two variables. β is the Lagrange multiplier that controls the trade-off between two MI terms. By optimizing the loss $\mathcal{L}_{IB}$, the IB principle compresses the noise data in X while retaining enough features in X to predict Y.

Methodology

Overview of the Proposed Method

Technically, we found that the decline in cross-domain transfer learning ability for few-shot NER is reflected in: 1) entity span detection errors; 2) inaccurate unseen entity type classification. To solve the above problems, we propose a robust two-stage few-shot NER method with Boundary Discrimination and Correlation Purification (BDCP). Figure 2 illustrates the architecture of the proposed BDCP method. Following Ma et al. (2022), the few-shot NER task is decomposed into span detection and entity typing. Overall, the span detection stage is designed to detect the entity spans in the input text, then those entity spans are classified into corresponding unseen entity types in the entity typing stage.

Different from Ma et al. (2022), in the span detection stage, for the purpose of improving the adversarial robustness of entity span detection, entity boundary discriminative module is introduced to provide a highly discriminative boundary representation space. In the entity typing stage, aiming at improving the adversarial robustness of entity typing, the correlations between entities and contexts are purified by minimizing the interference information and facilitating correlation generalization. Note that, the span detection stage and the entity typing stage are learned serially. The proposed BDCP method does not introduce additional computational cost like adversarial training and data augmentation.

Adversarial Examples Generating

Textual adversarial attack is conducted on the examples in the support set $\mathcal{S}_t$ and the query set $\mathcal{Q}_t$ of the target domain. The textual adversarial attack algorithm BERT-Attack (Li et al. 2020) is used to perform synonym substitution and generate adversarial examples[2]. The cross-domain transfer learning ability is reflected through the performance of the few-shot NER model in the query set $\mathcal{Q}_t$ of the target domain. Textual adversarial attacks in the real world usually exist randomly in textual data, hence we generate adversarial examples against the entire original examples without distinguishing entities and contexts (Lin et al. 2021).

Boundary Discriminative Span Detection

Span detection is regarded as a token-level label classification process. The entity boundary discriminative module $\boldsymbol{\mathcal{G}}$ is introduced to alleviate the problem of entity boundary detection errors, which provides an additional robust representation space for entity span boundaries. Entity boundary discriminative module $\boldsymbol{\mathcal{G}}$ contains N_b components blocks $\boldsymbol{\mathcal{G}}_k$, where N_b denotes the number of entity boundary classes. $\boldsymbol{\mathcal{G}}_k = \{u_i\}_{i=1}^{N_c}$ is the set of components $u_i \in \mathbb{R}^C$ assigned to boundary class k, where N_c represents the number of components assigned to each boundary class.

Boundary Assignment. First, the adversarial example $\boldsymbol{x'} = \{w_i\}_{i=1}^{n}$ is input into an encoder f_θ to generate token representations $\boldsymbol{h} = \{h_i\}_{i=1}^{n}$:

$$\boldsymbol{h} = f_\theta(\boldsymbol{x'}). \tag{3}$$

Each token representation h_i is matched amongst the components in each block $\boldsymbol{\mathcal{G}}_k$ according to the cosine similarity, thus obtaining the closest component $\boldsymbol{u}_i^*$ corresponding to h_i:

$$\boldsymbol{u}_i^* = \underset{u_j \in \boldsymbol{\mathcal{G}}_k}{\arg\max} \frac{h_i \boldsymbol{.} u_j}{||h_i||\,||u_j||}, \tag{4}$$

where $\boldsymbol{.}$ represents dot product. Inspired by the mixture based feature space (Afrasiyabi, Lalonde, and Gagné 2021), we improve the angular margin-based softmax function (Deng et al. 2019) with a temperature variable τ:

$$p_\theta(v_j|h_i, \boldsymbol{\mathcal{G}}) = \frac{e^{\cos(\angle(h_i,u_j)+m)/\tau}}{e^{\cos(\angle(h_i,u_j)+m)/\tau} + \sum\limits_{u_l \in \{\boldsymbol{\mathcal{G}} \setminus u_j\}} e^{\cos(\angle(h_i,u_l))/\tau}}, \tag{5}$$

where $\angle(h_i, u_j) = arccos(h_i^\top u_i/(||h_i||||u_j||))$, v_j denotes the pseudo-label associated to u_j, and m represents a margin.

Next, every token representation h_i is assigned to the closest component $\boldsymbol{u}_i^*$ in $\boldsymbol{\mathcal{G}}$ utilizing the assignment loss $\mathcal{L}_a$:

$$\mathcal{L}_a = -\frac{1}{L}\sum_{i=1}^{L} \log p_\theta(v_i^*|h_i, \boldsymbol{\mathcal{G}}), \tag{6}$$

where L represents the number of all tokens in a batch, and v_i^* denotes the one-hot pseudo-label corresponding to u_i^*.

Diverse Assignment. Each entity boundary class is supposed to be mapped to multiple components rather than a single one in $\boldsymbol{\mathcal{G}}_k$, which can improve the adversarial robustness. Because in this way, more generalized boundary representation space can be leveraged to detect the entity boundaries. Training the backbone model and the entity boundary augmentation matrix only on the assignment loss $\mathcal{L}_a$ usually results in that the tokens of the entity boundary class k are assigned to a single component $u_i \in \boldsymbol{\mathcal{G}}_k$. To avoid this, the diversity loss $\mathcal{L}_d$ is designed to facilitate the diversity of component selection for every entity boundary class.

The diversity loss $\mathcal{L}_d$ pushes the token representation h_i towards the centroid of components associated with its entity boundary class. The centroid c_k for entity boundary class k is defined as:

$$\boldsymbol{c}_k = (1/|\mathcal{G}_k|) \sum\nolimits_{u_j \in \mathcal{G}_k} u_j, \tag{7}$$

where $|\mathcal{G}_k|$ denotes the number of components for each entity boundary labels y_i. For the centroids set $\boldsymbol{C} = \{c_k\}_{k=1}^{L_e}$, the diversity loss $\mathcal{L}_d$ is calculated as:

$$\mathcal{L}_d = -\frac{1}{L}\sum_{i=1}^{L} \log p_\theta(y_i|h_i, sg[\boldsymbol{C}]), \tag{8}$$

where sg denotes stopgradient that protects specific variables from backpropagation. It prevents the components of each boundary class from collapsing into a single point.

Following Ma et al. (2022), we also use averaged cross-entropy loss with a maximum term as training loss:

$$\begin{aligned}\mathcal{L}_c = &\frac{1}{L}\sum_{i=1}^{L} CrossEntropy(y_i, p(w_i)) \\ &+ \alpha \max_{i \in \{1,2,\ldots,L\}} CrossEntropy(y_i, p(w_i)),\end{aligned} \tag{9}$$

where the maximum term is leveraged to alleviate insufficient learning for tokens with higher loss. p represents the

[2]More details about the attack algorithm and generation implementation are listed in Appendix, which can be available at http://arxiv.org/abs/2312.07961

fully connected layer with $softmax$ activation function. α is the weighting factor.

Final Objective. Overall, the training loss in boundary discriminative span detection stage is the combination of Eq. (6), Eq. (8) and Eq. (9):

$$\mathcal{L}_{sp} = \mathcal{L}_c + \gamma_1 \mathcal{L}_a + \gamma_2 \mathcal{L}_d, \tag{10}$$

where γ_1 and γ_2 are weighting factors for assignment loss and diversity loss, respectively.

Correlation Purified Entity Typing

In the entity typing stage, the entity spans obtained in the span detection stage are classified into corresponding entity types. In this paper, the correlations refer to the correlations between entities and contexts for simplicity.

Textual adversarial attacks bring interferences to the correlations, which has an adverse effect on cross-domain transfer learning. Therefore, we design correlation purified entity typing to filter the interference information existing in the correlations and facilitate the correlation generalization, which can alleviate the perturbations caused by textual adversarial attacks. Inspired by Tishby and Zaslavsky (2015) and Wang et al. (2022b), we explicitly minimize the interference information and maximize the interactive information between entities and contexts.

For entity typing, the adversarial example x' is input into another encoder g_θ to generate token representations $\tilde{\boldsymbol{h}} = \{\tilde{h}_i\}_{i=1}^n$:

$$\tilde{\boldsymbol{h}} = g_\theta(x'). \tag{11}$$

The positions of the entity spans can be obtained from the span detection stage. Subsequently, the entity span representation $\tilde{h}_s$ is computed by averaging all token representations inside the entity span. The context representation $\tilde{h}_c$ is computed by averaging the rest of the token representations in x' other than the entity spans.

Intuitively, for few-shot NER in the textual adversarial scenarios, the correlations between entities and contexts are essential to predict unseen entities, and the interferences in adversarial examples are interfering features. Motivated by the chain rule of mutual information and the contrastive strategy (Federici et al. 2020; Wang et al. 2022b), the mutual information $I(x; \tilde{h}_s)$ between entity span $\tilde{h}_s$ and original example x can be decomposed into two parts:

$$I(x; \tilde{h}_s) = \underbrace{I(\tilde{h}_s; x')}_{predictable} + \underbrace{I(x; \tilde{h}_s|x')}_{specific}, \tag{12}$$

where x' denotes to the adversarial example. $I(\tilde{h}_s; x')$ refers to the information in $\tilde{h}_s$ that is predictable for x', i.e. non-entity-specific information. $I(x; \tilde{h}_s|x')$ indicates the information in $\tilde{h}_s$ that is unique to x but is unpredictable for x', i.e. entity-specific information.

Consequently, the entity-specific information is interfering, and any token representation h that encompasses the information jointly shared by x and x' would also contain the requisite label information. Then Eq. (12) can be approximated as follows:

$$maximize\ I(\tilde{h}_s; y) \sim I(\tilde{h}_s; x'), \tag{13}$$

$$minimize\ I(x; \tilde{h}_s|y) \sim I(x; \tilde{h}_s|x'), \tag{14}$$

Correlation Facilitating. To facilitate correlation generalization between entities and contexts, Eq. (13) is utilized to maximize the generalization information of textual representations. Similar to Wang et al. (2022b), it has been proved that $I(\tilde{h}_s, \tilde{h}_c)$ is a lower bound of $I(\tilde{h}_s, x')$ (proof is detailed in Appendix). InfoNCE (van den Oord, Li, and Vinyals 2018) can be leveraged to approximate $I(\tilde{h}_s, \tilde{h}_c)$, which is a lower bound of mutual information. It can be known from previous work (van den Oord, Li, and Vinyals 2018) that when the number of tokens is lager, the lower bound is closer to InfoNCE. During training, minimizing the InfoNCE loss can maximize the lower bound of mutual information. Thus, the goal of facilitating correlation generalization is optimized by:

$$\mathcal{L}_p = -\mathbb{E}_1 \left[g_p(\tilde{h}_s, \tilde{h}_c) - \mathbb{E}_2 (\log \sum_{\tilde{h}_i \in \tilde{h}_c} \exp(g_p(\tilde{h}_s, \tilde{h}_c))) \right], \tag{15}$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are two different activation functions. g_p is the compatibility scoring function implemented by a linear neural network.

Correlation Purification. Aiming at alleviating the adverse effects of interferences in adversarial examples, Eq. (14) is exploited to minimize the interference information. To this end, we minimize an upper bound of $I(x; \tilde{h}_s|x')$ (proof is detailed in Appendix), which is formulated as:

$$p_{ib}(t|h) = \mathcal{N}(t|f_{ib}^{\mu}(h), f_{ib}^{M}(h)) \tag{16}$$

$$\mathcal{L}_r = D_{KL}[p_{ib}(t_s|\tilde{h}_s)||p_{ib}(t_c|\tilde{h}_c))] \tag{17}$$

where p_{ib} denotes the information bottleneck layer, t refers to internal representation, f_{ib}^{μ} and f_{ib}^{v} are multilayer perceptrons (MLP) that compute the mean μ and the covariance matrix M of t, respectively, and $\mathcal{N}$ stands for the reparameterization trick (Kingma and Welling 2014). D_{KL} refers to Kullback-Leibler divergence.

Final Objective. We utilize ProtoNet (Snell, Swersky, and Zemel 2017) to calculate class prototypes and use cross-entropy loss to compute prototype loss $\mathcal{L}_t$, same as Ma et al. (2022). Due to space limitation, the calculation processes are listed in Appendix. Finally, the training loss in correlation purified entity typing stage is the combination of $\mathcal{L}_t$, $\mathcal{L}_p$ and $\mathcal{L}_r$:

$$\mathcal{L}_{sp} = \mathcal{L}_t + \gamma_3 \mathcal{L}_p + \gamma_4 \mathcal{L}_r, \tag{18}$$

where γ_3 and γ_4 are weighting factors for the correlation facilitation loss and interference information filtering loss. In addition, we adopt the same meta-learning strategy as Ma et al. (2022) to train and evaluate the model in two stages.

Experiments

Experimental Settings

Datasets. Comprehensive experiments are conducted on two groups of datasets to evaluate the adversarial robustness of the proposed method: (1) **Few-NERD** (Ding et al. 2021) contains 8 coarse-grained and 66 fine-grained entity types

Models	INTRA				INTER			
	5-1		10-5		5-1		10-5	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
ProtoBERT (Fritzler et al. 2019)	23.45	12.45	34.61	29.08	44.44	38.30	53.97	45.23
StructShot (Yang et al. 2020)	35.92	21.62	26.39	19.71	57.33	42.78	49.39	42.75
CONTAINER (Das et al. 2022)	40.43	24.58	47.49	40.15	55.95	41.45	57.12	48.19
Decomposed (Ma et al. 2022)	52.04	33.73	56.84	45.16	68.77	49.36	68.32	54.68
RockNER (Lin et al. 2021)	52.17	35.29	57.25	46.50	68.92	51.54	68.43	57.64
RanMASK (Zeng et al. 2023)	52.33	36.16	56.79	44.18	67.58	52.37	68.61	58.86
Our BDCP	**52.63**	**40.76**	**57.46**	**50.71**	**69.59**	**56.84**	**68.87**	**64.97**

Table 1: Performance on Few-NERD.

Models	5-shot							
	News		Wiki		Social		Mixed	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
MatchingNet (Vinyals et al. 2016)	19.85	13.95	5.58	4.52	6.61	5.13	8.08	6.58
SimBERT (Hou et al. 2020)	32.01	25.56	10.63	8.27	8.20	6.74	21.14	14.62
L-TapNet+CDT (Hou et al. 2020)	45.35	31.23	11.65	9.69	23.30	15.38	20.95	13.44
Decomposed (Ma et al. 2022)	58.18	42.87	31.36	23.39	31.02	22.46	45.55	34.19
RockNER (Lin et al. 2021)	58.32	44.51	31.63	24.45	31.52	24.18	45.78	36.48
RanMASK (Zeng et al. 2023)	58.54	45.19	32.06	22.64	31.67	24.69	44.35	37.51
Our BDCP	**58.76**	**51.36**	**32.17**	**27.91**	**31.84**	**27.29**	**46.29**	**40.73**

Table 2: Performance on Cross-Dataset.

with hierarchical structure for few-shot NER[3], addressing two tasks named INTRA and INTER. Specifically, entities in the train/dev/test splits belong to different coarse-grained entity types in INTRA, while fine-grained entity types are mutually disjoint and the coarse-grained entity types are shared in INTER. (2) four sub-datasets from four domains are used in **Cross-Dataset** (Hou et al. 2020): News (CoNLL-2003) (Sang 2002), Wiki (GUM) (Zeldes 2017), Social (WNUT-2017) (Derczynski et al. 2017), Mixed (Ontonotes) (Pradhan et al. 2013)[4].

For the Few-NERD dataset, the episodes in Ma et al. (2022) are used, where each episode contains one N-way K-shot few-shot NER task. Totally, 20,000/1,000/5,000 episodes are employed for training/validation/testing. Accordingly, 5-way 1-shot and 10-way 5-shot setups are utilized for INTRA and INTER. For the Cross-Dataset, two sub-datasets are used to construct training episodes. One sub-dataset and one of the remaining sub-datasets are utilized to construct validation episodes and test episodes, respectively. In the 5-shot setup on Cross-Dataset, 200/100/100 episodes are used for training/validation/testing, as in Hou et al. (2020). The sub-datasets for training, validation and test are mutually disjoint from each other, so that all the models can be tested in cross-domain scenarios.

Baselines. The typical metric-learning based baselines are considered first. For Few-NERD, the proposed method is compared with typical few-shot NER methods ProtoBERT (Fritzler, Logacheva, and Kretov 2019), StructShot (Yang and Katiyar 2020), CONTAINER (Das et al. 2022) and Decomposed (Ma et al. 2022). For Cross-Dataset, the proposed method is compared with SimBERT (Hou et al. 2020), L-TapNet+CDT (Hou et al. 2020) and Decomposed (Ma et al. 2022). Second, since there are currently few robust few-shot NER methods, the direct adaptations of the following textual defense methods are also compared with our BDCP model: RockNER (Lin et al. 2021), RanMASK (Zeng et al. 2023) for both Few-NERD and Cross-Dataset. More details about the baselines are listed in Appendix. We report the F1 scores of the models handling clean and adversarial examples.

Implementation Details. Two separate BERT-base-uncased models (Devlin et al. 2019) are used to implement encoders f_θ and g_θ. The batch size, dropout probability and maximum sequence length are set to 64, 0.2, 128, respectively. The model is trained using the AdamW optimizer (Loshchilov and Hutter 2019), and the initial learning rates for two stages are set to 3e-5 and 1e-4, respectively. The temperature variable τ and the margin m are 0.025 and 0.01. The number of components for each entity boundary class in $\mathcal{G}$ is 15. The weighting factors γ_1, γ_2, γ_3 and γ_4 are set to 0.1, 0.1, 1e-3,1e-5, respectively. For clean examples, the results of typical few-shot NER models are reported from Ma et al. (2022), and we reproduce the rest of the models to report results. For adversarial examples, we reproduce all the baseline models to obtain results.

[3]https://github.com/thunlp/Few-NERD

[4]https://github.com/AtmaHou/FewShotTagging

Experimental Results

Performance Comparison. Table 1 and Table 2 report the overall comparison results of different models on Few-NERD and Cross-Dataset datasets, under both Clean and Attack scenarios. Here we can observe that:

1) All baseline models suffer from a significant drop in performance under textual adversarial attack, indicating the vulnerability of the cross-domain transfer learning ability for existing few-shot NER methods. For example, the typical Decomposed model (Ma et al. 2022) dramatically drops an average of 15.76 and 10.80 points on the Few-NERD and Cross-Dataset datasets, respectively. The reason lies in that those models learn the preferences and usage patterns for specific words during training, and are fooled by the interferences brought by textual adversarial attacks.
2) RockNER and RanMASK improve robustness through data augmentation and random masking respectively. Compared with the typical Decomposed model, the performance of RockNER is slightly improved, which shows that increasing data scale is conducive to improving the robustness of few-shot NER. However, random masking may reduce the useful information where there is a lack of labeled samples. Hence RanMASK cannot improve the cross-domain transfer learning ability in that case (e.g. 10-way 5-shot setting in INTRA).
3) By utilizing entity boundary discrimination and correlation purification, the proposed BDCP model outperforms all the baseline methods in both Clean and Attack scenarios. Under textual adversarial attack scenario, compared with the typical Decomposed model, the performance of our BDCP model is improved by 7.59% and 6.10% in terms of F1 on the Few-NERD and Cross-Dataset datasets, respectively. The advantages of the BDCP model in adversarial robustness lie in that: (a) the boundary discrimination module improves the robustness of span detection by providing highly distinguishing and diverse entity boundary representation space; (b) the correlation purification module mitigates the interference of attacks by restraining unpredictable information and the KL divergence of the internal representations, entities and contexts.

Models	INTRA		INTER	
	5-1	10-5	5-1	10-5
Base	33.73	45.16	49.36	54.68
+ assignment	34.86	45.87	51.03	55.67
+ components	36.25	47.24	52.58	57.26
+ facilitating	33.81	46.20	50.95	55.53
+ filter	35.38	47.66	53.42	58.70
+ purify	37.41	48.12	54.34	60.19
BDCP	**40.76**	**50.71**	**56.84**	**64.97**

Table 3: Ablation study results on Few-NERD dataset.

Ablation Study. We design ablation studies to analyze the effects of different modules on adversarial robustness in the BDCP method. Table 3 shows the results in different settings, taking the Few-NERD dataset containing adversarial examples as an example. "Base" model refers to Decomposed model (Ma et al. 2022). "+components" indicates that both assignment loss and diversity loss are added. "+purify" means simultaneously using correlation facilitation loss and interference information filtering loss. It can be observed that the "+components" model outperforms the "+assignment" model, indicating that diverse assignment for boundary discrimination can effectively improve the robustness of entity span detection. Moreover, for correlation purification, adding both correlation facilitation loss and interference information filtering loss to the base model performs better than adding any single module, which demonstrates that those two modules are beneficial to improve the adversarial robustness and can boost each other.

Furthermore, the adversarial robustness is significantly improved after adding all the above modules, that is, our BDCP model performs best. It illustrates that when handling adversarial examples in few-shot NER, our BDCP model can prevent the cross-domain transfer learning from rote memorizing the surface features of entities and exploiting preferences in the data.

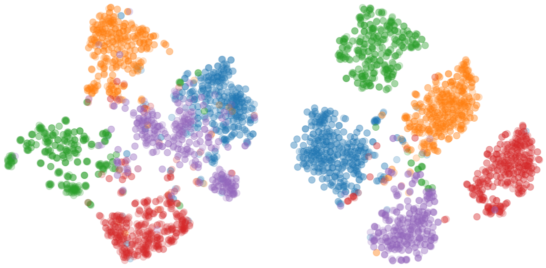
Visualization

In order to intuitively illustrate the effectiveness of boundary discrimination and correlation purification, t-SNE (Van der Maaten and Hinton 2008) is used to reduce the dimensionality of the span representations from the BDCP model and the Decomposed model, and visualize the span representations of different entity types, as shown in Figure 3.

It can be observed that when handling textual adversarial examples in few-shot NER, the span representations of different entity types generated by the Decomposed model are less discriminative than the BDCP model. Additionally, there are more outlier span representations for entity typing. In contrast, the BDCP model can better disperse the span representations of different entity types, and has a clearer decision boundary than Decomposed model. Therefore, the BDCP model can provide textual representation space with better adversarial robustness and generalization for few-shot NER through boundary discrimination and correlation purification, under textual adversarial attack scenario.

Conclusion

In this paper, we have explored and evaluated the adversarial robustness of cross-domain transfer learning ability in few-shot NER for the first time. Extensive experiments indicate that existing few-shot NER methods are vulnerable when handling textual adversarial examples. To address this issue, we propose a robust few-Shot NER method with boundary discrimination and correlation purification. Specifically, in the span detection stage, the entity boundary discriminative module is introduced to diversify the assignment of token representations to corresponding closest components, thus providing a highly discriminative boundary representation



(a) Decomposed (b) Our BDCP

Figure 3: t-SNE Visualization of span representations on Few-NERD INTER 5-way 1-shot query set that containing adversarial examples. The representations are obtained from Decomposed (Ma et al. 2022) model and our BDCP model, respectively. Different colors represent different entity types.

space. In the entity typing stage, correlation purification is performed by minimizing interference information and facilitating correlation generalization to alleviate the perturbations caused by textual adversarial attacks. Experiments on various few-shot NER datasets containing adversarial examples demonstrate the adversarial robustness of the BDCP model. In future work, we will explore the robustness of adversarial training in few-shot NER.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (62072039). We thank the anonymous reviewers for their valuable comments and feedback.

References

Afrasiyabi, A.; Lalonde, J.; and Gagné, C. 2021. Mixture-based Feature Space Learning for Few-shot Image Classification. In *ICCV*, 9021–9031. IEEE.

Agarwal, O.; Yang, Y.; Wallace, B. C.; and Nenkova, A. 2021. Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve. *Comput. Linguistics*, 47(1): 117–140.

Das, S. S. S.; Katiyar, A.; Passonneau, R. J.; and Zhang, R. 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. In *ACL*, 6338–6353. Association for Computational Linguistics.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 4690–4699. Computer Vision Foundation / IEEE.

Derczynski, L.; Nichols, E.; van Erp, M.; and Limsopatham, N. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *NUT@EMNLP*, 140–147. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.; and Liu, Z. 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. In *ACL/IJCNLP*, 3198–3213. Association for Computational Linguistics.

Fang, J.; Wang, X.; Meng, Z.; Xie, P.; Huang, F.; and Jiang, Y. 2023. MANNER: A Variational Memory-Augmented Model for Cross Domain Few-Shot Named Entity Recognition. In *ACL*, 4261–4276. Association for Computational Linguistics.

Federici, M.; Dutta, A.; Forré, P.; Kushman, N.; and Akata, Z. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*. OpenReview.net.

Fritzler, A.; Logacheva, V.; and Kretov, M. 2019. Few-shot classification in named entity recognition task. In *SAC*, 993–1000. ACM.

Hou, Y.; Che, W.; Lai, Y.; Zhou, Z.; Liu, Y.; Liu, H.; and Liu, T. 2020. Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network. In *ACL*, 1381–1393. Association for Computational Linguistics.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *NeurIPS*, 125–136.

Kato, T.; Abe, K.; Ouchi, H.; Miyawaki, S.; Suzuki, J.; and Inui, K. 2020. Embeddings of Label Components for Sequence Labeling: A Case Study of Fine-grained Named Entity Recognition. In *ACL (student)*, 222–229. Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2013. One-shot learning by inverting a compositional causal process. In *NIPS*, 2526–2534.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *HLT-NAACL*, 260–270. The Association for Computational Linguistics.

Li, J.; Chiu, B.; Feng, S.; and Wang, H. 2022a. Few-Shot Named Entity Recognition via Meta-Learning. *IEEE Trans. Knowl. Data Eng.*, 34(9): 4245–4256.

Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; and Li, F. 2022b. Unified Named Entity Recognition as Word-Word Relation Classification. In *AAAI*, 10965–10973. AAAI Press.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *EMNLP*, 6193–6202. Association for Computational Linguistics.

Li, Z.; Xu, J.; Zeng, J.; Li, L.; Zheng, X.; Zhang, Q.; Chang, K.; and Hsieh, C. 2021. Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution. In *EMNLP*, 3137–3147. Association for Computational Linguistics.

Lin, B. Y.; Gao, W.; Yan, J.; Moreno, R.; and Ren, X. 2021. RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. In *EMNLP*, 3728–3737. Association for Computational Linguistics.

Liu, H.; Xu, Z.; Zhang, X.; Xu, X.; Zhang, F.; Ma, F.; Chen, H.; Yu, H.; and Zhang, X. 2023. SSPAttack: A Simple and Sweet Paradigm for Black-Box Hard-Label Textual Adversarial Attack. In *AAAI*, 13228–13235. AAAI Press.

Liu, Q.; Zheng, R.; Rong, B.; Liu, J.; Liu, Z.; Cheng, Z.; Qiao, L.; Gui, T.; Zhang, Q.; and Huang, X. 2022. Flooding-X: Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning. In *ACL*, 5634–5644. Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, J.; Jin, S.; Liang, J.; and Zhang, C. 2021. Robust Few-Shot Learning for User-Provided Data. *IEEE Trans. Neural Networks Learn. Syst.*, 32(4): 1433–1447.

Ma, R.; Lin, Z.; Chen, X.; Zhou, X.; Wang, J.; Gui, T.; Zhang, Q.; Gao, X.; and Chen, Y. W. 2023. Coarse-to-fine Few-shot Learning for Named Entity Recognition. In *ACL (Findings)*, 4115–4129. Association for Computational Linguistics.

Ma, T.; Jiang, H.; Wu, Q.; Zhao, T.; and Lin, C. 2022. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. In *ACL (Findings)*, 1584–1596. Association for Computational Linguistics.

Piran, Z.; Shwartz-Ziv, R.; and Tishby, N. 2020. The Dual Information Bottleneck. *CoRR*, abs/2006.04641.

Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *CoNLL*, 143–152. ACL.

Sang, E. F. T. K. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL*. ACL.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. *CoRR*, abs/1703.00810.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *NIPS*, 4077–4087.

Tishby, N.; and Slonim, N. 2000. Data Clustering by Markovian Relaxation and the Information Bottleneck Method. In *NIPS*, 640–646. MIT Press.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *ITW*, 1–5. IEEE.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *NIPS*, 3630–3638.

Wang, P.; Xu, R.; Liu, T.; Zhou, Q.; Cao, Y.; Chang, B.; and Sui, Z. 2022a. An Enhanced Span-based Decomposition Method for Few-Shot Sequence Labeling. In *NAACL-HLT*, 5012–5024. Association for Computational Linguistics.

Wang, Q.; Boudreau, C.; Luo, Q.; Tan, P.; and Zhou, J. 2019. Deep Multi-view Information Bottleneck. In *SDM*, 37–45. SIAM.

Wang, X.; Dou, S.; Xiong, L.; Zou, Y.; Zhang, Q.; Gui, T.; Qiao, L.; Cheng, Z.; and Huang, X. 2022b. MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective. In *ACL*, 5590–5600. Association for Computational Linguistics.

Wang, X.; Yang, Y.; Deng, Y.; and He, K. 2021. Adversarial Training with Fast Gradient Projection Method against Synonym Substitution Based Text Attacks. In *AAAI*, 13997–14005. AAAI Press.

Yang, Y.; and Katiyar, A. 2020. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. In *EMNLP*, 6365–6375. Association for Computational Linguistics.

Yoon, S. W.; Seo, J.; and Moon, J. 2019. TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7115–7123. PMLR.

Yu, D.; He, L.; Zhang, Y.; Du, X.; Pasupat, P.; and Li, Q. 2021. Few-shot Intent Classification and Slot Filling with Retrieved Examples. In *NAACL-HLT*, 734–749. Association for Computational Linguistics.

Zeldes, A. 2017. The GUM corpus: creating multilayer resources in the classroom. *Lang. Resour. Evaluation*, 51(3): 581–612.

Zeng, J.; Xu, J.; Zheng, X.; and Huang, X. 2023. Certified Robustness to Text Adversarial Attacks by Randomized [MASK]. *Comput. Linguistics*, 49(2): 395–427.

Zhou, Y.; Jiang, J.; Chang, K.; and Wang, W. 2019. Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification. In *EMNLP/IJCNLP*, 4903–4912. Association for Computational Linguistics.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *ICLR*. OpenReview.net.