FactMix: Using a Few Labeled In-domain Examples to Generalize to Cross-domain Named Entity Recognition

Linyi Yang[1,4*], Lifan Yuan[2*†], Leyang Cui[3], Wenyang Gao[1], Yue Zhang[1,4]
[1] School of Engineering, Westlake University
[2] Huazhong University of Science and Technology
[3] Tencent AI Lab
[4] Institute of Advanced Technology, Westlake Institute for Advanced Study
[1]{yanglinyi, gaowenyang, zhangyue}@westlake.edu.cn
[2]lievanyuan173@gmail.com [3]leyangcui@tencent.com

Abstract

Few-shot Named Entity Recognition (NER) is imperative for entity tagging in limited resource domains and thus received proper attention in recent years. Existing approaches for few-shot NER are evaluated mainly under in-domain settings. In contrast, little is known about how these inherently faithful models perform in cross-domain NER using a few labeled in-domain examples. This paper proposes a two-step rationale-centric data augmentation method to improve the model's generalization ability. Results on several datasets show that our model-agnostic method significantly improves the performance of cross-domain NER tasks compared to previous state-of-the-art methods, including the data augmentation and prompt-tuning methods. Our codes are available at https://github.com/lifan-yuan/FactMix.

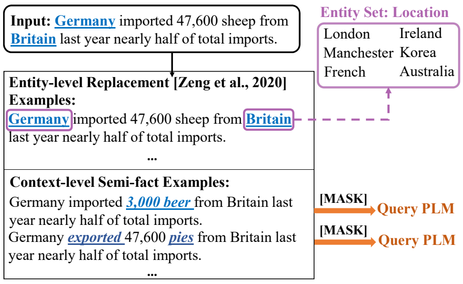

Input: Germany imported 47,600 sheep from Britain last year nearly half of total imports.
Entity Set: Location
London
Ireland
Manchester
Korea
French
Australia
Entity-level Replacement [Zeng et al., 2020] Examples:
Germany imported 47,600 sheep from Britain last year nearly half of total imports.
...
Context-level Semi-fact Examples:
Germany imported 3,000 beer from Britain last year nearly half of total imports.
Germany exported 47,600 pies from Britain last year nearly half of total imports.
...
[MASK]
Query PLM
[MASK]
Query PLM

Figure 1: The demonstration of two components of FactMix, namely context-level semi-fact and entity-level semi-fact examples.

1 Introduction

Named Entity Recognition (NER) is a subtask of natural language processing, which detects the mentions of named entities in input text, such as location, organization, and person (Sang and De Meulder, 2003; Yang et al., 2017; Cui et al., 2021). It has attracted research from academia and industry due to its broadened usage in customer services and document parsing as a core task in natural language understanding (Nadeau and Sekine, 2007; Ma and Hovy, 2016; Cui and Zhang, 2019; Yamada et al., 2020). However, training data for NER is available only for limited domains. It has been shown that such labeled data introduces challenges for a model to generalize to new domains (Snell et al., 2017; Ma et al., 2021a; Lin et al., 2021).

To address this problem, a line of research considers how to allow a model to effectively learn from a few labeled examples in a new target domain (Zhang et al., 2021; Ma et al., 2021b; Das et al., 2021; Chen et al., 2022; Wang et al., 2022a,b). However, such methods still require manual labeling for target domains, which makes them difficult to generalize to zero-shot diverse domain settings. A different line of research in NER considers data augmentation, using automatically constructed labeled examples to enrich training data. Zeng et al. (2020) consider using entity replacement to generate intervened new instances. We follow this line of work and consider a new setting – how to generate NER instances for data augmentations effectively – so that a few labeled examples in a source domain can generalize to arbitrary target domains.

Cross-domain NER poses unique challenges in practice. First, as a structured learning problem, it is essential to understand dependencies within the labels instead of classifying each token independently (Dai and Adel, 2020). While examples from different domains usually have different dependency patterns, which inevitably brings challenges for fine-tuning few-shot NER models to cross-domain tasks (Liu et al., 2021). Second, non-entity tokens in NER do not hold unified semantic meanings, but they could become noisy when combined with entity tokens in the training set. Such compositional generalization challenges

[*] Equal contribution. Random order of the authorship.
[†] Work done at Westlake University as an intern.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5360–5371
October 12–17, 2022.

have proven to be manifest in performance decay problems in various NLP tasks, such as sentiment analysis (Kaushik et al., 2019) and machine translation (Li et al., 2021), especially when faced with out-of-domain data.

As a consequence of the challenges above, spurious patterns between non-entity tokens and labels learned by models could obstruct the generalization of few-shot NER models in cross-domain settings. For example, given that "Jane monitored the patient's heart rate", 'Jane' is labeled as a person. The NER model will learn the relationship between the word 'Jane' and 'monitor' for the prediction. Suppose a NER model is trained on a medical domain and tested on the movie review. The correlation between 'Jane' and 'monitor' could become the 'spurious pattern' (Kaushik et al., 2019; Yang et al., 2021). From a causal perspective, spurious correlations are caused by confounding factors rather than a cause-effect relation.

To deal with these challenges and avoid spurious patterns, we present a novel model-agnostic, two-step, rationale-enhanced approach called **FactMix**, where we care about the efficacy of data augmentations for improving in-domain and out-of-domain (OOD) performance. We aim to leverage the contrast among – original, context-level semi-fact, and entity-level semi-fact instances – for teaching the model to capture more causal label dependencies between entities and the context. As Figure 1 shows, FactMix consists of two parts, namely context-level semi-fact generations and entity-level semi-fact instances generations. It is motivated by the natural intuition that models are much easier to learn from two-step contrastive examples compared to the one-step semi-fact augmentation (Zeng et al., 2020) [1].

The semi-factual generation component aims to alleviate the pitfall of non-entity tokens, which the previous data augmentation approach has not considered. We conduct synonym substitutions for non-entity tokens only. In particular, we mask the non-entity tokens, leverage the masked language models to predict the masked tokens, and replace the original tokens with predicted tokens. This replacement operation potentially introduces out-of-context information produced by the pre-trained masked language model when generating augmented examples. The entity-level semi-fact examples are generated by replacing the existing entity words in the training set. Finally, the augmented data generated by two steps will be mixed up together for training models. FactMix is a fully automatic method that does not require any additional hand-labeled data or human interventions and can be plugged for any few-shot NER models with different tuning strategies, including the standard fine-tuning and recent prompt-tuning.

Our method supoorts a new cross-domain NER setting, which is difficult from existing work. In particular, existing few-shot NER work considers in-domain fine-tuning (Ma et al., 2021a) and in-domain prompt-tuning (Cui et al., 2021). While our method also considers using only a source domain dataset for training models that generalize to target domains. Experimental results show that FactMix can achieve an average 3.16% performance gain in the in-domain fine-tuning setting compared to the state-of-the-art entity-level semi-fact generation approach (Zeng et al., 2020) and an average 6.85% improvement for prompt-tuning compared to EntLM (Ma et al., 2021b). Improvements in such a scale hint that FactMix builds a novel benchmark. To the best of our knowledge, we are the first to explore the cross-domain few-shot NER setting using fine-tuning and prompt-tuning methods.

2 Related Work

Cross-domain NER focuses on transferring NER models across different text styles (Pan et al., 2013; Xu et al., 2018; Liu et al., 2021; Chen et al., 2021). Current NER models cannot guarantee well-generalizing representation for out-of-domain data and result in sub-optimal performance. To address this issue, Lee et al. (2018) continue fine-tuning the model trained on the source domain by using the data from the target domain. Yang et al. (2017) jointly train NER models in both the source domain and target domain. Jia et al. (2019) and Jia and Zhang (2020) perform cross-domain knowledge transfer by using the language model. These methods rely on NER annotation or raw data in the target domain. In contrast, we propose a data argumentation method that only boosts cross-domain performance by using the source-domain corpus.

Few-shot NER aims to recognize pre-defined named entities by only using a few labeled ex-

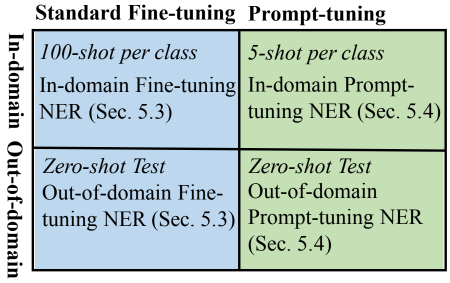
[1] Zeng et al. (2020) use "counterfactual" to denote the setting, where augmented data contains different entities with *the same* type compared with the original data. However, strictly speaking, "counterfactual" refers to augmented data that contains *different* types of entities with a minimum change of the input that can flip the predicted label. Hence, we use semi-fact instead in our paper

amples and is commonly used for evaluating structured prediction models in recent (Ravi and Larochelle, 2016; Snell et al., 2017; Das et al., 2021). Wiseman and Stratos (2019) and Yang and Katiyar (2020) propose distance-based methods, which copy the label of nearest neighbors. Huang et al. (2021) further investigates the efficacy of the self-training method on external data based on the distance-based methods. Cui et al. (2021) and Ma et al. (2021a) adopt prompt-based methods by using BART and BERT, respectively. These methods focus on designing few-shot-friendly models without any external guidance. In contrast, we augment both entity-level semi-fact and context-level semi-fact examples to boost the model performance on the new cross-domain few-sot setting.

The area of **Few-shot Cross-domain Learning** is motivated by the ability of humans to learn object categories from a few examples at a rapid pace, which is called rationale-based learning. Inductive bias (Baxter, 2000; Zhang et al., 2020) has been identified for a long time as a critical component. Benefits from the rapid development of large-scale pre-trained language models, few-shot learning, and out-of-distribution generalization become rapidly growing fields of NLP research (Brown et al., 2020; Shen et al., 2021; Chen et al., 2022). However, these two research directions have been separately explored in down-streaming tasks but rarely discussed together, except in the very recent study of sentiment analysis (Lu et al., 2022). To the best of our knowledge, we are the first to consider this setting for NER.

Data Augmentation through deformation has been known to be effective in various text classification tasks (Feng et al., 2021; Li et al., 2022), such as sentiment analysis (Yang et al., 2021; Lu et al., 2022) and natural language inference (Kaushik et al., 2021; Wu et al., 2022). In the task of NER, self-training has been applied to automatically increase the amount of training data (Wang et al., 2020). Paul et al. (2019) propose to combine self-training with noise handling on the self-labeled data to increase the robustness of the NER model. Bansal et al. (2020) and Wang et al. (2021) develop self-training and meta-learning techniques for training NER models with few labels, respectively.

In addition to self-training methods, prompt-based (Lee et al., 2021; Ma et al., 2021b) and causal-enhanced (Zeng et al., 2020) approaches have also surfaced in this domain, which are two important baselines for our work. Zeng et al. (2020) consider using the human intervention to generate the augmented data to improve few-shot NER models, and Ma et al. (2021b) aims to leverage the template-free prompt for boosting the performance of few-shot NER models. Nevertheless, both methods only focus on the in-domain accuracy while ignoring the cross-domain generalization of few-shot NER models.



	Standard Fine-tuning	Prompt-tuning
In-domain	*100-shot per class* In-domain Fine-tuning NER (Sec. 5.3)	*5-shot per class* In-domain Prompt-tuning NER (Sec. 5.4)
Out-of-domain	*Zero-shot Test* Out-of-domain Fine-tuning NER (Sec. 5.3)	*Zero-shot Test* Out-of-domain Prompt-tuning NER (Sec. 5.4)

Figure 2: The categorization of experiment settings.

3 Settings

We investigate the effectiveness of FactMix using different methods under several settings. We first introduce task settings in Section 3.1, then show the standard fine-tuning method and prompt-based method in Section 3.2 and Section 3.3, respectively.

3.1 Task Settings

The input of the NER system is a sentence $\mathbf{x} = x_1, \ldots, x_n$, which is a sequence of n words and the output is a sequence of NER tags $\mathbf{y} = y_1, \ldots, y_n$, where $y_i \in \mathcal{Y}$ for each word and $\mathcal{Y}$ is selected from a pre-defined label set$\{B - X, I - X, S - X, E - X...O\}$. B, I, E, S represent the beginning, middle, ending, and single-word entity, respectively. X indicates the entity type, such like PER and LOC, and O refers to the non-entity tokens. We use $\mathcal{D}_{ori}$ and $\mathcal{D}_{ood}$ to represent the original dataset and out-of-domain dataset, respectively.

Given small labelled instances of $\mathcal{D}_{ori}$, we first train a model $\mathcal{M}_{ori}$ through the standard fine-tuning method. We test the performance of $\mathcal{M}_{ori}$ on $\mathcal{D}_{ood}$ and $\mathcal{D}_{ood}$ under *In-domain Few-shot Setting* and *Out-of-domain Zero-shot Setting*, respectively, which can be seen in Figure 2.

3.2 Standard Fine-tuning Methods

Following Devlin et al. (2018), we feed contextualized word embeddings into a linear classification

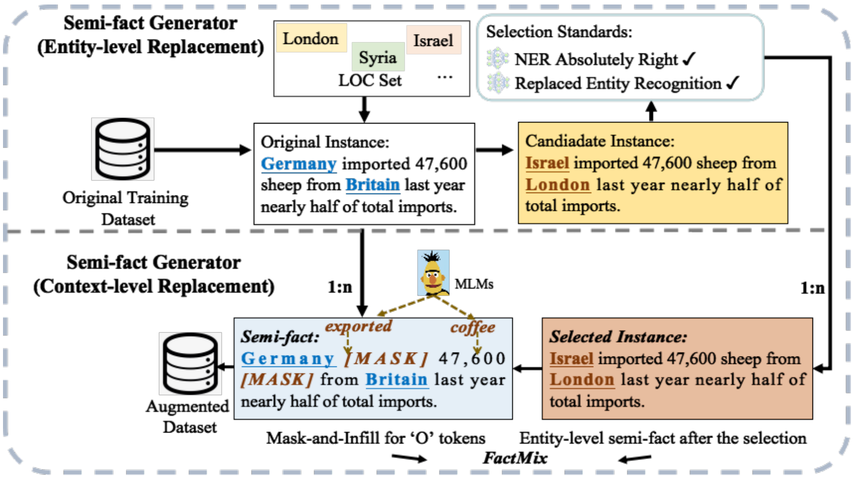

Semi-fact Generator
(Entity-level Replacement)
London
Israel
Syria
LOC Set
...
Selection Standards:
NER Absolutely Right ✓
Replaced Entity Recognition ✓
Original Training Dataset
Original Instance:
Germany imported 47,600 sheep from Britain last year nearly half of total imports.
Candiadate Instance:
Israel imported 47,600 sheep from London last year nearly half of total imports.
Semi-fact Generator
(Context-level Replacement)
1:n
MLMs
1:n
Semi-fact:
exported
coffee
Germany [MASK] 47,600 [MASK] from Britain last year nearly half of total imports.
Selected Instance:
Israel imported 47,600 sheep from London last year nearly half of total imports.
Augmented Dataset
Mask-and-Infill for 'O' tokens
Entity-level semi-fact after the selection
FactMix

Figure 3: The pipeline of the two-step FactMix approach operated on the source domain, which consists with entity-level semi-fact generations and context-level semi-fact generations.

with the softmax function to predict the probability distribution of entity types. Formally, we first feed the input $\mathbf{x}$ into the feature encoder PLM_θ to get the corresponding contextualized word embeddings $\mathbf{h}$:

$$\mathbf{h} = PLM_\theta(\mathbf{x}), \tag{1}$$

where $\mathbf{h}$ is the sequence of contextualized word embeddings based on pre-trained language models (PLMs), i.e., BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). We optimize the cross entropy loss $\mathcal{L}_{NER}$ by using AdamW (Loshchilov and Hutter, 2018), which is formulated as:

$$\mathcal{L}_{NER} = -\sum_{c=1}^{N} y_{o,c} \log\left(p_{o,c}\right), \tag{2}$$

where N is the number of classes, y is the binary indicator (0 or 1) depending on if the gold label c is the correct prediction for observation o, and p is the predicted probability for observation o of c.

3.3 Prompt-tuning Methods

Prompt-tuning NER reformulates classification tasks by using the mask-and-infill technique based on human-defined templates to generate label words. We perform the template-based and template-free prompt tuning as two additional experimental scenes to verify the validity of our method. Unlike the standard fine-tuning, no new parameters are introduced in this setting.

Template-based Approach Formally, we adopt the prompt template function $F_{\text{prompt}}(\cdot)$ proposed by a very recent work (Ma et al., 2021b) to converts the input $\mathbf{x}$ to a prompt input $x_{\text{prompt}} = F_{\text{prompt}}(x)$, and pre-defined label words $\mathcal{P}$ from the label set $\mathcal{Y}$ are generated through a mapping function $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{P}$. In particular, two slots need to be infilled for each instance: the input slot [X] is filled by the original input $\mathbf{x}$ directly, and the prompt slot [Z] is filled by the label word. To be note that [Z] is predicted by the masked language model (MLM) for prompt-based tuning in this work. The probability distribution over the label set $\mathcal{Y}$ can be optimized by the softmax function for predicting masked tokens using pre-trained models.

Template-free Approach In order to reduce the computational cost of the decoding process for template-based prompt tuning, Ma et al. (2021b) propose an entity-oriented LM (EntLM) objective for fine-tuning NER. Following Ma et al. (2021b), we first construct a label word set $\mathcal{P}_f$ by the label word engineering, which is also connected with the label set through a mapping function $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{P}_f$. Next, we replace entity tokens at entity positions with corresponding label word $\mathcal{M}(y_i)$. Finally, the target input can be represented as $\mathbf{x}^{\mathbf{Rep}} = \{x_1, ..., \mathcal{M}(y_i), ..., x_n\}$. We train the language model by maximizing the probability $P\left(\mathbf{x}^{\mathbf{Rep}} \mid \mathbf{x}\right)$. The loss function for generating the prompt can be formulated as:

$$\mathcal{L}_{RepLM} = -\sum_{i=1}^{N} \log P\left(x_i = x_i^{Rep} \mid \mathbf{x}\right), \tag{3}$$

where N is the number of classes. Initial parameters of the predictive model are obtained from pre-trained language models.

4 Method

FactMix automatically generates semi-fact examples for both standard fine-tuning and prompt-tuning. The pipeline of our approach is shown in Figure 3 and is made up of three components: (1) *entity-level semi-fact generator*; (2) *context-level semi-fact generator*; (3) *augmented data selection and mixing*. Briefly, a key innovation in this work is using a mixed semi-fact generator to improve the single entity-level data augmentation approach by adding the intermediate thinking process in a human-thinking manner.

4.1 Semi-factual Generation

We randomly remove one *O* token in each sentence. Specifically, we introduce out-of-context information by randomly masking an *O* word and then filling the span using the Masked Language Model (MLM), i.e., BERT (Devlin et al., 2018). Intuitively, we can generate numerous semi-factual samples because the MLM model can fill the masked span with multiple predictions. More importantly, choosing the number and order of the selected words is a combinatorial permutation problem. However, in practice, we find that more augmented data can not always lead to a better result; and for each semi-factual sample, we only replace one word or two-word phrase in a sentence using the top one mask-and-infill prediction of MLM.

Formally, given an input of NER as $\mathbf{x} = x_1, ..., x_i, ..., x_n$, where x_i is the chosen *O* word. We first mask x_i by replacing it with the [MASK] token, and thus get $\mathbf{x} = x_1, ..., [MASK], ..., x_n$. Then we fill the [MASK] token using *BERT-base-cased*[2] model and finally obtain a semi-factual example $\mathbf{x_{semi}} = x_1, ..., x_i{}', ..., x_n$. For instance, as seen in Fig. 3, *sheep* may first be masked and then infilled by an out-of-context word *coffee*, which can be generated by PLMs.

The intervention of the selected word may inflect the entity tag of other words and introduce extra noises into the dataset. Thus, we adopt a denoising mechanism that can filter out noisy examples by leveraging the predictive model trained on the original dataset that contains prior knowledge for NER tasks. Different from Zeng et al. (2020), who filter only those samples whose replaced entities cannot be predicted correctly, we use a stricter constraint to preserve only those samples where all tokens are predicted accurately.

4.2 Entity-level Semi-fact Generation

We generate entity-level semi-fact examples by interventions on the existing entity words. Specifically, for each training sample, we randomly select one of its entity words and replace it with words of the same type in a prepared $Entity\_Base$. For cases where data is not extremely scarce, e.g. in the fine-tuning setting in our experiments, the $Entity\_Base$ can be constructed by extracting and categorizing all entity words in the original dataset. Otherwise, e.g. in the 5-shot prompt-tuning setting in our experiments, the $Entity\_Base$ should be constructed from other available datasets.

Formally, given the input as $\mathbf{x} = x_1, \ldots, x_j, \ldots, x_n$, and x_j as the chosen entity word. We assume that the label of x_j is B-LOC and extract all the B-LOC entities in the $Entity\_Base$ and denote them as B-LOC Set. Next, a word in B-LOC Set is chosen to replace x_j and denoted as $x_j{}'$. In this way, the generated semi-fact sample is $\mathbf{x_{cf}} = x_1, \ldots, x_j{}', \ldots, x_n$. For example, as seen in Fig. 3, the B-LOC entity word *German* is replaced by *Israel* in B-LOC Set. All augmented samples are labeled as the same tag with original ones for saving manual efforts.

4.3 Mix Up

In the last step, we combine two types of automatically generated data by a mix-up strategy. Although the FactMix method can generate an unlimited amount of data theoretically, past experience (Lu et al., 2022) suggests that more fact-based data instances can not always bring performance benefits accordingly.

Following Zeng et al. (2020), we set the maximum augmentation ratio as 1:8 for the entity-level semi-fact data generation. While for context-level semi-fact generations, we set the ratio as 1:5. The optimal augmentation ratios for these two kinds of augmentations are jointly selected by the grid search on the development set of in-domain data. Finally, we obtain the final FactMix augmented training data, which can be represented as $\mathbf{x_{mix}} = Concat\{\mathbf{x_{semi}}, \mathbf{x_{cf}}\}$.

5 Experiments

As shown in Table 2, we conduct experiments under the scenarios of both fine-tuning and prompt-tuning, using in-domain and out-of-domain evalua-

[2] https://huggingface.co/bert-base-cased

Domain	# Instances			Entity Types
	Train	Dev	Test	
Reuters	14,987	3,466	3,684	Person, Location, Organization, Miscellaneous
TechNews	-	-	2000	
AI	-	-	431	
Literature	-	-	416	
Music	-	-	456	
Politics	-	-	651	
Science	-	-	543	

Table 1: Statistics of datasets used in experiments.

tions. We are also interested in better understanding the contributions of the two-step data augmentation approach when it comes to prediction performance. Thus, we consider several ablation studies to better the relative contributions of entity-level and context-level semi-fact augmented data. Micro F1 is used as evaluation metric for all settings.

5.1 Methodology

Fine-tuning. Given that FactMix is a model-agnostic data augmentation approach, we adopt the standard fine-tuning method based on two pre-trained models with different parameter sizes: BERT-base, BERT-large, RoBERT-base, and RoBERT-large. All backbone models are implemented on the transformer package provided by Huggingface [3]. To fine-tune NER models in a few-shot setting, we randomly sample 100 instances per label from the original dataset to ensure that the model converges. We report the average performance of models trained by five-times training.

Prompt-tuning. We adopt the recent EntLM model proposed by Ma et al. (2021b) as the benchmark for prompt-tuning. Following Ma et al. (2021b), we conduct the prompt-based experiments using the 5-shot training strategy. Again, we conduct a comparison between the state-of-the-art prompt-tuning method and several variants of FactMix. We also analyze the separate contribution of the counterfactual generator and semi-fact generator by providing an ablation study based on the the base and large versions of the BERT-cased backbone. For the standard hold-out test, we report results on both development and test sets. We also select two representative datasets for the out-of-domain test in terms of the highest (TechNews) and lowest (Science) word overlap with the original training domain (Reuters).

Dataset	Backbone	In-domain Fine-tuning Results			
		Ori	**CF**	**Semi**	**FactMix**
CoNLL2003 (Dev)	BERT-base-cased	57.98	79.78	81.48	**83.13***
	BERT-large-cased	69.18	83.27	85.87	**85.73***
	RoBERTa-base	52.44	85.81	87.99	**88.51***
	RoBERTa-large	68.81	88.25	89.39	**89.95***
CoNLL2003 (Test)	BERT-base-cased	54.03	77.71	78.70	**80.10***
	BERT-large-cased	65.38	81.11	**83.04**	82.65
	RoBERTa-base	48.53	82.74	85.05	**85.33***
	RoBERTa-large	65.70	85.20	86.84	**86.91***

Table 2: The Micro F1 score of different models by using FactMix and related data augmentation methods – CF: Entity-level Semi-fact Generation (Zeng et al., 2020); Semi: Context-level Semi-fact Generation (Ours); FactMix (Ours) – using the in-domain few-shot fine-tuning. $*$ indicates the statistically significant under T-test, p<0.05.

5.2 Datasets

The statistics of both source domain and out-of-domain datasets are introduced in Table 1. As a common understanding, it is easy to collect a large unlabeled corpus for one domain, while the corpus size could be small for low-resource domains. Then, we introduce datasets used in experiments for in-domain tests and out-of-domain tests, respectively, as follows.

In-domain Dataset. We conduct the in-domain experiments on the widely used CoNLL2003 (Sang and De Meulder, 2003) dataset with a text style of Reuters News and categories of person, location, organization, and others.

Out-of-domain Datasets. We adopt the cross-domain dataset collected by Liu et al. (2021) with new domains of AI, Literature, Music, Politics, and Science. Vocabularies for each domain are created by considering the top 5K most frequent words (excluding stopwords). Liu et al. (2021) report that vocabulary overlaps between domains are generally small, which further illustrates that the overlaps between domains are comparably small and out-of-domain datasets are diverse. Notably, since the model trained on CoNLL2003 can only predict person, location, organization, and various entities, we set all the unseen labels in OOD datasets to O.

5.3 Results on Few-shot Fine-tuning

In-domain experimental results on a widely used CoNLL2003 dataset show that FactMix achieves an average **3.16%** performance gain in the in-domain fine-tuning setting (100 instances per class) and an average **2.81%** improvement for prompt-tuning (5 instances per class) compared to the state-of-the-art data augmentation approach. For OOD test results,

[3] https://huggingface.co/models

Dataset	Backbone	Fine-tuning OOD Results				Dataset	Fine-tuning OOD Results			
		Ori	CF	Semi	FactMix		Ori	CF	Semi	FactMix
TechNews	BERT-base-cased	41.46	61.20	**65.20***	65.09*	Music	10.46	19.33	17.59	**19.49**
	BERT-large-cased	52.63	67.51	**69.98***	69.28		12.00	19.64	19.32	**19.97***
	RoBERTa-base	44.88	71.83	73.15	**73.62***		11.78	22.24	21.37	**23.75***
	RoBERTa-large	51.76	73.11	**74.89***	74.62		14.44	21.13	**22.93***	20.96
AI	BERT-base-cased	15.88	22.49	23.66	**24.67***	Politic	21.38	41.84	40.82	**43.60***
	BERT-large-cased	18.62	26.00	26.03	**26.25***		29.77	43.37	42.57	**43.84***
	RoBERTa-base	18.63	32.03	29.79	**32.09**		26.81	44.12	44.09	**44.66***
	RoBERTa-large	23.27	28.76*	29.77*	**30.06***		28.56	**45.87**	44.36	45.05
Literature	BERT-base-cased	12.85	22.89	23.05	**25.70***	Science	12.41	25.67	28.26	**29.72***
	BERT-large-cased	17.53	24.96	**26.25***	25.39		16.05	**28.75**	27.02	27.88
	RoBERTa-base	15.05	28.21	27.90	**28.89***		14.17	33.33	31.06	**34.13***
	RoBERTa-large	19.20	25.43	**26.76***	26.30*		17.25	31.36	29.89	**32.39***

Table 3: The average five times running results of Fine-tuning OOD over six datasets using various data augmentation approaches compared to the original training method (Standard Fine-tuning). CF: Entity-level Semi-fact Generation (Zeng et al., 2020); Semi: Context-level Semi-fact Generation (Ours); FactMix (Ours). $*$ indicates the statistically significant under T-test, p<0.05, when compared to CF.

Dataset	Backbone	Prompt-tuning In-domain Results				Dataset	Prompt-tuning OOD Results			
		EntLM	CF	Semi	FactMix		EntLM	CF	Semi	FactMix
CoNLL2003 (Dev)	BERT-base-cased	51.73	48.14	59.30	**62.40***	TechNews	47.16	52.36	50.96	**52.44***
	BERT-large-cased	60.95	58.42	49.53	**61.64***		**52.53**	48.32	32.48	48.64
CoNLL2003 (Test)	BERT-base-cased	54.00	55.61	57.23	**59.19***	Science	15.70	18.32	17.28	**18.62***
	BERT-large-cased	60.37	56.49	58.37	**60.80***		15.32	15.34	13.01	**16.80***

Table 4: The comparison among our methods, counterfactual data augmentation, and EntLM (Ma et al., 2021b) using prompt-tuning hold-out test and OOD test. $*$ indicates the statistically significant under T-test, p<0.05, when compared to EntLM.

FactMix increases absolute **14.19%** F1 score in average in fine-tuning compared to (Zeng et al., 2020) and **1.45%** increase in prompt-tuning compared to (Ma et al., 2021b).

In-domain Fine-tuning results are presented in Table 2 under the standard fine-tuning setting, using each of the baselines (Ori) and several variations of our FactMix approach. All results average five times running with randomly training instance selections.

FactMix achieves the best performance on both development and test sets, in terms of the highest Micro F1 score, excluding that the BERT-large model can achieve the best performance using our semi-fact augmentation approach only. Furthermore, we observe that improvements introduced by variants of the data augmentation approach are relatively significant when compared to models trained without data augmentations (**25.3%** absolute F1 improvements on average). FactMix also shows its superior performance compared to the previous state-of-the-art data augmentation method (Zeng et al., 2020) with a 2.1% absolute improvement in average. Finally, FactMix establishes a new state-of-the-art for the data augmentation approach in the cross-domain few-shot NER.

Out-of-domain Fine-tuning. We consider the performance of few-shot NER in the context of a more challenging cross-domain setting. The micro-f1 score of pre-trained models based on different augmentation methods is shown in Table 3. We find that the performance decay in technews is relatively lower than other domains since the technews domain also holds a relatively higher overlap with the training set (Reuters News). Again, our semi-factual generation and FactMix achieve the best performance in most settings. For instance, the RoBERTa-large model trained with Semi-fact Only and FactMix can achieve 74.89% and 74.62% F1, respectively, compared to only 51.76% F1 using the original training set. We also notice that all pre-trained methods manifest a significant drop in accuracy on other datasets, which share fewer overlaps with the training data than technews. For example, the RoBERTa-base model gets an **11.78%** F1 by using the standard fine-tuning, while it can be improved to **23.75%** with FactMix. Moreover, we can see that our methods, including Semi-fact and FactMix, achieve a significantly consistent improvement over different datasets compared to standard fine-tuning and the previous state-of-the-art method (Zeng et al., 2020), no matter the dataset distribution gap between domains. Finally, the ablation study shows that the mix-up strategy can effectively improve the performance of fine-tuning methods in most scenarios, compared to the single semi-fact augmentation method.

5.4 Results on Few-shot Prompt-tuning

To further understand the benefits of FactMix, in what follows, we also consider several ablation studies based on the few-shot prompt-tuning setting (5 instances per class).

In-domain Prompt-tuning. The results are shown in Table 4. We can see that FactMix achieves the best performance in 5-shot prompt-tuning on the development set and test set of CoNLL2003, compared to EntLM (Ma et al., 2021b) and the ablation part of FactMix. The overall Micro F1 score of prompt-tuning with FactMix is relatively lower than the results of 100-shot fine-tuning, i.e., **88.51 vs. 60.80** based on the BERT-large model. It is noteworthy that our approach shows its superior for all settings, while the previous data augmentation approach (Zeng et al., 2020) hurts the performance when using the BERT-large models, i.e., the F1 score decreases from 60.37 to 56.49 as shown in the test set. The stable performance further proves that two-step fact-based augmentations can significantly benefit NER models for both fine-tuning and prompt-tuning models.

Out-of-domain Prompt-tuning. The OOD results for prompt-tuning methods are also shown in Table 4. In general, we observe that prompt-based tuning methods have considerable potential for the cross-domain few-shot NER. While cross-domain results evaluated on the high-overlap dataset (TechNews) with the training domain are significantly higher than the low-overlap dataset (Science), i.e., **52.44 vs. 18.62** based on BERT-base. Furthermore, FactMix provides the best performance based on all of the pre-trained models, compared to EntLM and its variants. In contrast, EntLM performs better than FactMix on TechNews. It hints that our method could be more useful in a low-resource setting where the overlap between the original domain and target domain is relatively low.

5.5 Discussion

Benefiting from the generalized ability of pre-trained models, FactMix achieves much improved results on the few-shot in-domain test – 86.91%. More importantly, it shows decent scalability when combined with fine-tuning and prompt-tuning methods. To better understand the influence of the number of initial training examples and augmentation ratios, we illustrate the comparison of in-domain fine-tuning as follows.

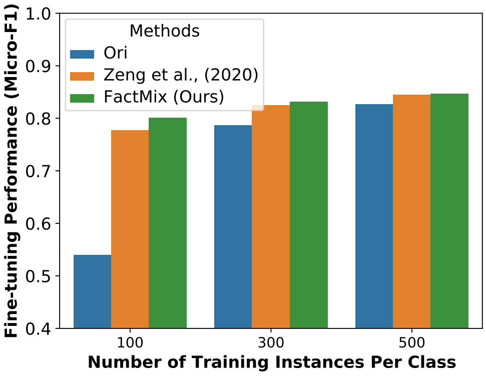


Figure 4: In-domain fine-tuning results are reported based on the BERT-base-cased model.

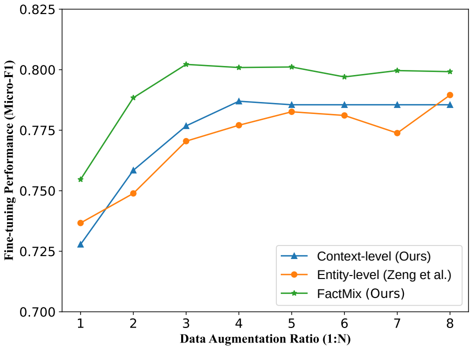


Figure 5: In-domain fine-tuning results based on BERT-base-cased using different augmentation ratios.

The Influence of Training Samples. The comparison of results based on the BERT-base-cased model is shown in Figure 4. We present the results of three different methods by using the different number of training examples varying from 100 to 500. Results show that FactMix holds the best performance when the size of training examples has been set as 100, 300, and 500. We also notice that the improvements introduced by FactMix decreased as the amount of raw training data per class increased from 100 to 300 when compared to the standard fine-tuning method. Finally, our method shows its superior for all settings when compared to the previous state-of-the-art data augmentation method (Zeng et al., 2020) for Few-shot NER.

The Influence of Augmentation Ratios. In-domain fine-tuning results using different augmentation ratios are shown in Figure 5. We consider three approaches in the evaluation, including semi-factual generation, FactMix, and the baseline method (Zeng et al., 2020). FactMix shows its absolute performance advantage using the augmentation ratio from one to eight. In particular, Micro-F1 scores of all methods increase with the increase of the number of augmented training instances when the augmentation ratio is less than 1:4, whereas the trend of increase gradually slow down when generating examples more than 1:4.

6 Conclusion

We proposed a joint context-level and entity-level semi-fact generation framework, FactMix, for better cross-domain NER using few labeled in-domain examples. Experimental results show that our method can not only boost the performance of pre-trained backbones in in-distribution and OOD datasets, but also show promising results combined with template-free prompt-tuning methods. As a single data augmentation method, FactMix can be useful for different NLP tasks to enable fast generalization, i.e., relation extraction, question answering, and sentiment analysis.

Acknowledgements

We acknowledge with thanks the discussion with Qingkai Min from Westlake University, as well as the many others who have helped. We would also like to thank anonymous reviewers for their insightful comments and suggestions to help improve the paper. This publication has emanated from research conducted with the financial support of the Pioneer and "Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHDX0003. Yue Zhang is the corresponding author.

References

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534.

Jonathan Baxter. 2000. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

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

Jiawei Chen, Qing Liu, Hongyu Lin, Xianpei Han, and Le Sun. 2022. Few-shot named entity recognition with self-describing networks. *arXiv preprint arXiv:2203.12252*.

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.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Leyang Cui and Yue Zhang. 2019. Hierarchically-refined label attention network for sequence labeling. 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 4115–4128, Hong Kong, China. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2021. Container: Few-shot named entity recognition via contrastive learning. *arXiv preprint arXiv:2109.07589*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik, Amrith Rajagopal Setlur, Eduard H. Hovy, and Zachary Chase Lipton. 2021. Explaining the efficacy of counterfactually-augmented data. *ArXiv*, abs/2010.02114.

Dong-Ho Lee, Mahak Agarwal, Akshen Kadakia, Jay Pujara, and Xiang Ren. 2021. Good examples make a faster learner: Simple demonstration-based learning for low-resource ner. *arXiv preprint arXiv:2110.08454*.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. 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 4767–4780.

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.

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

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. 2022. A rationale-centric framework for human-in-the-loop machine learning. *arXiv preprint arXiv:2203.12918*.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. 2021a. Template-free prompt tuning for few-shot NER. *CoRR*, abs/2109.13532.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. 2021b. Template-free prompt tuning for few-shot ner. *arXiv preprint arXiv:2109.13532*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Sinno Jialin Pan, Zhiqiang Toh, and Jian Su. 2013. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems (TOIS)*, 31(2):1–27.

Debjit Paul, Mittul Singh, Michael A Hedderich, and Dietrich Klakow. 2019. Handling noisy labels for robustly learning from self-training data for low-resource sequence labeling. *NAACL HLT 2019*, page 29.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *ICLR*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *ArXiv*, abs/2108.13624.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1737–1747.

Yidong Wang, Hao Chen, Yue Fan, SUN Wang, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. 2022a. Usb: A unified semi-supervised learning benchmark. *arXiv preprint arXiv:2208.07204*.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. 2022b. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Sam Wiseman and Karl Stratos. 2019. Label-agnostic sequence labeling by copying nearest neighbors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5363–5369, Florence, Italy. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets.

Jingjing Xu, Hangfeng He, Xu Sun, Xuancheng Ren, and Sujian Li. 2018. Cross-domain and semisupervised named entity recognition in chinese social media: A unified model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2142–2152.

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.

Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. *arXiv preprint arXiv:2106.15231*.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks.

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.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339.

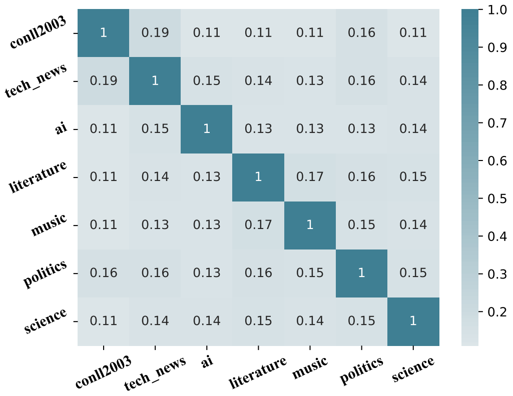


Figure 6: The word overlap between NER datasets from different domains.

A Appendix: Domain Distributions

The similarity between the dataset of source domain and six out-of-domain datasets is shown in Figure 6. We find that the technical news dataset shares the highest overlap ratio with the CoNLL2003 dataset, while the science domain shares the lowest overlap. Based on that, we select TechNews and Science as two representative datasets in prompt-tuning experiments. Also, the experimental results shown in Tables 3 and 4 demonstrate that cross-domain transfer between low-overlap domains still be a challenge problem, even for FactMix.