CrossNER: Evaluating Cross-Domain Named Entity Recognition

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji,
Samuel Cahyawijaya, Andrea Madotto, Pascale Fung

Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
zihan.liu@connect.ust.hk, pascale@ece.ust.hk

Abstract

Cross-domain named entity recognition (NER) models are able to cope with the scarcity issue of NER samples in target domains. However, most of the existing NER benchmarks lack domain-specialized entity types or do not focus on a certain domain, leading to a less effective cross-domain evaluation. To address these obstacles, we introduce a cross-domain NER dataset (CrossNER), a fully-labeled collection of NER data spanning over five diverse domains with specialized entity categories for different domains. Additionally, we also provide a domain-related corpus since using it to continue pre-training language models (*domain-adaptive pre-training*) is effective for the domain adaptation. We then conduct comprehensive experiments to explore the effectiveness of leveraging different levels of the domain corpus and pre-training strategies to do domain-adaptive pre-training for the cross-domain task. Results show that focusing on the fractional corpus containing domain-specialized entities and utilizing a more challenging pre-training strategy in domain-adaptive pre-training are beneficial for the NER domain adaptation, and our proposed method can consistently outperform existing cross-domain NER baselines. Nevertheless, experiments also illustrate the challenge of this cross-domain NER task. We hope that our dataset and baselines will catalyze research in the NER domain adaptation area. The code and data are available at https://github.com/zliucr/CrossNER.

Introduction

Named entity recognition (NER) is a key component in text processing and information extraction. Contemporary NER systems rely on numerous training samples (Ma and Hovy 2016; Lample et al. 2016; Chiu and Nichols 2016; Dong et al. 2016; Yadav and Bethard 2018), and a well-trained NER model could fail to generalize to a new domain due to the domain discrepancy. However, collecting large amounts of data samples is expensive and time-consuming. Hence, it is essential to build cross-domain NER models that possess transferability to quickly adapt to a target domain by using only a few training samples.

Existing cross-domain NER studies (Yang, Salakhutdinov, and Cohen 2017; Jia, Liang, and Zhang 2019; Jia and Zhang 2020) consider the CoNLL2003 English NER dataset (Tjong Kim Sang and De Meulder 2003) from

Reuters News as the source domain, and utilize NER datasets from Twitter (Derczynski, Bontcheva, and Roberts 2016; Lu et al. 2018), biomedicine (Nédellec et al. 2013) and CBS SciTech News (Jia, Liang, and Zhang 2019) as target domains. However, we find two drawbacks in utilizing these datasets for cross-domain NER evaluation. First, most target domains are either close to the source domain or not narrowed down to a specific topic or domain. Specifically, the CBS SciTech News domain is close to the Reuters News domain (both are related to news) and the content in the Twitter domain is generally broad since diverse topics will be tweeted about and discussed on social media. Second, the entity categories for the target domains are limited. Except the biomedical domain, which has specialized entities in the biomedical field, the other domains (i.e., Twitter and CBS SciTech News) only have general categories, such as person and location. However, we expect NER models to recognize certain entities related to target domains.

In this paper, we introduce a new human-annotated cross-domain NER dataset, dubbed CrossNER, which contains five diverse domains, namely, politics, natural science, music, literature and artificial intelligence (AI). And each domain has particular entity categories; for example, there are "politician", "election" and "political party" categories specialized for the politics domain. As in previous works, we consider the CoNLL2003 English NER dataset as the source domain, and five domains in CrossNER as the target domains. We collect ∼1000 development and test examples for each domain and a small size of data samples (100 or 200) in the training set for each domain since we consider a low-resource scenario for target domains. In addition, we collect the corresponding five unlabeled domain-related corpora for the domain-adaptive pre-training, given its effectiveness for domain adaptation (Beltagy, Lo, and Cohan 2019; Donahue et al. 2019; Lee et al. 2020; Gururangan et al. 2020).

We evaluate existing cross-domain NER models on our collected dataset and explore using different levels of domain corpus and masking strategies to continue pre-training language models (e.g., BERT (Devlin et al. 2019)). Results show that emphasizing the partial corpus with specialized entity categories in BERT's domain-adaptive pre-training (DAPT) consistently improves its domain adaptation ability. Additionally, in the DAPT, BERT's masked language modeling (MLM) can be enhanced by intentionally masking

contiguous random spans, rather than random tokens. Comprehensive experiments illustrate that the span-level pre-training consistently outperforms the original MLM pre-training for NER domain adaptation. Furthermore, experimental results show that the cross-domain NER task is challenging, especially when only a few data samples in the target domain are available. The main contributions of this paper are summarized as follows:

- We introduce CrossNER, a fully-labeled dataset spanning over five diverse domains, as well as the corresponding five domain-related corpora for studies of the cross-domain NER task.

- We report a set of benchmark results of existing strong NER models, and propose competitive baselines which outperform the current state-of-the-art model.

- To the best of our knowledge, we are the first to conduct in-depth experiments and analyses in terms of the number of target domain training samples, the size of domain-related corpus and different masking strategies in the DAPT for the NER domain adaptation.

Related Work

Existing NER Datasets CoNLL2003 (Tjong Kim Sang and De Meulder 2003) is the most popular NER dataset and is collected from the Reuters News domain. It contains four general entity categories, namely, person, location, organization and miscellaneous. The Email dataset (Lawson et al. 2010), Twitter dataset (Derczynski, Bontcheva, and Roberts 2016; Lu et al. 2018), and SciTech News dataset (Jia, Liang, and Zhang 2019) have the same or a smaller set of entity categories than CoNLL2003. WNUT NER (Strauss et al. 2016), from the Twitter domain, has ten entity types.[1] However, aside from four entity types that are the same as those in CoNLL2003, the other six types come from four domains, which means they are not concentrated on a particular domain. Different from these datasets, the OntoNotes NER dataset (Pradhan et al. 2012) consists of six genres (newswire, broadcast news, broadcast conversation, magazine, telephone conversation and web data). However, the six genres are either relatively close (e.g., newswire and broadcast news) or have broad content (e.g., web data and magazine). To the best of our knowledge, only Biomedical NER (Nédellec et al. 2013) and CORD-NER (Wang et al. 2020) focus on specific domains (biomedicine and COVID-19, respectively) which are diverse from the News domain and have specialized entity classes. However, annotations in CORD-NER are produced by models instead of annotators.

Cross-Domain NER Cross-domain algorithms alleviate the data scarcity issue and boost the models' generalization ability to target domains (Kim et al. 2015; Yang, Liang, and Zhang 2018; Lee, Dernoncourt, and Szolovits 2018; Lin and Lu 2018; Liu, Winata, and Fung 2020). Daumé III (2007) enhanced the adaptation ability by mapping the entity label space between the source and target domains. Wang

et al. (2018) proposed a label-aware double transfer learning framework for cross-specialty NER, while Wang, Kulkarni, and Preoţiuc-Pietro (2020) investigated different domain adaptation settings for the NER task. Liu et al. (2020b) introduced a two-stage framework to better capture entities for input sequences. Sachan et al. (2018); Jia, Liang, and Zhang (2019) injected target domain knowledge into language models for the fast adaptation, and Jia and Zhang (2020) presented a multi-cell compositional network for NER domain adaptation. Additionally, fast adaptation algorithms have been applied to low-resource languages (Lample and Conneau 2019; Liu et al. 2019, 2020a,c; Wilie et al. 2020), accents (Winata et al. 2020), and machine translation (Artetxe et al. 2018; Lample et al. 2018).

The CrossNER Dataset

To collect CrossNER, we first construct five unlabeled domain-specific (politics, natural science, music, literature and AI) corpora from Wikipedia. Then, we extract sentences from these corpora for annotating named entities. The details are given in the following sections.

Unlabeled Corpora Collection

Wikipedia contains various categories and each category has further subcategories. It serves as a valuable source for us to collect a large corpus related to a certain domain. For example, to construct the corpus in the politics domain, we gather Wikipedia pages that are in the politics category as well as its subcategories, such as political organizations and political cultures. We utilize these collected corpora to investigate domain-adaptive pre-training.

NER Data Collection

Pre-Annotation Process For each domain, we sample sentences from our collected unlabeled corpus, which are then given named entity labels. Before annotating the sampled sentences, we leverage the DBpedia Ontology (Mendes, Jakob, and Bizer 2012)[2] to automatically detect entities and pre-annotate the selected samples. By doing so, we can alleviate the workload of annotators and potentially avoid annotation mistakes. However, the quality of the pre-annotated NER samples is not satisfactory since some entities will be incorrectly labeled and many entities are not in the DBpedia Ontology. In addition, we utilize the hyperlinks in Wikipedia and mark tokens that have hyperlinks to facilitate the annotation process and assist annotators in noticing entities. This is because tokens having hyperlinks are highly likely to be the named entity.

Annotation Process Each data sample requires two well-trained NER annotators to annotate it and one NER expert to double check it and give final labels. The data collection proceeds in three steps. First, one annotator needs to detect and categorize the entities in the given sentences. Second, the other annotator checks the annotations made by the first

---

[1]Ten categories: organization, location, person, facility, movie, music artist, product, sports team, TV show and miscellaneous.

[2]The DBpedia Ontology contains 320 entity classes and categorizes 3.64 million entities.

Domain	Unlabeled Corpus			Labeled NER			Entity Categories
	# paragraph	# sentence	# tokens	# Train	# Dev	# Test	
Reuters	-	-	-	14,987	3,466	3,684	person, organization, location, miscellaneous
Politics	2.76M	9.07M	176.56M	200	541	651	politician, person, organization, political party, event, election, country, location, miscellaneous
Natural Science	1.72M	5.32M	98.50M	200	450	543	scientist, person, university, organization, country, location, discipline, enzyme, protein, chemical compound, chemical element, event, astronomical object, academic journal, award, theory, miscellaneous
Music	3.49M	9.82M	194.62M	100	380	456	music genre, song, band, album, musical artist, musical instrument, award, event, country, location, organization, person, miscellaneous
Literature	2.69M	9.17M	177.33M	100	400	416	book, writer, award, poem, event, magazine, person, location, organization, country, miscellaneous
Artificial Intelligence	97.04K	287.62K	5.20M	100	350	431	field, task, product, algorithm, researcher, metrics, university, country, person, organization, location, miscellaneous

Table 1: Data statistics of unlabeled domain corpora, labeled NER samples and entity categories for each domain.

annotator, makes markings if he/she thinks that the annotations **could** be wrong and gives another annotation. Finally, the expert first goes through the annotations again and checks for possible mistakes, and then makes the final decision for disagreements between the first two annotators. In order to ensure the quality of annotations, the second annotator concentrates on looking for possible mistakes made by the first annotator instead of labeling from scratch. In addition, the expert will give a second round check and confer with the first two annotators when he/she is unsure about the annotations. A total 63.64% of entities (the number of pre-annotated entities divided by the number of entities with hyperlinks) are pre-annotated based on the DBpedia Ontology, 73.33% of entities (the number of corrected entities divided by the number of entities with hyperlinks) are corrected in the first annotation stage, 8.59% of entities are annotated as possibly incorrect in the second checking stage, and finally, 8.57% of annotations (out of all annotations) are modified by the experts. The details are reported in the Appendix.

Data Statistics

The data statistics of the Reuters News domain (Tjong Kim Sang and De Meulder 2003) and the collected five domains are illustrated in Table 1. In general, it is easy to collect a large unlabeled corpus for one domain, while for some low-resource domains, the corpus size could be small. As we can see from the statistics of the unlabeled corpus, the size is large for all domains except the AI domain (only a few AI-related pages exist in Wikipedia). Since DAPT experiments usually require a large amount of unlabeled sentences (Wu et al. 2020; Gururangan et al. 2020), this data scarcity issue introduces a new challenge for the DAPT.

We make the size of the training set (from Table 1) relatively small since cross-domain NER models are expected to do fast adaptation with a small-scale of target domain data samples. In addition, there are domain-specialized entity types for each domain, resulting in a hierarchical category structure. For example, there are "politician" and "person" classes, but if a person is a politician, that person should be annotated as a "politician" entity, and if not, a "person" entity. Similar cases can be found for "scientist" and "person", "organization" and "political party", etc. We believe this hierarchical category structure will bring a challenge to
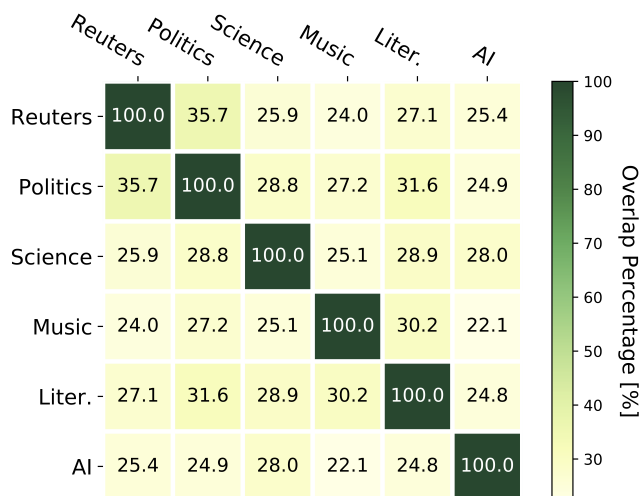

Figure 1: Vocabulary overlaps between domains (%). Reuters denotes the Reuters News domain, "Science" denotes the natural science domain and "Litera." denotes the literature domain.

this task since the model needs to better understand the context of inputs and be more robust in recognizing entities.

Domain Overlap

The vocabulary overlaps of the NER datasets between domains (including the source domain (Reuters News domain) and the five collected target domains) are shown in Figure 1. Vocabularies for each domain are created by considering the top 5K most frequent words (excluding stopwords). We observe that the vocabulary overlaps between domains are generally small, which further illustrates that the overlaps between domains are comparably small and the domains of our collected datasets are diverse. The vocabulary overlaps of the unlabeled corpora between domains are reported in the Appendix.

Domain-Adaptive Pre-training

We continue pre-training the language model BERT (Devlin et al. 2019) on the unlabeled corpus (i.e., DAPT) for the do-

main adaptation. The DAPT is explored in two directions. First, we investigate how different levels of the corpus influences the pre-training. Second, we explore the effectiveness between token-level and span-level masking in the DAPT.

Pre-training Corpus

When the size of the domain-related corpus is enormous, continuing to pre-train language models on it would be time-consuming. In addition, there would be noisy and domain-unrelated sentences in the collected corpus which could weaken the effectiveness of the DAPT. Therefore, we investigate whether extracting more indispensable content from the large corpus for pre-training can achieve comparable or even better cross-domain performance.

We consider three different levels of corpus for pre-training. The first is the **domain-level** corpus, which is the largest corpus we can collect related to a certain domain. The second is the **entity-level** corpus. It is a subset of the domain-level corpus and made up of sentences having plentiful entities. Practically, it can be extracted from the domain-level corpus based on an entity list. We leverage the entity list in DBpedia Ontology and extract sentences that contain multiple entities to construct the entity-level corpus. The third is the **task-level** corpus, which is explicitly related to the NER task in the target domain. To construct this corpus, we select sentences having domain-specialized entities existing in the DBpedia Ontology. The size of the task-level corpus is expected to be much smaller than the entity-level corpus. However, its content should be more beneficial than that of the entity-level corpus.

Taking this further, we propose to **integrate** the entity-level and the task-level corpus. Instead of simply merging these two corpora, we first upsample the task-level corpus (double the size in practice) and then combine it with the entity-level corpus. Hence, models will tend to focus more on the task-level sentences in the DAPT.

Span-level Pre-training

Inspired by Joshi et al. (2020), we propose to change the token-level masking (MLM) in BERT (Devlin et al. 2019) into span-level masking for the DAPT. In BERT, MLM first randomly masks 15% of the tokens in total, and then replaces 80% of the masked tokens with special tokens (`[MASK]`), 10% with random tokens and 10% with the original tokens. We follow the same masking strategy as BERT except the first masking step. In the first step, after the random masking, we move the individual masked index position into its adjacent position that is next to another masked index position in order to produce more masked spans, while we do not touch the continuous masked indices (i.e., masked spans). For example, the randomly masked sentence:

```
Western music's effect would [MASK] to
grow within the country [MASK] sphere
```
would become
```
Western music's effect would continue
to grow within the [MASK] [MASK] sphere.
```

Intuitively, span-level masking provides a more challenging task for pre-trained language models. For example, predicting "San Francisco" is much harder than predicting

"San" given "Francisco" as the next word. Hence, the span-level masking can facilitate BERT to better understand the domain text so as to complete the more challenging task.

Experiments

Experimental Settings

We consider the CoNLL2003 English NER dataset (Tjong Kim Sang and De Meulder 2003) from Reuters News, which contains person, location, organization and miscellaneous entity categories, as the source domain and five domains in CrossNER as target domains. Our model is based on BERT (Devlin et al. 2019) in order to have a fair comparison with the current state-of-the-art model (Jia and Zhang 2020), and we follow Devlin et al. (2019) to fine-tune BERT on the NER task. More training details are in the Appendix.

Before training on the source or target domains, we conduct the DAPT on BERT when the unlabeled domain-related corpus is leveraged. Moreover, in the DAPT, different types of unlabeled corpora are investigated (i.e., domain-level, entity-level, task-level and integrated corpora), and different masking strategies are inspected (i.e., token-level and span-level masking). Then, we carry out three different settings for the domain adaptation, which are described as follows:

- We ignore the source domain training samples, and fine-tune BERT directly on the target domain data.
- We first pre-train BERT on the source domain data, and then fine-tune it to the target domain samples.
- We jointly fine-tune BERT on both source and target domain data samples. Since the size of the data samples in the target domains is smaller than in the source domain, we upsample the target domain data to balance the source and target domain data samples.

Baseline Models

We compare our methods to the following baselines:

- BiLSTM-CRF (Lample et al. 2016) incorporate bidirectional LSTM (Hochreiter and Schmidhuber 1997) and conditional random fields for named entity recognition. We combine source domain data samples and the upsampled target domain data samples to jointly train this model (i.e., the joint training setting mentioned in the experimental settings). We use the word-level embeddings from Pennington, Socher, and Manning (2014) and the char-level embeddings from Hashimoto et al. (2017).
- Liu et al. (2020b) proposed a new framework, Coach, for slot filling and NER domain adaptation. It splits the task into two stages by first detecting the entities and then categorizing the detected entities.
- Jia, Liang, and Zhang (2019) integrated language modeling tasks and NER tasks in both source and target domains to perform cross-domain knowledge transfer. We follow their settings and provide the domain-level corpus for the language modeling tasks.
- Jia and Zhang (2020) proposed a multi-cell compositional LSTM structure based on BERT representations (Devlin et al. 2019) for domain adaptation, which is the current state-of-the-art cross-domain NER model.

Models	Masking	Corpus	Politics	Science	Music	Litera.	AI	Average
Fine-tune Directly on Target Domains (Directly Fine-tune)								
		w/o DAPT	66.56	63.73	66.59	59.95	50.37	61.44
BERT-based	Token-level	Domain-level	67.21	64.63	70.56	62.54	53.66	63.72
		Entity-level	67.59	65.97	70.64	63.77	53.94	64.38
		Task-level	67.30	65.04	70.37	62.10	53.19	63.60
		Integrated	**68.83**	**66.55**	72.42	63.95	**55.44**	65.44
	Span-level	Entity-level	68.58	66.70	71.62	64.67	55.65	65.44
		Task-level	68.37	65.84	70.66	63.85	54.48	64.64
		Integrated	**70.45**	**67.59**	**73.39**	**64.96**	**56.36**	**66.55**
Pre-train on the Source Domain then Fine-tune on Target Domains (Pre-train then Fine-tune)								
		w/o DAPT	68.71	64.94	68.30	63.63	58.88	64.89
BERT-based	Token-level	Domain-level	69.37	66.68	72.05	65.15	61.48	66.95
		Entity-level	70.32	67.03	71.55	65.76	61.52	67.24
		Task-level	70.21	65.99	71.74	65.32	60.29	66.71
		Integrated	**71.44**	**67.53**	**74.02**	**66.57**	**61.90**	**68.29**
	Span-level	Entity-level	71.85	68.04	73.34	66.28	61.66	68.23
		Task-level	70.77	67.41	73.01	66.58	61.68	67.89
		Integrated	**72.05**	**68.78**	**75.71**	**69.04**	**62.56**	**69.63**
Jointly Train on Both Source and Target Domains (Jointly Train)								
		w/o DAPT	68.85	65.03	67.59	62.57	58.57	64.52
BERT-based	Token-level	Domain-level	69.49	66.37	71.94	63.74	60.53	66.41
		Entity-level	70.01	66.55	71.51	63.35	61.29	66.54
		Task-level	70.14	66.06	70.70	62.68	60.14	65.94
		Integrated	**71.09**	**67.58**	72.57	**64.27**	62.55	**67.61**
	Span-level	Entity-level	71.90	68.04	71.98	64.23	61.63	67.55
		Task-level	71.31	67.75	71.17	63.24	60.83	66.86
		Integrated	**72.76**	**68.28**	**74.30**	**65.18**	**63.07**	**68.72**
Baseline Models								
BiLSTM-CRF (word)	-	-	52.52	44.6	40.77	35.69	38.24	42.36
BiLSTM-CRF (word + char)	-	-	56.60	49.97	44.79	43.03	43.56	47.59
Coach (word)	-	-	54.01	44.88	45.58	36.18	40.41	44.21
Coach (word + char)	-	-	61.50	52.09	51.66	48.35	45.15	51.75
Jia, Liang, and Zhang (2019)	-	-	68.44	64.31	63.56	59.59	53.70	61.92
Jia and Zhang (2020)	-	-	70.56	66.42	70.52	66.96	58.28	66.55
+ DAPT (Span-level & Integrated)	-	-	**71.45**	**67.68**	**74.19**	**68.63**	**61.64**	**68.71**

Table 2: F1-scores of our proposed methods in three settings and baseline models. Results are averaged over three runs.

Results & Analysis

Corpus Types & Masking Strategies

From Table 2, we can see that DAPT using the entity-level or the task-level corpus achieves better or on par results with using the domain-level corpus, while according to the corpus statistics illustrated in Table 3, the size of the entity-level corpus is generally around half or less than half that of the domain-level corpus, and the size of the task-level corpus is much smaller than the domain-level corpus. We conjecture that the content of the corpus with plentiful entities is more suitable for the NER task's DAPT. In addition, selecting sentences with plentiful entities is able to filter numerous noisy sentences and partial domain-unrelated sentences from the domain corpus. Picking sentences having domain-specialized entities also filters a great many sentences that are not explicitly related to the domain and makes the DAPT more effective and efficient. In general, DAPT using the task-level corpus performs slightly worse than using the entity-level corpus. This can be attributed to the large corpus size differences. Furthermore, integrating the entity-level and task-level corpora is able to consistently boost the adaptation performance compared to utilizing other cor-

pus types, although the size of the integrated corpus is still smaller than the domain-level corpus. This is because the integrated corpus ensures the pre-training corpus is relatively large, and in the meantime, focuses on the content that is explicitly related to the NER task in the target domain. The results suggest that the corpus content is essential for the DAPT, and *we leave exploring how to extract effective sentences for the DAPT for future work*. Surprisingly, the DAPT is still effective for the AI domain even though the corpus size in this domain is relatively small, which illustrates that the DAPT is also practically useful in a small corpus setting.

As we can see from Table 2, when leveraging the same corpus, the span-level masking consistently outperforms the token-level masking. For example, in the Pre-train then Fine-tune setting, DAPT on the integrated corpus and using span-level masking outperforms that using token-level masking by a 1.34% F1-score on average. This is because predicting spans is a more challenging task than predicting tokens, forcing the model to better comprehend the domain text and then to possess a more powerful capability to do the downstream tasks. Moreover, adding DAPT using the span-level masking and the integrated corpus to Jia and Zhang (2020) further improves the F1-score by 2.16% on average. *Never-*

	Politics	**Science**	**Music**	**Litera.**	**AI**
Domain-level	177M (1x)	99M (1x)	195M (1x)	177M (1x)	5.2M (1x)
Entity-level	67M (0.37x)	36M (0.36x)	96M (0.49x)	87M (0.49x)	2.7M (0.52x)
Task-level	16M (0.09x)	3.9M (0.04x)	26M (0.13x)	14M (0.08x)	0.2M (0.04x)
Integrated	99M (0.56x)	44M (0.44x)	148M (0.76x)	115M (0.65x)	3.1M (0.60x)

Table 3: Number of tokens of different corpus types. The number in the brackets represents the size ratio between the corresponding corpus and the domain-level corpus.
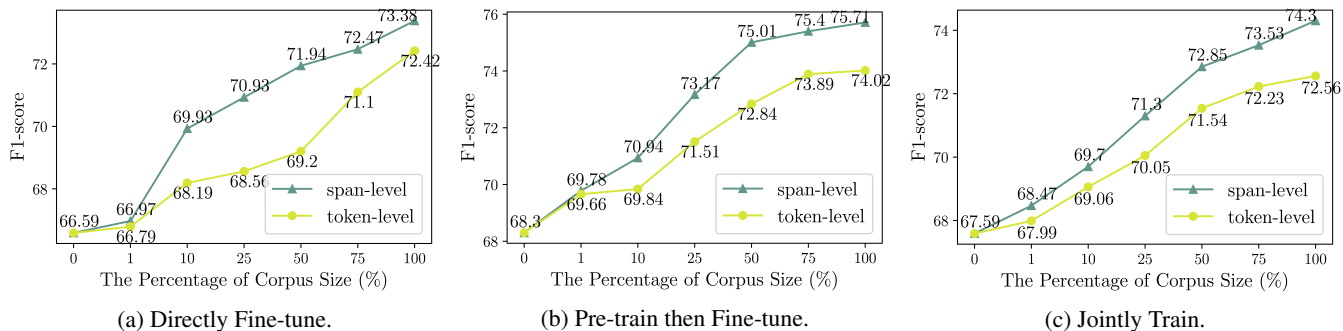

Figure 2: Comparisons among utilizing different percentages of the music domain's integrated corpus and masking strategies in the DAPT for the three settings.

theless, we believe that exploring more masking strategies or DAPT methods is worthwhile. We leave this for future work.

From Table 2, we can clearly observe the improvements when the source domain data samples are leveraged. For example, compared to the Directly Fine-tune, Pre-train then Fine-tune (w/o DAPT) improves the F1-score by 3.45% on average, and Jointly Train (w/o DAPT) improves the F1-score by 3.08% on average. We notice that Pre-train then Fine-tune generally leads to better performance than Jointly Train. We speculate that jointly training on both the source and target domains makes it difficult for the model to concentrate on the target domain task, leading to a sub-optimal result, while for the Pre-train then Fine-tune, the model learns the NER task knowledge from the source domain data in the pre-training step and then focuses on the target domain task in the fine-tuning step. Finally, we can see that our best model can outperform the existing state-of-the-art model in all five domains. However, the averaged F1-score of the best model is not yet perfect (lower than 70%), which highlights the need for more advanced cross-domain models.

Performance vs. Unlabeled Corpus Size

Given that a large-scale domain-related corpus might sometimes be unavailable, we investigate the effectiveness of different corpus sizes for the DAPT and explore how the masking strategies will influence the adaptation performance. As shown in Figure 2, as the size of unlabeled corpus increases, the performance generally keeps improving. This implies that the corpus size is generally essential for the DAPT, and within a certain corpus size, the larger the corpus is, the better the domain adaptation performance the DAPT will produce. Additionally, we notice that in the Pre-train then Fine-tune setting, the improvement becomes compa-

rably less when the percentage reaches 75% or higher. We conjecture that it is relatively difficult to improve the performance when it reaches a certain amount.

Furthermore, little performance improvements are observed for both the token-level and span-level masking strategies when only a small-scale corpus (1% of the music integrated corpus, 1.48M) is available. As the corpus size increases, the span-level masking starts to outperform the token-level masking. We notice that in Directly Fine-tune, the performance discrepancy between the token-level and span-level is first increasing and then decreasing. And the performance discrepancies are generally increasing in the other two settings. We hypothesize that the span-level masking can learn the domain text more efficiently since it is a more challenging task, while the token-level masking requires a larger corpus to better understand the domain text.

Performance vs. Target Domain Sample Size

From Figure 3, we can see that the performance drops when the number of target domain samples is reduced, the span-level pre-training generally outperforms token-level pre-training, and the task becomes extremely difficult when only a few data samples (e.g., 10 samples) in the target domain are available. Interestingly, as the target domain sample size decreases, the advantage of using source domain training samples becomes more significant, for example, Pre-train then Fine-tune outperforms Directly Fine-tune by ∼10% F1 when the sample size is reduced to 40 or lower than 40. This is because these models are able to gain the NER task knowledge from the large amount of source domain examples and then possess the ability to quickly adapt to the target domain. Additionally, using DAPT significantly improves the performance in the Pre-train then Fine-tune setting when target

Figure 3: Few-shot F1-scores (averaged over three runs) in the music domain. We use the integrated corpus for the DAPT.

Models	Genre	Song	Band	Album	Artist	Country	Loc.	Org.	Per.	Misc.
Directly Fine-tune										
w/o DAPT	77.35	26.57	71.65	60.61	80.71	84.93	67.82	59.36	7.97	17.20
Span-level + Integrated	78.89	42.63	**83.03**	65.93	**85.70**	83.86	77.17	69.43	10.09	19.20
Pre-train then Fine-tune										
w/o DAPT	79.12	42.04	68.45	61.79	76.75	**88.85**	78.47	69.70	10.15	**29.80**
Span-level + Integrated	**80.82**	**58.67**	82.72	**69.28**	84.58	85.61	**80.54**	**75.16**	**12.59**	28.67

Table 4: F1-scores (averaged over three runs) for the categories in the music domain over the Directly Fine-tune and Pre-train then Fine-tune settings. Span-level+Integrated denotes that the span-level masking and integrated corpus are utilized for the DAPT. "Loc.", "Org.", "Per." and "Misc." denote "Location", "Organization", "Person" and "Miscellaneous", respectively.

domain samples are scarce (e.g., 10 samples). This can be attributed to the boost in the domain adaptation ability made by the DAPT, which allows the model to quickly learn the NER task in the target domain. Furthermore, we notice that with the decreasing of the sample size, the performance discrepancy between Pre-train then Fine-tune and Jointly Train is getting larger. We speculate that in the Jointly Train setting, the models focus on the NER task in both the source and target domains. This makes the models tend to ignore the target domain when the sample size is too small, while for the Pre-train then Fine-tune setting, the models can focus on the target domain in the fine-tuning stage to ensure the good performance in the target domain.

Fine-grained Comparison

In this section, we further explore the effectiveness of the DAPT and leveraging NER samples in the source domain. As shown in Table 4, the performance is improved on almost all categories when the DAPT or the source domain NER samples are utilized. We observe that using source domain NER data might hurt the performance in some domain-specialized entity categories, such as "artist" ("musical artist") and "band". This is because that since "artist" is a subcategory of "person" and models pre-trained on the source domain tend to classify artists as "person" entities. Similarly, "band" is a subcategory of "organization", which

leads to the same misclassification issue after the source domain pre-training. When the DAPT is used, the performance on some domain-specialized entity categories is greatly improved (e.g., "song", "band" and "album"). We notice that the performance on the "person" entity is relatively low compared to other categories. It is because that the hierarchical category structure could cause models to be confused between "artist" and "person" entities, and we find out that 84.81% of "person" entities are misclassified as "artist" in the best model we have.

Conclusion

In this paper, we introduce CrossNER, a humanly-annotated NER dataset spanning over five diverse domains with specialized entity categories for each domain. In addition, we collect the corresponding domain-related corpora for the study of DAPT. A set of benchmark results of existing strong NER models is reported. Moreover, we conduct comprehensive experiments and analyses in terms of the size of the domain-related corpus and different pre-training strategies in the DAPT for the cross-domain NER task, and our proposed method consistently outperforms existing baselines. Nevertheless, the performance of our best model is not yet perfect, especially when the number of target domain training samples is limited. We hope that our dataset will facilitate further research in the NER domain adaptation field.

References

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2018. Unsupervised Neural Machine Translation. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=Sy2ogebAW.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pre-trained Language Model for Scientific Text. 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)*, 3606–3611.

Chiu, J. P.; and Nichols, E. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4: 357–370.

Daumé III, H. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 256–263.

Derczynski, L.; Bontcheva, K.; and Roberts, I. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1169–1179.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Donahue, C.; Mao, H. H.; Li, Y. E.; Cottrell, G. W.; and McAuley, J. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. *arXiv preprint arXiv:1907.04868* .

Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; and Di, H. 2016. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, 239–250. Springer.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 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*, 8342–8360. Online: Association for Computational Linguistics.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1923–1933.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Jia, C.; Liang, X.; and Zhang, Y. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2464–2474.

Jia, C.; and Zhang, Y. 2020. Multi-Cell Compositional LSTM for NER Domain Adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5906–5917.

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8: 64–77.

Kim, Y.-B.; Stratos, K.; Sarikaya, R.; and Jeong, M. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 473–482.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270.

Lample, G.; and Conneau, A. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* .

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rkYTTf-AZ.

Lawson, N.; Eustice, K.; Perkowitz, M.; and Yetisgen-Yildiz, M. 2010. Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 71–79.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4): 1234–1240.

Lee, J. Y.; Dernoncourt, F.; and Szolovits, P. 2018. Transfer Learning for Named-Entity Recognition with Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lin, B. Y.; and Lu, W. 2018. Neural Adaptation Layers for Cross-domain Named Entity Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2012–2022.

Liu, Z.; Shin, J.; Xu, Y.; Winata, G. I.; Xu, P.; Madotto, A.; and Fung, P. 2019. Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables. 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)*, 1297–1303.

Liu, Z.; Winata, G. I.; and Fung, P. 2020. Zero-Resource Cross-Domain Named Entity Recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 1–6. Online: Association for Computational Linguistics. doi:

10.18653/v1/2020.repl4nlp-1.1. URL https://www.aclweb.org/anthology/2020.repl4nlp-1.1.

Liu, Z.; Winata, G. I.; Lin, Z.; Xu, P.; and Fung, P. 2020a. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8433–8440.

Liu, Z.; Winata, G. I.; Xu, P.; and Fung, P. 2020b. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 19–25. Online: Association for Computational Linguistics.

Liu, Z.; Winata, G. I.; Xu, P.; Lin, Z.; and Fung, P. 2020c. Cross-lingual Spoken Language Understanding with Regularized Representation Alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7241–7251.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1990–1999.

Ma, X.; and Hovy, E. 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)*, 1064–1074.

Mendes, P. N.; Jakob, M.; and Bizer, C. 2012. DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC*, 1813–1817. Citeseer.

Nédellec, C.; Bossy, R.; Kim, J.-D.; Kim, J.-J.; Ohta, T.; Pyysalo, S.; and Zweigenbaum, P. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, 1–7.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, 1–40.

Sachan, D. S.; Xie, P.; Sachan, M.; and Xing, E. P. 2018. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine Learning for Healthcare Conference*, 383–402.

Strauss, B.; Toma, B.; Ritter, A.; De Marneffe, M.-C.; and Xu, W. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, 138–144.

Tjong Kim Sang, E. F.; and De Meulder, F. 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-Volume 4*, 142–147.

Wang, J.; Kulkarni, M.; and Preoţiuc-Pietro, D. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8476–8488.

Wang, X.; Song, X.; Guan, Y.; Li, B.; and Han, J. 2020. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218* .

Wang, Z.; Qu, Y.; Chen, L.; Shen, J.; Zhang, W.; Zhang, S.; Gao, Y.; Gu, G.; Chen, K.; and Yu, Y. 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)*, 1–15.

Wilie, B.; Vincentio, K.; Winata, G. I.; Cahyawijaya, S.; Li, X.; Lim, Z. Y.; Soleman, S.; Mahendra, R.; Fung, P.; Bahar, S.; and Purwarianti, A. 2020. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Winata, G. I.; Cahyawijaya, S.; Liu, Z.; Lin, Z.; Madotto, A.; Xu, P.; and Fung, P. 2020. Learning Fast Adaptation on Cross-Accented Speech Recognition. In *Proc. Interspeech 2020*, 1276–1280. doi:10.21437/Interspeech.2020-0045. URL http://dx.doi.org/10.21437/Interspeech.2020-0045.

Wu, C.-S.; Hoi, S.; Socher, R.; and Xiong, C. 2020. Todbert: Pre-trained natural language understanding for task-oriented dialogues. *arXiv preprint arXiv:2004.06871* .

Yadav, V.; and Bethard, S. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2145–2158.

Yang, J.; Liang, S.; and Zhang, Y. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3879–3889.

Yang, Z.; Salakhutdinov, R.; and Cohen, W. W. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.