Style Transfer as Data Augmentation: A Case Study on Named Entity Recognition

Shuguang Chen
University of Houston
schen52@uh.edu

Leonardo Neves
Snap Inc.
lneves@snap.com

Thamar Solorio
University of Houston
tsolorio@uh.edu

Abstract

In this work, we take the named entity recognition task in the English language as a case study and explore style transfer as a data augmentation method to increase the size and diversity of training data in low-resource scenarios. We propose a new method to effectively transform the text from a high-resource domain to a low-resource domain by changing its style-related attributes to generate synthetic data for training. Moreover, we design a constrained decoding algorithm along with a set of key ingredients for data selection to guarantee the generation of valid and coherent data. Experiments and analysis on five different domain pairs under different data regimes demonstrate that our approach can significantly improve results compared to current state-of-the-art data augmentation methods. Our approach is a practical solution to data scarcity, and we expect it to be applicable to other NLP tasks. [1]

1 Introduction

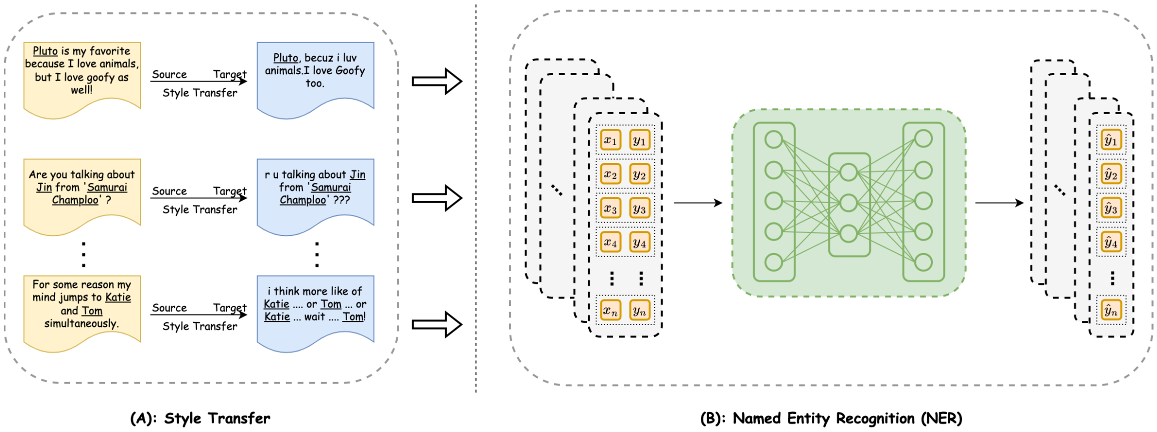
Large-scale pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) have shown impressive performances on a wide variety of NLP tasks. Following the paradigm of self-supervised pre-training and fine-tuning, these models have achieved state-of-the-art performance in many NLP benchmarks such as question answering (Yang et al., 2019; Yamada et al., 2020), machine translation (Provilkov et al., 2020; Lewis et al., 2020), and text summarization (Zaheer et al., 2020; Aghajanyan et al., 2021). However, due to the discrepancy of training objectives between language modeling and downstream tasks, the performance of such models may be limited by data scarcity in low-resource domains (Jiang et al., 2020; Gururangan et al., 2020).

Data augmentation is effective in addressing data scarcity. Previous work (Wei and Zou, 2019; Morris et al., 2020) has mainly focused on using in-domain data to synthesize new datasets for training. When applied to low-resource domains, however, these approaches may not lead to significant improvement gains as the data in low-resource domains is not comparable to that in high-resource domains in terms of size and diversity. Recently, many studies (Xia et al., 2019; Dehouck and Gómez-Rodríguez, 2020) have revealed the potential of leveraging data from high-resource domains to improve low-resource tasks. Despite their impressive results, directly using abundant data from high-resource domains can be problematic due to the difference in data distribution (e.g., language shift) and feature misalignment (e.g., class mismatch) between domains (Wang et al., 2018; Zhang et al., 2021).

In this work, we explore the potential of employing style transfer as a way of data augmentation in cross-domain settings. Style transfer on natural language aims to change the style-related attributes of text while preserving its semantics, which makes it a reasonable alternative for the purpose of data augmentation. Here, we take the named entity recognition (NER) task as a case study to investigate its effectiveness. The general pipeline is shown in Figure 1. Compared to the text classification task, the NER task is more difficult as it requires to jointly modify the tokens and their corresponding labels. One of the critical challenges here is the lack of parallel style transfer data annotated with NER labels. Our workaround to the lack of data is to figure out how to leverage a non-NER parallel style transfer dataset and a nonparallel NER dataset. Both datasets contain pairs of sentences in source and target styles, respectively. This scenario is much more realistic since resources for style transfer tend to be task-agnostic. At the same time, it is easier to come

[1] We released the code at https://github.com/RiTUAL-UH/DA_NER.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1827–1841
December 7-11, 2022 ©2022 Association for Computational Linguistics


Pluto is my favorite because I love animals, but I love goofy as well!
Source
Target
Style Transfer
Pluto, becuz i luv animals.I love Goofy too.
Are you talking about Jin from 'Samurai Champloo' ?
Source
Target
Style Transfer
r u talking about Jin from 'Samurai Champloo' ???
For some reason my mind jumps to Katie and Tom simultaneously.
Source
Target
Style Transfer
i think more like of Katie .... or Tom ... or Katie ... wait .... Tom!
x_1 y_1
x_2 y_2
x_3 y_3
x_4 y_4
x_n y_n
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\hat{y}_4$
$\hat{y}_n$

(A): Style Transfer

(B): Named Entity Recognition (NER)

Figure 1: The general pipeline of employing style transfer as data augmentation. For each sample from the source domain, we transfer it to the target domain by changing its style-related attributes as synthetic data, and then use it to train NER models. In the figure, the source and target domains are newswire and social media, respectively, and the underline words are entity instances.

by nonparallel task-specific datasets with different styles. Moreover, a solution that can successfully exploit data in this manner will be relevant in general data augmentation scenarios beyond NER.

We formulate style transfer as a paraphrase generation problem following previous work (Krishna et al., 2020; Qi et al., 2021) and propose a novel neural architecture that uses PLMs as the generator and discriminator. The generator aims to transform text guided by task prefixes while the discriminator aims to differentiate text between different styles. We jointly train the generator and discriminator on both parallel and nonparallel data to learn transformations in a semi-supervised manner. We apply paraphrase generation directly on the parallel data to bridge the gap between source and target styles. For nonparallel data, we use a cycle-consistent reconstruction to re-paraphrase back the paraphrased sentences to its original style. Additionally, to guarantee the generation of valid and coherent data, we present a constrained decoding algorithm based on prefix trees along with a set of key ingredients for data selection to reduce the noise in synthetic data. We systematically evaluate our proposed methods on 5 different domain pairs under different data regimes. Experiments and analysis show that our proposed method can effectively generate synthetic data by imitating the style of a target low-resource domain and significantly outperform previous state-of-the-art methods.

In summary, our contributions are as follows:

1. We propose a novel approach for data augmentation that leverages style transfer to improve the low-resource NER task and show that our approach can consistently outperform previous state-of-the-art methods in different data regimes.
2. We present a constrained decoding algorithm along with a set of key ingredients to guarantee the generation of valid and coherent data.
3. Our proposed solution is practical and can be expected to be applicable to other low-resource NLP tasks as it does not require any task-specific attributes.

2 Related Work

Style Transfer Style transfer aims to adjust the stylistic characteristics of a sentence while preserving its original meaning. It has been widely studied in both supervised (Jhamtani et al., 2017; Niu et al., 2018; Rao and Tetreault, 2018; Wang et al., 2019, 2020) and unsupervised manners (Yang et al., 2018; Li et al., 2018; Prabhumoye et al., 2018; John et al., 2019; Dai et al., 2019; He et al., 2020; Krishna et al., 2020; Liu et al., 2021b). Although style transfer is frequently utilized in many NLP tasks such as dialogue systems (Niu and Bansal, 2018; Zhu et al., 2021), sentiment transfer (Shen et al., 2017; Malmi et al., 2020), and text debiasing (Ma et al., 2020; He et al., 2021), its application on data augmentation still remains understudied by the NLP community. To facilitate research in this direction, we study style transfer as data augmentation and propose a novel approach to explore transferring the style-related attributes of text to add more variations in the training data.

Data Augmentation Data augmentation has been recently receiving increasing and widespread attention in the field of NER. The mainstream methods can be group into two categories: rule-based approaches (Bodapati et al., 2019; Dai and Adel, 2020; Lin et al., 2021; Simoncini and Spanakis, 2021), and model-based approaches (Ding et al., 2020; Nie et al., 2020; Zeng et al., 2020; Chen et al., 2020; Wenjing et al., 2021; Liu et al., 2021a; Wang and Henao, 2021; Zhao et al., 2021). Although previous methods have shown promising results, they may not perform well in low-resource domains as the size and diversity of the training data are limited. Recent work have studied data augmentation by leveraging the data from high-resource domains. Chen et al. (2021) investigated data transformation with implicit textual patterns while Zhang et al. (2021) explored replacing entities between domains with cross-domain anchor pairs. Motivated by their impressive results, we further study how to bridge the gap of data difference between domains and explore a better to leverage the data in high-resource domains for data augmentation.

3 Problem Formulation and Preliminaries

Considering a nonparallel NER dataset $\mathcal{D}$ consisting of source data $\mathcal{D}_{src}$ from a high-resource domain and target data $\mathcal{D}_{tgt}$ from a low-resource domain, training a model directly on $\mathcal{D}_{src}$ and evaluating on $\mathcal{D}_{tgt}$ is expected to give low prediction performance due to the stylistic differences (e.g., lexicons and syntax) between domains. In this work, we transform the data from a source domain to a target domain by changing text style and use the resulting transformed data to improve NER performance on $\mathcal{D}_{tgt}$. To this end, we assume access to a dataset $\mathcal{P}$ that contains pairs of parallel source and target sentences, respectively, to provide supervision signals for style transfer and a pre-trained generative model G_θ based on an encoder-decoder (Vaswani et al., 2017) architecture. Given an input sequence $x = \{x_1, x_2, ..., x_N\}$ of length N, the generator G_θ is pre-trained to maximize the conditional probability in an autoregressive manner:

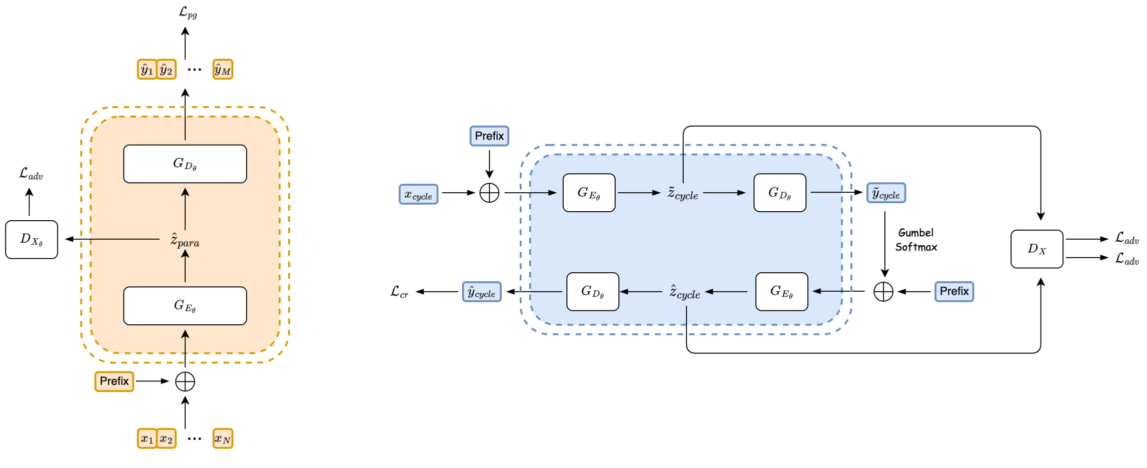
$$p_{\theta_G}(\hat{y}|x) = \prod_{i=1}^{M} p_{\theta_G}(\hat{y}_i|\hat{y}_{<i}, x)$$

where $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_M\}$ is the generated sentence of length M that has the same meaning as x but a different style.

Data Preparation Applying pre-trained generative models requires converting NER tokens and labels into a linearized format. In this work, we assume that the NER dataset follows the standard BIO labeling schema (Tjong Kim Sang and Veenstra, 1999). Previous work (Ding et al., 2020; Chen et al., 2021) has explored pre-pending each label to its corresponding token so that the model can capture the dependency and relationship between tokens and labels. However, we argue that this approach requires prohibitively large amounts of training data to adequately model the labeling schema. Recent work has also found that this format introduces too many hallucinated tokens and thus makes the learning problem significantly harder as the model needs to track token indices implicitly (Maynez et al., 2020; Raman et al., 2022). In the scenario where the number of annotated data is limited, the model may be confused with the labeling schema and tend to generate noisy samples. To mitigate this issue, we propose to linearize the data by only adding <START_ENTITY_TYPE> and <END_ENTITY_TYPE> special tokens to the beginning and the end of each entity span. For instance, the sample "A rainy day in [New]$_{\texttt{B-LOC}}$ [York]$_{\texttt{I-LOC}}$" will be converted to "A rainy day in <START_LOC> New York <END_LOC>". Furthermore, we also prepend task prefixes to the beginning of the sentences to guide the direction of style transfer, where a prefix is a sequence of words for task description specifying the source and target styles of transfer (e.g., "transfer source to target: ").

4 Method

In this work, we explore style transfer as a data augmentation method to improve the performance of NER systems on low-resource domains. We propose an adversarial learning framework to generate a paraphrase of the source data in a high-resource domain whose style conforms to the target data in a low-resource domain. The proposed framework comprises two main components: (1) *paraphrase generation*, which aims to rephrase the sentence to a different style with supervision, and (2) *cycle-consistent reconstruction*, which aims to transfer the sentence to a different style and then back to its original style with no



(a): Paraphrase Generation

(b): Cycle-consistent Reconstruction

Figure 2: The general architecture of our proposed method. Figure (a) shows the architecture for paraphrase generation, which aims to rephrase the sentence to a different style with supervision. Figure (b) shows the architecture for cycle-consistent reconstruction, which aims to transfer the sentence to a different style and then back to its original style. G_{E_θ} and G_{D_θ} refer to the encoder and decoder of the generator G_θ respectively while D_{X_θ} is the discriminator.

supervision. The overall architecture is shown in Figure 2. Besides, we design a constrained decoding algorithm to guarantee the generation of valid samples and present a set of key ingredients to select high-quality generated sentences.

4.1 Adversarial Learning with PLMs

Paraphrase Generation (PG) Recent work (Krishna et al., 2020; Qi et al., 2021) has demonstrated the effectiveness of reformulating style transfer as a controlled paraphrase generation problem. Here, we follow the same idea to perform the style transfer task in a supervised manner by using a gold standard annotated corpus $\mathcal{P}$. Specifically, as shown in Figure 2, given a sentence $x = \{x_1, x_2, ..., x_N\}$ of length N concatenated with a task prefix that specifies the source and target styles of transfer, the generator G_θ encodes it into a sequence of latent representations $\hat{z}_{para}$ and then decodes these representations into its paraphrase $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_M\}$ of length M which has the same meaning as x but a different style. The generator G_θ is trained with explicit supervision to maximize the log likelihood objective:

$$\mathcal{L}_{\text{pg}} = -\frac{1}{M}\sum_{i=1}^{M} y_i \cdot \log(\hat{y}_i)$$

Cycle-consistent Reconstruction (CR) Considering a nonparallel NER dataset consisting of data from $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$ in the source and target styles, respectively, we aim to change the style of text as a way to augment data. Previous mechanism enables the generator G_θ to learn different mapping functions, i.e., $G_\theta(x_{src}) \rightarrow \hat{y}_{tgt}$ and $G_\theta(x_{tgt}) \rightarrow \hat{y}_{src}$. Intuitively, the learned mapping functions should be reverses of each other. The sentence transferred by one mapping function should be able to be transferred back to its original representation using the other mapping function. Such cycle-consistency can not only encourage content preservation between the input and output, but also reduce the search space of mapping functions (Shen et al., 2017; Prabhumoye et al., 2018). To this end, as shown in Figure 2, we first use the generator G_θ to generate the paraphrase $\tilde{y}_{cycle}$ of the input sentence x_{cycle} concatenated with a prefix. As the gradients cannot be back-propagated through discrete tokens, we use Gumbel Softmax (Jang et al., 2017) for $\tilde{y}_{cycle}$ as a continuous approximation to recursively sample the tokens from the probability distribution:

$$p_{\theta_G}(\hat{y}_i|\hat{y}_{<i}, x) = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{|V|}\exp((\log(\pi_j) + g_j)/\tau)}$$

where π are class probabilities for tokens and $|V|$ denotes the size of vocabulary. g are i.i.d. samples drawn from Gumbel$(0, 1)$ distribution and τ is the temperature hyperparameter (Hinton et al., 2015). $\tau \rightarrow 0$ approximates a one-hot representation while $\tau \rightarrow \infty$ approximates a uniform distribution. Then we concatenate the paraphrase $\tilde{y}_{cycle}$ with a different prefix as the input to the generator G_θ and let it transfer the paraphrase back to the original sentence $\hat{y}_{cycle}$. The training objective for cycle-consistent reconstruction is formulated as:

$$\mathcal{L}_{\text{cr}} = \mathbb{E}_{x_{src} \sim X_{src}}[-\log p_{\theta_G}(\tilde{y}_{tgt}|x_{src})] + \mathbb{E}_{x_{tgt} \sim X_{tgt}}[-\log p_{\theta_G}(\tilde{y}_{src}|x_{tgt})]$$

Additionally, the generator G_θ is adversarially trained with a style discriminator D_θ, which takes as the input the latent representations of either the input sentence or its paraphrase, and discriminates the input between the source and target styles:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{x_{src} \sim X_{src}}[-\log D_\theta(G_\theta(x_{src}))] + \mathbb{E}_{x_{tgt} \sim X_{tgt}}[-\log(1 - D_\theta(G_\theta(x_{tgt})))]$$

Overall Training Objective The overall training objective can be formalized as:

$$\mathcal{L}(\theta_G, \theta_D) = \lambda_{\text{pg}}\mathcal{L}_{\text{pg}} + \lambda_{\text{cr}}\mathcal{L}_{\text{cr}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$$

where λ_{pg}, λ_{cr}, and λ_{adv} reflect the relative importance of $\mathcal{L}_{\text{pg}}$, $\mathcal{L}_{\text{cr}}$, and $\mathcal{L}_{\text{adv}}$, respectively.

The training process begins with the paraphrase generation as the first stage: the generator G_θ is trained with the paraphrase generation objective while the discriminator D_θ is trained with the adversarial learning objective. In the second stage, both paraphrase generation and cycle-consistent reconstruction are involved: the cycle-consistent reconstruction objective is further incorporated to train the generator G_θ.

4.2 Data Transformation

Constrained Decoding When given the input sequence x and the already generated tokens $\{y_1, y_2, ..., y_{i-1}\}$, a straightforward approach to generate the next word y_i is to greedily select the word with the highest probability at the timestep i. However, this greedy decoding approach may result in the sub-optimal decision and cannot guarantee to generate valid NER samples (e.g, mismatch of entity types and incomplete sentences).

To address these issues, we propose to apply a constrained decoding algorithm based on prefix trees (Cao et al., 2021; Lu et al., 2021) to control data transformation. Figure 3 presents our proposed algorithm. Specifically, the decoding starts with a <BOS> token and ends with a <EOS> token. At each decoding step, we apply top-k (Fan et al., 2018) and top-p (Holtzman et al., 2020) algorithms to navigate the search space. The prefix tree maintains two candidate vocabularies for text spans (i.e, <Text> node) and entity spans (i.e, <B_ENT> and <I_ENT> nodes), respectively. Based on the previous generated token, the constrained decoding dynamically prunes the vocabulary to lead the model to generate valid tokens. For example, if the previously generated token is <Text>, the model can only generate either <EOS> or <B_ENT> as the next token. Otherwise, it makes the sample invalid and noisy. We also adapt a slightly larger temperature (i.e., τ in Gumbel Softmax) to smooth the probability distribution towards the tokens that most likely conform to the target style.

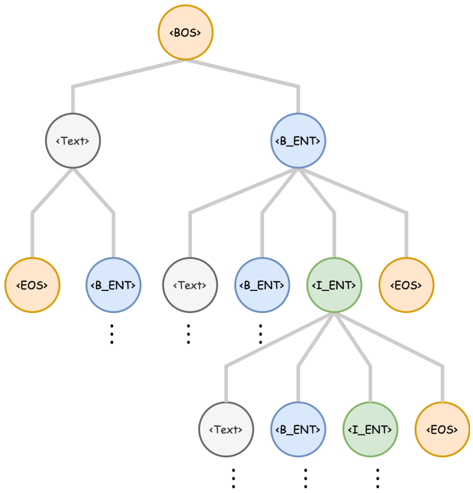


Figure 3: Prefix tree based constraints decoding algorithm for sentence generation. <BOS> and <EOS> represent the start and the end of each sentence. <Text> refers to the text spans while <B_ENT> and <I_ENT> refer to the start and inside of entity spans respectively.

Data Selection Even with a valid structure, the generated sentences may still remain unreliable as it may have a low quality due to degenerate repetition and incoherent gibberish (Holtzman et al., 2020; Welleck et al., 2020). To mitigate this issue, we further perform data selection with the following metrics:

- *Consistency*: a confidence score from a pretrained style classifier as the extent a generated sentence is in the target style.
- *Adequacy*: a confidence score from a pretrained NLU model on how much semantics is preserved in the generated sentence.
- *Fluency*: a confidence score from a pretrained NLU model indicating the fluency of

the generated sentence.

- *Diversity*: the edit distance between original sentences and the generated sentences at the character level.

For each sentence, we over-generate *k*=10 candidates. We calculate the above metrics (see Appendix C for more details) and assign a weighted score of these metrics to each candidate. Then we use the score to rank all candidates and select the best one for training NER systems.

5 Experiments

In this section, we present the experimental setup and results. We extensively evaluate our proposed method on five different domain pairs and compare our proposed method with state-of-the-art systems on data augmentation for NER in cross-domain scenarios.

5.1 Experimental Setup

Datasets We focus exclusively on formality style transfer which aims to transfer formal sentences to informal sentences. We use the parallel style transfer data from the GYAFC [2] (Rao and Tetreault, 2018) as $\mathcal{P}$. This corpus contains pairs of formal and informal sentences collected from Yahoo Answers. For the nonparallel NER data $\mathcal{D}$, we use a subset of OntoNotes 5.0 [3] corpus as source data $\mathcal{D}_{src}$ and Temporal Twitter Corpus [4] (Rijhwani and Preotiuc-Pietro, 2020) as target data $\mathcal{D}_{tgt}$. Here, we only consider the English datasets. The source data involves five different domains in the formal style: broadcast conversation (`BC`), broadcast news (`BN`), magazine (`MZ`), newswire (`NW`), and web data (`WB`) while the target data involves only social media (`SM`) domain in the informal style. The source and target data are nonparallel and we consider a total of 18 different entity types (e.g, *PERSON* and *LOCATION*). The data statistics and a list of entity types are shown in Appendix A. The details of data preprocessing and filtering are described in Appendix B.

Base Models For style transfer, we use a pre-trained T5_{base} model (Raffel et al., 2020) to initialize both the generator G_θ and discriminator D_θ. For NER, we use a sequence labeling framework consisting of a pre-trained $\text{BERT}_{\text{base}}$ model (Devlin et al., 2019) as the text encoder and a linear layer as the classifier to assign labels for each token. We use Huggingface Transformers library (Wolf et al., 2020) to implement all models. The details of hyper-parameters and fine-tuning are described in Appendix D.

Data Regimes To better understand the effectiveness of our proposed method, we undertake experiments in three different data regimes by varying the amount of training data in the target domain, namely FEW-SHOT, LOW-RESOURCE, and FULL-SET scenarios. For all scenarios, we assume full access to $\mathcal{P}$ and $\mathcal{D}_{src}$ but different access to $\mathcal{D}_{tgt}$. In the FEW-SHOT scenario, we adopt a N-way $K \sim 2K$-shot setting following Ding et al. (2021). We randomly select samples from $\mathcal{D}_{tgt}$ and ensure that each entity class contains $K \sim 2K$ examples. The K is set to 10 in our experiments and thus we will have 10 $\sim$ 20 samples from the target data. In the LOW-RESOURCE scenario, we simulate low-resource settings by randomly selecting 1024 samples from $\mathcal{D}_{tgt}$. For the FULL-SET scenario, we assume a full access to $\mathcal{D}_{tgt}$, i.e., use all of samples from the target data.

Compared Methods We investigate the following methods on data augmentation for NER in cross-domain scenarios for comparison: (1) **Adaptive Data Augmentation (ADA)** (Zhang et al., 2021) which proposes to augment sentences by replacing the entity in the source data with the entity in the target data that belongs to same entity class, and (2) **Cross-domain Data Augmentation (CDA)** (Chen et al., 2021) which studies to augment sentences by transforming data representations through an aligned feature space between the source and target data. We apply each method on source data to obtain the same amount of generated pseudo data. Each sample in the pseudo data corresponds to a sample in the source data. To make a fair comparison, we use the same base model (i.e., $\text{BERT}_{\text{base}}$ +Linear) but different training data generated from each method. The validation and test data are from the target domain. We do five runs for all experiments and report the average micro F1 score as the evaluation metric.

5.2 Main Results

Using Same Amount of Pseudo Data Here, we randomly select 1K, 2K, 3K, and 4K sam-

[2] `https://github.com/raosudha89/GYAFC-corpus`

[3] `https://catalog.ldc.upenn.edu/LDC2013T19`

[4] `https://github.com/shrutirij/temporal-twitter-corpus`

Method	Few-shot Setting				Low-resource Setting				Full-set Setting			
	1K	2K	3K	4K	1K	2K	3K	4K	1K	2K	3K	4K
BC $\rightarrow$ ***SM***												
Source (Baseline)	33.13	35.07	36.37	36.52	33.13	35.07	36.37	36.52	33.13	35.07	36.37	36.52
ADA (Zhang et al., 2021)	33.51	34.16	34.71	35.04	34.35	34.86	35.45	35.81	35.34	35.79	35.94	36.05
CDA (Chen et al., 2021)	35.21	37.14	39.48	39.59	36.59	40.75	42.23	43.26	45.04	49.81	52.06	53.81
Ours	39.95	41.58	43.10	43.71	48.77	50.71	52.26	53.37	52.84	58.04	59.35	60.32
BN $\rightarrow$ ***SM***												
Source (Baseline)	33.77	35.25	36.40	36.85	33.77	35.25	36.40	36.85	33.77	35.25	36.40	36.85
ADA (Zhang et al., 2021)	31.81	34.31	35.35	36.08	34.39	35.66	36.10	36.48	35.57	36.40	36.97	37.39
CDA (Chen et al., 2021)	31.92	35.65	37.55	39.22	36.44	39.23	40.45	41.07	41.81	47.66	50.83	52.52
Ours	36.47	40.04	40.72	41.38	44.87	47.57	50.29	51.52	53.56	56.81	59.01	59.40
MZ $\rightarrow$ ***SM***												
Source (Baseline)	26.04	31.38	32.17	33.36	26.04	31.38	32.17	33.36	26.04	31.38	32.17	33.36
ADA (Zhang et al., 2021)	28.92	32.83	34.19	35.07	30.35	33.66	35.10	35.78	32.97	35.13	35.87	35.99
CDA (Chen et al., 2021)	33.14	36.05	36.93	37.75	37.55	38.31	39.14	40.04	41.32	45.83	46.22	46.35
Ours	34.71	38.10	40.71	41.97	47.88	52.11	52.82	53.24	51.04	55.43	57.12	58.00
NW $\rightarrow$ ***SM***												
Source (Baseline)	34.15	35.64	36.57	36.77	34.15	35.64	36.57	36.77	34.15	35.64	36.57	36.77
ADA (Zhang et al., 2021)	34.05	34.70	35.47	35.94	34.78	35.84	36.18	36.45	35.29	36.77	37.50	37.62
CDA (Chen et al., 2021)	36.76	38.70	40.04	41.55	36.03	38.03	38.98	39.26	41.32	43.84	45.25	46.70
Ours	41.66	45.97	48.73	49.40	47.58	50.65	52.00	52.86	54.07	56.43	57.14	57.29
WB $\rightarrow$ ***SM***												
Source (Baseline)	8.23	12.93	15.08	15.96	8.23	12.93	15.08	15.96	8.23	12.93	15.08	15.96
ADA (Zhang et al., 2021)	9.93	15.96	17.12	18.57	12.40	17.93	18.13	18.17	18.19	21.64	23.34	24.62
CDA (Chen et al., 2021)	15.95	17.06	18.91	19.89	15.04	21.12	24.64	26.45	24.61	27.75	29.64	30.07
Ours	17.43	24.96	25.83	26.56	22.65	26.90	28.60	29.27	27.59	37.27	38.90	39.70
Target (SM)	-	-	-	-	-	-	-	-	68.42	73.49	75.36	76.63

Table 1: Performance comparison of different data augmentation methods with same amount of pseudo data for training. Scores are calculated with the micro F1 metric.

ples generated by each method as the training data to fine-tune the model. The baseline is established by fine-tuning the model on the same amount of data from the source domain. The validation and test data are from the target domain. Table 1 presents the performance comparison of different data augmentation methods with the same amount of pseudo data. Overall, our proposed method significantly outperforms the previous best method. On average, the F1 score increases by 4.7%, 10.1%, and 9.3% in FEW-SHOT, LOW-RESOURCE, and FULL-SET settings, respectively. Regarding the effect of data augmentation in cross-domain settings, we find that simply replacing the entities may decrease the model performance, especially when we only have limited amounts of target data. Although this method can comparatively address the word discrepancy by sharing entities across domains, it potentially increases the overlap between the training and test data, and thus reduces the model's generalization ability. Additionally, learning text differences between domains from only nonparallel samples suffers from the problem of data scarcity and can result in the generation of invalid and/or low-quality data. In contrast, our proposed method is more stable and consistently outperforms the baseline in different settings. We attribute this improvement to the fact that the proposed method can significantly increase the diversity from the perspective of words and entities while barely bringing semantic changes to the original text.

Using Large Amount of Pseudo Data Theoretically, we could generate an infinite amount of pseudo data for training. Thus, we undertake experiments using more pseudo data combined with target data for training. Here, we make comparison with three different method to support the effectiveness of our proposed method: (1) **S + T**: fine-tune on source and target data together, (2) **T only**: fine-tune on only target data, and (3) **S** $\rightarrow$ **T**: fine-tune on source data first and then target data. For our proposed method **P + T**, we gradually increase the number of generated samples combined

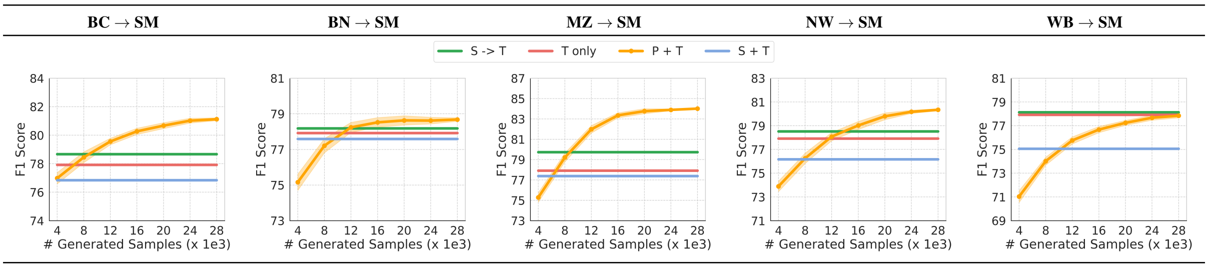


Table 2: Performance comparison of different data augmentation methods with large amount of pseudo data for training in FULL-SET setting: (1) **S + T**: fine-tune on source and target data together, (2) **T only**: fine-tune on only target data, (3) **S → T**: fine-tune on source data first and then target data, and (4) **P + T**, fine-tune on pseudo data and target data together while the number of samples from pseudo data is different in different steps. Scores are calculated with the micro F1 metric.

with target data as for training. We present the results in Figure 2. Notably, combining source data directly with target data could hurt the model performance. One possible explanation for such poor performance is that the distribution (e.g., lexicons and syntax) of source and target data can be very different, which may encourage the model to learn irrelevant patterns and thus lead to under-fitting. We also notice a similar phenomenon while only using a few amounts of samples combined with target samples. We argue that, in such cases, the model can have an inductive bias towards memorization instead of generalization, and thus perform poorly on test data. Additionally, fine-tuning on the source data first and then target data **S → T** can achieve better results than simply fine-tuning on source and target data together **S + T** or only target data **T only**. Nevertheless, with more and more generated samples for training, our proposed method can significantly boost the model performance comparing against other methods in four domain pairs (BC, BN, MZ, and NW) while remains very competitive in WB.

5.3 Analysis

Ablation Study In Table 3, we present ablation studies in the FULL-SET setting. For each domain pair, we generate the training data by randomly transferring 1K samples from the source domain while the validation and test data are from the target domain SM. We consider a series of ablation studies: (1) no cycle-consistent reconstruction, which indicates that we train the model with only the style transfer datasets, (2) no style discriminator, (3) no constrained decoding (i.e., no guarantee on the generation of valid sentences, and (4) no data selection. From Table 3, we can see that cycle-consistent reconstruction is critical for our proposed method to transfer the knowledge between domains. Without this component, the F1 score decreases by 39.15% on average. Additionally, constrained decoding also plays an important role in avoiding label mistakes, which can significantly hurt model performance. Moreover, the style discriminator is effective to enable the model to find generalized patterns while data selection can further boost the model performance. Overall, the ablation results demonstrate that each strategy in our proposed method is crucial in achieving promising results.

Method	BC	BN	MZ	NW	WB
Ours	52.84	53.56	51.04	54.07	27.59
- CR	15.30	13.82	7.11	4.15	2.98
- style discriminator	35.79	23.87	29.39	19.58	11.47
- constrained decoding	18.91	7.56	12.66	15.81	13.73
- data selection	46.48	52.31	48.12	49.30	20.81

Table 3: Ablation study with different domain as source data and SM as target data. CR refers to cycle-consistent reconstruction. Scores are calculated with the micro F1 metric.

Case Study Table 6 shows some hand-picked examples of formal source sentences and their corresponding informal target sentences generated from our proposed method. We observe that the generated sentences are embedded with some target domain characteristics (e.g., misspellings, grammar errors, language variations, and emojis) which significantly enhances entity context yet remains semantically related and coherent. This indicates that our proposed method can learn and transfer the text styles. Besides, we find that some entities in the original sentences can be replaced with not only those of the same entity type but also those of a different entity type. We also noticed that the proposed method tends to generate short sentences, imitating the target data to make the context ambiguous. However, we also note that the model may ignore the entity spans in the original sentences, which has a negative impact on the diversity of entities if the data is very limited. Besides, our model cannot deal well with the mismatch of labeling schema (i.e., the same entity span is labeled into different entity types in the source and target data) and has a limited ability to recognize rare entities. In the future, we plan to continue exploring approaches to address these issues.

6 Conclusion

In this paper, we propose an effective approach to employ style transfer as a data augmentation method. Specifically, we present an adversarial learning framework to bridge the gap between different text styles and transfer the entity knowledge from high-resource domains to low-resource domains. Additionally, to guarantee the generation of valid and coherent data, we design a constrained decoding algorithm along with a set of key ingredients for data generation and selection. We undertake experiments on five different domain pairs. The experimental results show that our proposed method can effectively transfer the data across domains to increase the size and diversity of training data for low-resource NER tasks. For the future work, we plan to explore style transfer as data augmentation in cross-lingual settings (i.e., transfer entity knowledge across languages instead of just domains). Additionally, since our approach is based on pre-trained language models, it would be interesting to explore leveraging pre-trained knowledge for data augmentation.

Limitations

Based on our studies, we find the following main limitations: (1) *mismatch of annotation schema*: we observe that the annotation schema between some NER datasets conflict with each other. The same entity span can be labeled into different entity types. For example, “America" is an instance of *GPE* in the OntoNotes 5.0 dataset but *LOCATION* in the WNUT17 dataset. This phenomenon introduces noise and make it difficult for models to understand entity types and learn transformations. (2) *mismatch of labeling schema*: the labeling schema in different NER datasets can be very different. For instance, OntoNotes 5.0 dataset contains 18 coarse-grained entity types while FEW-NERD contains 9 coarse-grained and 66 fine-grained entity types. Using such datasets as source and target data may not lead to a significant improvement gains. We hope our findings can inform potential avenues of improvement on data augmentation for NER and inspire the further work in this research direction.

Acknowledgements

This work was partially supported by the National Science Foundation (NSF) under grant #1910192. We would like to thank the anonymous reviewers for their valuable suggestions and the members from the RiTUAL lab at the University of Houston for their valuable feedback.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mathieu Dehouck and Carlos Gómez-Rodríguez. 2020. Data augmentation via subtree swapping for dependency parsing of low-resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3818–3830, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. 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)*, pages 3198–3213, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4173–4181, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021a. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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)*, pages 5834–5846, Online. Association for Computational Linguistics.

Yixin Liu, Graham Neubig, and John Wieting. 2021b. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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)*, pages 2795–2806, Online. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a seq2seq task. *ArXiv*, abs/2203.08378.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Walter Simoncini and Gerasimos Spanakis. 2021. SeqAttack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 308–318, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179, Bergen, Norway. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578, Hong Kong, China. Association for Computational Linguistics.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhu Wenjing, Liu Jian, Xu Jinan, Chen Yufeng, and Zhang Yujie. 2021. Improving low-resource named entity recognition via label-aware data augmentation and curriculum denoising. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1131–1142, Huhhot, China. Chinese Information Processing Society of China.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:*

System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7298–7309.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

Tao Zhang, Congying Xia, Philip S. Yu, Zhiwei Liu, and Shu Zhao. 2021. PDALN: Progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5441–5451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyan Zhao, Haibo Ding, and Zhe Feng. 2021. GLaRA: Graph-based labeling rule augmentation for weakly supervised named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3636–3649, Online. Association for Computational Linguistics.

Qingfu Zhu, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2021. Neural stylistic response generation with disentangled latent variables. 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)*, pages 4391–4401, Online. Association for Computational Linguistics.

A Data Statistics

Table 4 presents the data statistics of GYAFC corpus (Rao and Tetreault, 2018), OntoNotes 5.0 corpus and Temporal Twitter corpus (Rijhwani and Preotiuc-Pietro, 2020). We consider totally 18 different entity types following the annotation schema of OntoNotes 5.0 corpus, including *WORK_OF_ART*, *ORG*, *FAC*, *LAW*, *PERCENT*, *PRODUCT*, *MONEY*, *DATE*, *PERSON*, *GPE*, *QUANTITY*, *CARDINAL*, *NORP*, *TIME*, *EVENT*, *ORDINAL*, *LOC*, *LANGUAGE*.

Dataset	**Domain**	**Train**	**Dev**	**Test**
GYAFC Corpus	Formality	10,4562	10,268	4,849
OntoNotes 5.0 Corpus	BC	11,879	2,117	2,211
	BN	10,683	1,295	1,357
	MZ	6,911	642	780
	NW	34,970	5,896	2,327
	WB	15,554	2,316	2,307
Temporal Twitter Corpus	SM	10,000	500	1,500

Table 4: Data Statistics of GYAFC corpus, OntoNotes 5.0 corpus and Temporal Twitter corpus.

B Data Preprocessing and Filtering

Due to the limitation of computational resources, we set a max length of 64 to filter out long linearized sentences in both style transfer dataset $\mathcal{P}$ and NER dataset $\mathcal{D}$ for training the proposed framework to generate pseudo data. We also uses a pre-trained BERT$_{\text{large}}$ model (Devlin et al., 2019) to assign pseudo NER tags for sentences in the style transfer dataset $\mathcal{P}$ as weak supervision. The pre-trained BERT$_{\text{large}}$ model is only trained with the source data and has no access to the target data. The predicted NER tags are selected only if it comes with a confidence score (i.e., predicted probability) higher than 0.9 in both parallel source and target sentence. This results in approximately

10% of sentences in the style transfer dataset $\mathcal{P}$ having pseudo NER tags. For the NER dataset $\mathcal{D}$, we simply adapt original tags without further providing pseudo labels.

C Score Calculation in Data Selection

We simply fine-tune a T5$_{\text{base}}$ model (Raffel et al., 2020) for style classification as the style classifier to obtain the consistency score. The adequacy and fluency scores are obtained from the softmax confidence a pretrained NLU model [5].

D Hyper-parameters and Fine-tuning

Table 5 lists the hyper-parameters for both style transfer and NER tasks. All hyper-parameters are kept the same across different experiments for fine-tuning and/or generating. For the hardware, we use 8 NVIDIA V100 GPUs with a memory of 24GB. By adjusting the training batch size, our experiments should be compatible with any GPU that has a memory higher than 10GB.

Parameter	**Style Transfer**		**NER**
	Fine-tuning	**Generating**	**Fine-tuning**
Model Type	T5$_{\text{base}}$	T5$_{\text{base}}$	BERT$_{\text{base}}$
Optimizer	AdamW	-	Adam
Learning Rate	$1e-4$	-	$5e-5$
Training Epochs	5	-	10
Batch Size	6	-	32
Weight Decay	0.01	-	0.01
Top-k	50	50	-
Top-p	0.98	0.98	-
λ_{pg}	1.0	-	-
λ_{cr}	0.5	-	-
λ_{adv}	1.25	-	-
τ	1.0	1.5	-
$\lambda_{\text{Consistency}}$	-	1.0	-
$\lambda_{\text{Adequacy}}$	-	1.0	-
λ_{Fluency}	-	0.1	-
$\lambda_{\text{Diversity}}$	-	0.5	-

Table 5: Hyper-parameters for style transfer and NER tasks.

E Examples of Generated Data

Table 6 shows some examples of formal source sentences and their corresponding informal target sentences generated from our proposed method for case study.

[5] https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

#	Original Sentences (formal)	Generated Sentences (informal)
Good Examples		
1	The **[IRS]**$_{\textbf{ORGANIZATION}}$ denied cost deductions because few of the temps got to go aboard.	the **[IRS]**$_{\textbf{ORGANIZATION}}$ denying cost depossions... bcoz few of the guys got to go on board
2	Oh yes, the **[Atlantic]**$_{\textbf{ORGANIZATION}}$ was also pushed back **[300 yards]**$_{\textbf{QUANTITY}}$.	Yea-huh the **[Atlantic]**$_{\textbf{ORGANIZATION}}$ was pushed back of **[300 feet]**$_{\textbf{QUANTITY}}$ 😐
3	Her neighbour, **[Lin]**$_{\textbf{PERSON}}$, laughs in agreement, “Tea processing, it just kills you!"	Her friend **[Alexander Lu]**$_{\textbf{PERSON}}$ likes it the same way but the time that Tea process kills you is very bad lol
4	Thus although he is without employment, he can while away his time by collecting stones and studying **[the Book of Changes]**$_{\textbf{WORK\_OF\_ART}}$ and fengshui.	So as if he has no jobs now he can keep his time by collection of jags and studying **[the Yingko Ceramics Museum’s]**$_{\textbf{ORGANIZATION}}$ and fengshui.
5	**[Ting]**$_{\textbf{PERSON}}$ concludes : “He has lost his direction in terms of education policy, and he underestimates the education community."	**[Ting]**$_{\textbf{PERSON}}$: He’s lost his plan for education
Bad Examples		
6	The departing Mr. **[Cathcart]**$_{\textbf{PERSON}}$ says he has no worries about **[Kidder]**$_{\textbf{ORGANIZATION}}$ ’s future.	He says he has no worries
7	I don’t think the checks are worth $ **[15]**$_{\textbf{MONEY}}$ apiece, ” he says.	he says It’s weird they’re for $ 15 per check
8	Well the rescue vehicle if it’s required will have only **[four]**$_{\textbf{CARDINAL}}$ people on it.	well if necessary the rescue car.. WILL have only four + people on it
9	it’s going to be in small gathering places in middle **[America]**$_{\textbf{GPE}}$ with people saying some pretty horrible things	Small gathering places in the middle **[America]**$_{\textbf{LOC}}$ with people saying some pretty awful things
10	So **[al Jazeera TV]**$_{\textbf{ORGANIZATION}}$ station has also adopted this structure.	Because **[al Jazeera]**$_{\textbf{PERSON}}$ TV **[Imad]**$_{\textbf{PERSON}}$ has too follow that structure

Table 6: Examples of data augmentation with data transformation for case study. The sentences are tagged with their corresponding entities in brackets (e.g., [IRS]$_{\text{ORGANIZATION}}$).