DoSEA: A Domain-specific Entity-aware Framework for Cross-Domain Named Entity Recognition

Minghao Tang[1,2], **Peng Zhang**[2,3]*, **Yongquan He**[4], **Yongxiu Xu**[1,2],
Chengpeng Chao[1,2] and **Hongbo Xu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, UCAS, Beijing, China
[3]School of Cyberspace Security, NJUST, Nanjing, China
[4]Meituan, Beijing, China
{tangminghao,hbxu}@iie.ac.cn, heyongquan@meituan.com

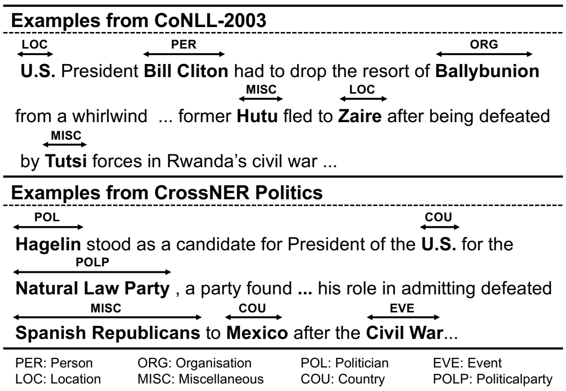
Abstract

Cross-domain named entity recognition aims to improve performance in a target domain with shared knowledge from a well-studied source domain. The previous sequence-labeling based method focuses on promoting model parameter sharing among domains. However, such a paradigm essentially ignores the domain-specific information and suffers from entity type conflicts. To address these issues, we propose a novel machine reading comprehension based framework, named DoSEA, which can identify domain-specific semantic differences and mitigate the subtype conflicts between domains. Concretely, we introduce an entity existence discrimination task and an entity-aware training setting, to recognize inconsistent entity annotations in the source domain and bring additional reference to better share information across domains. Experiments on six datasets prove the effectiveness of our DoSEA. Our source code can be obtained from https://github.com/mhtang1995/DoSEA.

1 Introduction

Named entity recognition(NER) is a fundamental task in natural language processing and has been extensively studied in various domains. However, acquiring a high performance NER model heavily relies on labor-intensive annotated data (Huang et al., 2015; Devlin et al., 2019). Thus, there is a growing interest in cross-domain NER, which aims to exploit the information on a well-studied source domain to improve the performance in a target domain (Pan and Yang, 2010). Following Daumé III (2007), we focus on the supervised cross-domain NER setting, which utilizes abundant annotated samples from the source domain and small annotated samples from the target domain.

Previous studies (Kim et al., 2015; Lin and Lu, 2018; Wang et al., 2018b; Jia and Zhang, 2020) typically treat cross-domain NER task as a sequence labeling problem, classifying each word as a type of entity. However, cross-domain NER is challenging due to the entity type difference between domains, since the target domain contains specific entity types. As Figure 1 shows, CrossNER Politics dataset (Liu et al., 2021) contains specific entity types(e.g., Event, Politician, Country and Political Party), which are not labeled in the CoNLL-2003 dataset (Sang and Meulder, 2003). Thus, the sequence-labeling based method commonly adopts separate model structures(CRF or softmax layer) for each domain, which primarily limits the parameter sharing across domains.



Examples from CoNLL-2003

U.S. President **Bill Cliton** had to drop the resort of **Ballybunion** from a whirlwind ... former **Hutu** fled to **Zaire** after being defeated by **Tutsi** forces in Rwanda's civil war ...

Examples from CrossNER Politics

Hagelin stood as a candidate for President of the **U.S.** for the **Natural Law Party** , a party found ... his role in admitting defeated **Spanish Republicans** to **Mexico** after the **Civil War**...

Figure 1: Examples from CoNLL-2003 and CrossNER Politics dataset.

A series of fine-tune methods (Lee et al., 2017; Lin and Lu, 2018) and multi-task learning methods (Wang et al., 2018b; Jia and Zhang, 2020) have been proposed for promoting parameter sharing. The fine-tune method first trains a model using source domain samples, then fine-tunes the model using target domain samples with an initialized label decoder. However, it depends on the sizes of target domain samples to learn a strong label decoder. Conversely, the multi-task learning method simultaneously trains a model for both domains under the jointly training strategy, and it essentially adds auxiliary tasks to facilitate parameter

[*] Peng Zhang is the corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 2147–2156
October 12–17, 2022.

sharing. Specifically, Wang et al. (2018b) adds KL-divergence in the features of identical entity types between each domain's CRF layer. Jia and Zhang (2020) models entity type features as separate cell states in a multi-cell compositional LSTM structure, then shares the same entity type's features across domains. However, they ignore the domain-specific information and suffer from entity type conflicts. For example, CoNLL-2003 and CrossNER Politics both contain "Location" entities, but the latter requires a distinction between "Location" and "Country", which is not considered in the former. In this case, previous work (Liu et al., 2021) indicates the subtype conflicts, since the cross-domain NER model may tend to classify "Country" as “Location” entities.

To this end, we propose a novel framework named Domain-specific Entity-aware(DoSEA) network for cross-domain NER, which aims to mitigate the negative impacts raised by the domain-specific entity types. Specifically, instead of assigning a separate model structure for each domain, our DoSEA formalizes NER as a machine reading comprehension (MRC) task (Li et al., 2020), which can naturally combine domain-related questions with annotated samples, to share all model parameters across domains. Moreover, we propose an entity existence discrimination (EED) task and an entity-aware training setting to recognize inconsistent entity annotations and handle the subtype conflicts. The EED task is designed to determine whether each type of entity exists in a sentence. As for the entity-aware training setting, we utilize the EED task to aware the presence of inconsistent annotated entities in source domain samples, then transform these entities to the target domain-specific entities by leveraging the explicit hierarchical relationship. The above procedures not only alleviate the subtype conflicts but also bring additional reference to better share information across domains. The main contributions of this paper are summarized as follows:

- We propose a novel framework named DoSEA for cross-domain NER, to handle the issues raised by the specific entity types from the target domain.
- In our method, we design an entity existence discrimination task and an entity-aware training setting to alleviate the subtype conflicts, which can identify the presence of each type of entity and transform the inconsistent annotated entities into hierarchical target domain-specific entities in source domain samples.
- Experimental results on six datasets show the superiority of our DoSEA over the state-of-the-art methods.

2 Related Work

MRC for NER Task. The goal of the machine reading comprehension(MRC) task is extracting answer spans from a sentence through a given question. There have been successful attempts to formalize other task as MRC task, such as NER (Li et al., 2020), relation extraction (Li et al., 2019) and event extraction (Liu et al., 2020a). As for the NER task, Li et al. (2020) first formulate NER as MRC to handle both flat and nested NER tasks. Xue et al. (2020) proposed a coarse-to-fine pre-training framework for NER task based on the MRC framework. Zhang et al. (2021) add MRC task in the training process of zero-resource NER task.

Cross-domain NER. Multi-task learning methods (Yang et al., 2017; Wang et al., 2018b; Jia et al., 2019; Jia and Zhang, 2020) have been popular in cross-domain NER, which is used to add the auxiliary task to improve performance. Jia et al. (2019) jointly trains the NER and LM tasks in a parameter generator network. Jia and Zhang (2020) proposed a multi-cell compositional LSTM structure for cross-domain NER, which models each entity type as a separate cell state. Fine-tune methods (Lee et al., 2017; Rodríguez et al., 2018; Cui et al., 2021) also show strong performance, which pre-trained a model in a source domain and then fine-tune the model in a target domain.

Some works try to achieve accurate transfer learning for cross-domain NER (Ruder and Plank, 2017; Wang et al., 2018a; Chen et al., 2021). Wang et al. (2018a) classifies source domain samples by similarity metrics and assigns different weights for training. Chen et al. (2021) proposes a data augmentation approach to transform the data representation across domains. A few works consider the relationship between entity types (Kim et al., 2015; Qu et al., 2016). Qu et al. (2016) considers the mentioned relationship between the source and target entity types, such as "Professor" and "Student". Compared with them, our proposed DoSEA is built on the MRC model, which aims to handle the issues raised by the specific entity types from the target domain.

3 Methodology

Figure 2 shows that DoSEA has three components, including a context encoder, a multi-task layer and an entity-aware training setting. The multi-task layer contains entity existence discrimination(EED) task and entity span prediction(ESP) task. In the entity-aware training setting, we adopt different training processes for source and target domains.

3.1 Problem formulation

Given a sentence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, where n denotes the word length of sentence $\mathbf{x}$. An entity $\mathbf{et}^t = (x^t_{start}, \cdots, x^t_{end})$ is a substring of $\mathbf{x}$ satisfying start≤end, where t represents the entity type. Besides, we define $y_t \in \{0, 1\}$ as the ground-truth of whether the t type of entity exists in sentence $\mathbf{x}$.

Combining with Questions. Given sentence $\mathbf{x}^r$ from domain $r \in \{\mathcal{S}, \mathcal{T}\}$, we need to combine every entity query questions $\mathbf{Q}^r = \{\mathbf{q}^r_1, \mathbf{q}^r_2, ..., \mathbf{q}^r_m\}$ with sentence $\mathbf{x}^r$, where m denotes the number of entity types of domain r. The entity annotation guidelines are used as references to construct questions. In particular, we use questions from the target domain for common entity types between domains. Therefore, we obtain a set of quadruples $(\mathbf{q}^r_t, y^r_t, [\mathbf{et}^{t,r}_1, \cdots, \mathbf{et}^{t,r}_l], \mathbf{x}^r)$ in each domain $r \in \{\mathcal{S}, \mathcal{T}\}$, where l denotes the number of t type of entities in sentence $\mathbf{x}^r$.

Meanwhile, if source domain $\mathcal{S}$ contains the entity type which has a hierarchical relationship with the entity type specific to target domain $\mathcal{T}$, the questions of these entity subtypes are also combined with source domain sentence $\mathbf{x}^{\mathcal{S}}$ for entity-aware training. Thus, we obtain a set of quadruples $(\mathbf{q}^{\mathcal{T}}_{sub_t}, unknow, unknow, \mathbf{x}^{\mathcal{S}})$ in domain $\mathcal{S}$, where $\mathbf{q}^{\mathcal{T}}_{sub_t} \in \mathbf{Q}^{\mathcal{T}}$ denotes the question of entity subtype sub_t.

3.2 Context Encoder

Normally, we combine question $\mathbf{q}^r_t$ and sentence $\mathbf{x}^r$ as a string $\{$`[CLS]`$, \mathbf{q}^r_t,$ `[SEP]`$, \mathbf{x}^r\}$, where `[CLS]` and `[SEP]` are special tokens. Then, the combined string is sent into the input embedding layer. We use BERT (Devlin et al., 2019) as the input embedding layer to generate the contextualized word embeddings. Since question $\mathbf{q}^r_t$ is a natural language sequence that may contain entity examples and disturb accurate entity span prediction, we only retain the sentence embeddings $\mathbf{V} = [\mathbf{v}^d_1, \mathbf{v}^d_2, ..., \mathbf{v}^d_n]$ for the next steps, where d is the output dimension of BERT. To encode sentence-level features, the retained embeddings are fed into a standard bi-directional LSTM layer (Graves and Schmidhuber, 2005). The hidden output of BiLSTM can be expressed as follows:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_i &= \text{LSTM}(\overrightarrow{\mathbf{h}}_{i-1}, v^d_i)\\ \overleftarrow{\mathbf{h}}_i &= \text{LSTM}(\overleftarrow{\mathbf{h}}_{i+1}, v^d_i)\end{aligned} \tag{1}$$

where $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ denote the forward and backward output of BiLSTM. The final representation of a word is $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$

3.3 Multi-task Layer

Given the sentence-level features, the purpose of the multi-task layer is to distinguish the existence of each type of entity and predict entity spans.

Entity Existence Discrimination. To model the relationship between sentence and various entity types, we introduce an entity existence discrimination task to identify whether a sentence contains each type of entity. Logically, this task is sensitive to the semantic feature and certain keywords of the sentence, then aggregates this information to make a particular prediction.

In general, the special characters `[CLS]` without semantic property is often used to represent the semantic features of the whole sentence. However, using the semantic feature alone to predict the existence of entities is not enough, because it lacks connections to the entity types. Therefore, we first use an entity type embedding layer to generate entity type features $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_m]$. Then, to capture the relationship between sentence and entity types, the sentence features and entity type features are incorporated by leveraging the attention mechanism (Vaswani et al., 2017). Given sentence-level features $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$, $\mathbf{h}^{ENT}_j$ represents the incorporated feature associating with entity type t as follow:

$$\mathbf{h}^{ENT}_t = \sum_i^n \alpha_{i,t} \mathbf{W}_v \mathbf{h}_i \tag{2}$$

$$\alpha_{i,t} = \frac{1}{z_t} (\mathbf{W}_q \mathbf{e}_t)^T \mathbf{W}_k \mathbf{h}_i \tag{3}$$

where $\mathbf{W}_q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_k \in \mathbb{R}^{d\times d}$, $\mathbf{W}_v \in \mathbb{R}^{d\times d}$ are parameter matrices and z_t is the normalized factor:

$$z_t = \sum_i^n (\mathbf{W}_q \mathbf{e}_t)^T \mathbf{W}_k \mathbf{h}_i \tag{4}$$

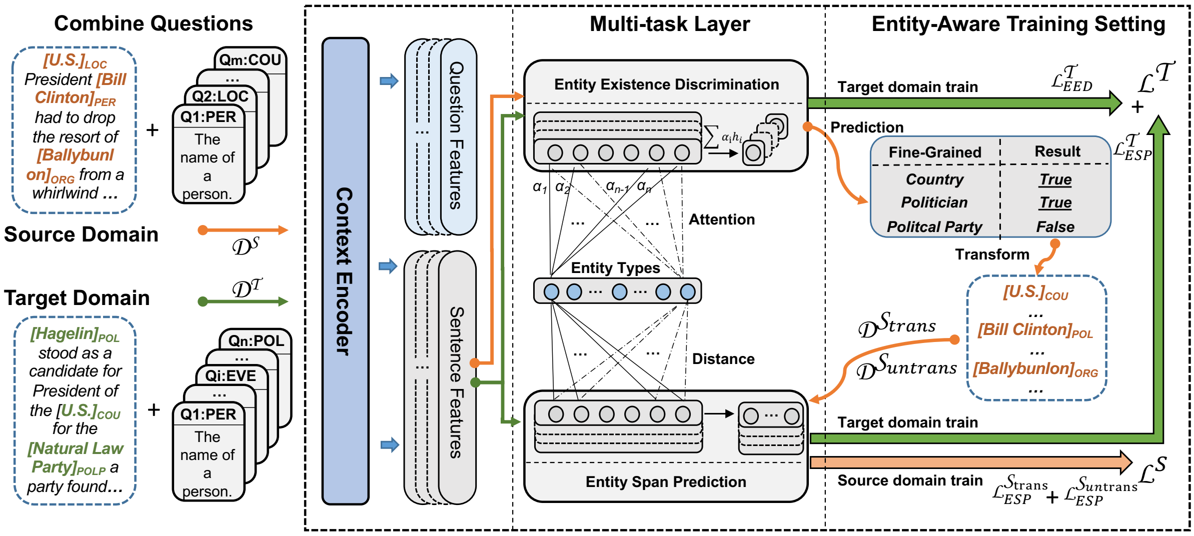

Combine Questions
[U.S.]LOC President [Bill Clinton]PER had to drop the resort of [Ballybunlon]ORG from a whirlwind ...
+
Qm:COU
...
Q2:LOC
Q1:PER
The name of a person.
Source Domain
$\mathcal{D}^S$
Target Domain
$\mathcal{D}^T$
[Hagelin]POL stood as a candidate for President of the [U.S.]COU for the [Natural Law Party]POLP a party found...
+
Qn:POL
...
Qi:EVE
...
Q1:PER
The name of a person.
Context Encoder
Question Features
Sentence Features
Multi-task Layer
Entity Existence Discrimination
$\sum \alpha_i h_i$
α_1 α_2 α_{n-1} α_n
Attention
Entity Types
Distance
Entity Span Prediction
Entity-Aware Training Setting
Target domain train
$\mathcal{L}^{\mathcal{T}}_{EED}$
$\mathcal{L}^{\mathcal{T}}$
+
Prediction
Fine-Grained
Result
Country
True
Politician
True
Politcal Party
False
$\mathcal{L}^{\mathcal{T}}_{ESP}$
Transform
[U.S.]COU
...
[Bill Clinton]POL
...
[Ballybunlon]ORG
...
$\mathcal{D}^{S_{trans}}$
$\mathcal{D}^{S_{untrans}}$
Target domain train
Source domain train
$\mathcal{L}^{S_{trans}}_{ESP}$ + $\mathcal{L}^{S_{untrans}}_{ESP}$
$\mathcal{L}^S$

Figure 2: The Domain-specific entity-aware framework(DoSEA) for cross-domain NER. DoSEA has three components: a context encoder, a multi-task layer and an entity-aware training setting. There are different training processes for source and target domains.

The weight $\alpha_{i,t}$ reflects the degree of relevance between word x_i and entity type t. Finally, a contacted feature $\hat{\mathbf{h}} = [\mathbf{h}^{CLS}; \mathbf{h}^{ENT}]$ is fed into a softmax layer to predict the probability of the existence of entities queried by question q_j:

$$\hat{y}_t = \text{softmax}(\mathbf{W}_c \hat{\mathbf{h}} + \mathbf{b}_c) \tag{5}$$

where $\mathbf{W}_c \in \mathbb{R}^{d \times 2}$, $\mathbf{b}_c \in \mathbb{R}^2$ are parameter matrices.

Therefore, the cross-entropy loss for entity existence discrimination task is denoted as follows:

$$\mathcal{L}_{EED} = -\frac{1}{m} \sum_{t}^{m} y_t \log(p(\hat{y}_t)) \tag{6}$$

Entity Span Prediction. The classical MRC-NER model directly uses word embeddings to predict the start and end positions of entities. However, to enhance the task relationship between the EED and ESP tasks, the entity type embeddings $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_m]$ are shared in both tasks.

Specifically, we improve the entity prediction task in that the start and end positions of entities are predicted by the absolute distance between word embeddings and entity type embeddings. The entity type embedding $\mathbf{e}_t$ is shared as follows:

$$\mathbf{h}_i^{'} = |\text{norm}(\mathbf{h}_i) - \mathbf{e}_t| \tag{7}$$

where norm$(\cdot)$ is instance normalization function (Ulyanov et al., 2016), and $|\cdot|$ means the absolute value. Therefore, $\mathbf{h}_i^{'}$ represents the absolute distance between word and entity type representations. To extract entity spans, $\mathbf{h}_i^{'}$ is fed into two softmax layers to predict the probability of each token being a start or end position of entities queried by question $\mathbf{q}_t$:

$$P_s(y_{i,t}^s | x_i) = \text{softmax}(\mathbf{W}_s \hat{\mathbf{h}}_i^{'} + \mathbf{b}_s) \tag{8}$$

$$P_e(y_{i,t}^e | x_i) = \text{softmax}(\mathbf{W}_e \hat{\mathbf{h}}_i^{'} + \mathbf{b}_e) \tag{9}$$

where $\mathbf{W}_s, \mathbf{W}_e \in \mathbb{R}^{d \times 2}$, $\mathbf{b}_s, \mathbf{b}_e \in \mathbb{R}^2$ are parameter matrices.

As mentioned above, the training samples are a set of quadruples $(\mathbf{q}_t, y_t, [\mathbf{et}_1^t, \cdots, \mathbf{et}_l^t], \mathbf{x}))$ for DoSEA. Meanwhile, entities $[\mathbf{et}_1^t, \cdots, \mathbf{et}_l^t]$ can be paired with two label sequences $[y_{1,t}^s, ..., y_{n,t}^s]$, $[y_{1,t}^e, ..., y_{n,t}^e]$, which represent the ground-truth label of each token x_i being the start position or end position of the entities queried by question $\mathbf{q}_t$. The cross-entropy losses of start and end positions prediction are denoted as follows:

$$\mathcal{L}_s = -\frac{1}{mn} \sum_{t}^{m} \sum_{i}^{n} y_{i,t}^s \log(P_s(\hat{y}_{i,t}^s | x_i)) \tag{10}$$

$$\mathcal{L}_e = -\frac{1}{mn} \sum_{t}^{m} \sum_{i}^{n} y_{i,t}^e \log(P_e(\hat{y}_{i,t}^e | x_i)) \tag{11}$$

The total loss of entity span prediction task is denoted as follows:

$$\mathcal{L}_{ESP} = \mathcal{L}_s + \mathcal{L}_e \tag{12}$$

3.4 Entity-aware Training Setting

As Figure 2 shows, considering the domain-specific entity types, we design different training processes for $\{\mathcal{S}, \mathcal{T}\}$ domains.

Target Domain Training. We adopt normally jointly training process for target domain $\mathcal{T}$, in which the EED task and the ESP task are training together. Therefore, the training loss for target domain $\mathcal{T}$ is defined as follows:

$$\mathcal{L}^{\mathcal{T}} = \mathcal{L}^{\mathcal{T}}_{ESP} + \gamma \mathcal{L}^{\mathcal{T}}_{EED} \tag{13}$$

where γ is auxiliary task weight.

Source Domain Training. In order to avoid learning inaccurate information, we don't train the EED task with samples from source domain $\mathcal{S}$. In contrast, given the source domain samples, the EED task is utilized to recognize the entities with inconsistent annotations between domains. Then we transform these inconsistent entities into hierarchical entities specific to target domain.

Specifically, we can obtian a set of supertype-subtype pair samples as discussed in the previous Section 3.1, in which a pair of the sample consists of $(\mathbf{q}^{\mathcal{S}}_{sup_t}, y^{\mathcal{S}}_{sup_t}, [\mathbf{et}^{sup_t,\mathcal{S}}_1, \cdots, \mathbf{et}^{sup_t,\mathcal{S}}_l], \mathbf{x}^{\mathcal{S}})$ and $(\mathbf{q}^{\mathcal{T}}_{sub_t}, unknow, unknow, \mathbf{x}^{\mathcal{S}})$. In the beginning, we send the subtype sample into the EED task for acquiring the existence prediction result $\hat{y}^{\mathcal{S}}_{sub_t}$. Then, we presume that if the sentence $\mathbf{x}^{\mathcal{S}}$ contains supertype entities and the hierarchical subtype entities are predicted to exist, these supertype entities can be transformed into subtype entities. Finally, we obtain the transformed samples to train the ESP task together, which consists of $(\mathbf{q}^{\mathcal{S}}_{sup_t}, None, \mathbf{x}^{\mathcal{S}})$ and $(\mathbf{q}^{\mathcal{T}}_{sub_t}, [\mathbf{et}^{sub_t,\mathcal{S}}_1, \cdots, \mathbf{et}^{sub_t,\mathcal{S}}_l], \mathbf{x}^{\mathcal{S}})$. The training loss for source domain $\mathcal{S}$ is defined as follows:

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}^{\mathcal{S}_{untrans}}_{ESP} + \delta \mathcal{L}^{\mathcal{S}_{trans}}_{ESP} \tag{14}$$

where δ is the data weight, and the source domain samples are divided into untransformed and transformed parts.

4 Experimental Settings

4.1 Datasets.

We take CoNLL-2003 dataset (Sang and Meulder, 2003) as the source domain for all experiments. We use CrossNER datasets (Liu et al., 2021) and MIT Movie Review dataset (Liu et al., 2013) as the target domain datasets. Statistics of these datasets are shown in Table 1.

Domain	Entity Type	Train.	Dev.	Test.
CoNLL-2003 Dataset				
Newswire	4	15.0K	3.5K	3.7K
CrossNER Datasets				
Politics	9	0.2K	0.5K	0.6K
Science	17	0.2K	0.5K	0.5K
Music	13	0.1K	0.4K	0.5K
Literature	11	0.1K	0.4K	0.4K
Artificial Intelligence	12	0.1K	0.4K	0.4K
MIT Movie Review Dataset				
Movie	12	9.7K	-	2.3K

Table 1: Statistics of datasets.

Hierarchical Entity Pairs. CoNLL-2003 is annotated with "Person", "Location", "Organization" and "Miscellaneous" entities. CrossNER datasets consist of five different domains: Politics, Science, Music, Literature and Artificial Intelligence(AI). Moreover, they all contain four overlapped entity types and hierarchical entity subtypes with CoNLL-2003. Especially, Politics domain contains "Politician", "Political Party" and "Country" entities. Science domain contains "Scientist", "University" and "Country" entities. Music domain contains "Artist", "Band" and "Country" entities. Literature domain contains "Writer" and "Country" entities. AI domain contains "Researcher", "University" and "Country" entities. Entity types in MIT Movie Review and CoNLL-2003 are non-overlapping, but MIT Movie Review contains "Actor", "Character" and "Director" entities which are subtypes of "Person" entities.

4.2 Baseline Methods

In the beginning, we consider a classical method, BiLSTM-CRF (Huang et al., 2015), which combined the bi-directional LSTM network and conditional random fields(CRF) for sequence labeling task. Based on this, there are two improved methods Coach (Liu et al., 2020b) and Multi-Cell LSTM (Jia and Zhang, 2020). Coach proposed a two-step approach for cross-domain NER, it first detects whether the tokens are entities or not, then predicts the specific entity types. Multi-Cell LSTM investigated a multi-cell compositional LSTM model structure, which models each entity type using a separate cell state.

Models	Pol.	Sci.	Mus.	Lite.	AI.	Mov.	Avg.
BiLSTM-CRF	53.89	49.12	43.65	41.87	43.18	77.52	51.54
BiLSTM-CRF-joint†	56.60	49.97	44.79	43.03	43.56	78.11	52.68
Coach†	61.50	52.09	51.66	48.35	45.15	78.59	56.22
BERT-Tagger†	66.56	63.73	66.59	59.95	50.37	79.37	64.43
BERT-Tagger-joint†	68.85	65.03	67.59	62.57	58.57	80.04	67.11
Multi-Cell LSTM(BERT)†	70.56	66.42	70.52	66.96	58.28	82.22	69.16
TemplateNER	65.41	62.93	64.67	64.55	57.64	78.56	65.62
MRC-NER	70.23	67.25	70.64	62.53	62.77	83.28	69.45
MRC-NER-joint	72.37	67.70	71.87	66.67	64.65	85.87	71.52
DoSEA w/o $\mathcal{L}_{EED}$+$\mathcal{L}_{ESP}^{Strans}$	72.41	68.20	71.93	66.74	64.77	86.19	71.71
DoSEA w/o $\mathcal{L}_{EED}$	73.31	70.61	72.55	67.35	65.23	86.74	72.63
DoSEA w/o $\mathcal{L}_{ESP}^{Strans}$	73.46	70.13	72.39	67.89	65.24	86.91	72.67
DoSEA(Ours)	**75.52***	**71.69***	**73.10***	**68.59***	**66.03***	**87.31***	**73.71***

Table 2: Cross-domain experiment results on six domain datasets compared to the baseline methods. † indicates the results on CrossNER datasets are from Liu et al. (2021). "joint" postfix means the model jointly trains on both domains. "w/o" is a abbreviation of "without".

We also compare BERT-based methods, including BERT-Tagger (Devlin et al., 2019) and Multi-Cell LSTM(BERT) method that leverages the outputs of BERT as contextualized word embeddings. As the basic model for our proposed framework, MRC-NER (Li et al., 2020) is considered as a baseline method too. In addition, we compare a fine-tune method named TemplateNER (Cui et al., 2021), which is a template-based method by using BART (Lewis et al., 2020) and also shows effectiveness in cross-domain NER. However, we don't compare our method with Liu et al. (2021), because they continue pre-training the language model BERT with abundant domain-related corpus, which is unfair to compare with each other.

4.3 Implementation details

For all methods, word embeddings are fine-tuned in the training process. When training BiLSTM-CRF and Coach, we use the word-level embedding from Pennington et al. (2014) and char-level embedding from Hashimoto et al. (2017) as the input layer. For BERT-based methods, we use the base-sized BERT pre-trained on the Wikipedia corpus to output contextualized word embeddings.

Since the size of training samples in CrossNER is far smaller than CoNLL-2003, we upsample the training samples in the target domain to keep the balance between domains. In the training step, we set the learning rate as 5e-5, entity type embedding dimension d_h as 768, task weight γ as 0.1 and data weight δ as 0.2.

Domain	Separately			Jointly		
	P	R	F1	P	R	F1
Pol.	83.16	77.82	80.40	85.44	80.31	**82.79**
Sci.	83.84	74.92	79.13	86.39	76.10	**80.92**
Mus.	75.86	71.56	73.65	78.71	75.79	**77.22**
Lite.	80.69	69.35	74.59	84.18	72.15	**77.71**
AI.	78.29	67.99	72.74	80.93	70.03	**75.09**
Mov.	90.87	88.21	89.52	92.56	89.66	**91.09**

Table 3: The entity existence discrimination task results on six domain datasets. "Separately" means the EED task is trained alone. "Jointly" means the EED task is jointly training with the ESP task.

5 Results and Discussion

5.1 Main Results

Cross-domain NER. Table 2 shows the overall performances of the proposed DoSEA against baseline methods on six domain datasets. Our proposed DoSEA significantly outperforms prior state-of-the-art methods on all target domains. To be specific, the F1 scores of DoSEA advance the previous best method by +3.15, +3.99, +1.23, +1.92 and +1.38 pp.(*e.g.*, percentage points.) on Politics, Science, Music, Literature, Artificial Intelligence and Movie domains, respectively.

Compared to the sequence-labeling based methods(e.g., BiLSTM-CRF, Coach, Multi-Cell LSTM and BERT-Tagger), results on the MRC-NER method show the best performance while jointly training across domains. We speculate the main

Model	ELE.	COU.	POLP.	POL.	EVE.	ORG.	LOC.	PER.	MISC.
BERT-Tagger-joint	92.07	69.46	78.09	67.75	45.64	63.12	68.55	13.37	47.11
MRC-NER-joint	92.43	70.46	81.48	69.49	**55.59**	69.49	72.55	49.67	50.03
DoSEA(Ours)	**92.61**	**72.00**	**81.51**	**74.66**	45.97	**72.92**	**81.26**	**58.90**	**50.13**

Table 4: Fine-grained comparisons on the Politics domain dataset. "ELE.", "COU.", "POLP.", "POL.", "EVE.", "ORG.", "LOC.", “PER.” and “MISC.” denote "Election", "Country", "Political Party", "Politician", "Event", "Organization", “Location”, “Person” and “Miscellaneous”, respectively.

reason is that the MRC-NER model parameters are all shared across domains, while the sequence-labeling based methods use an independent CRF layer for each domain. However, our method shows further performance improvement compared with the MRC-NER method. It demonstrates the effectiveness of our proposed DoSEA, which depth alleviates the subtype conflicts between each entity supertype-subtype pair among domains.

Auxiliary Task. We additionally analyze the performance on the entity existence discrimination task, which plays a crucial role in the entity-aware training setting. As shown in Table 3, the F1 scores of the EED task are 82.79, 80.92, 77.22, 77.71, 75.09 and 91.09 in Politics, Science, Music, Literature, Artificial Intelligence and Movie domains, respectively.

Although we only use annotated samples from the target domain to directly train the EED task, this task still achieves quite good performance compared to the results on cross-domain NER. Meanwhile, we also study the task relationship between the EED and ESP tasks, in which we don’t share the entity type features and separately train the EED task. However, the performance suffers a significant decline in all target domains in that case. Therefore, the experimental results indicate the positive effect of the jointly training strategy and provide support for the entity-aware training setting.

5.2 Ablation Study

We conduct ablation studies to explore the effectiveness of each component in the DoSEA framework. To be specific, we consider three settings in the ablation study. (1) We first consider eliminating $\mathcal{L}_{ESP}^{S_{trans}}$ from Eq.14 when using source domain samples to train the DoSEA model. In this case, F1-scores overall target domains suffer a significant decline. (2) To explore the interaction between tasks, $\mathcal{L}_{EED}$ is removed from Eq.13 when using target domain samples to train the DoSEA model. In particular, we use the separate EED model to generate the prediction results about whether the domain-specific subtype entities exist in the source domain samples. Furthermore, the cross-domain NER results suffer a severe drop of about an average of 0.98 pp on the target domains. (3) When we both remove the $\mathcal{L}_{ESP}^{S_{trans}}$ and $\mathcal{L}_{EED}$, DoSEA obtains a similar performance as the MRC-NER method. Eventually, these empirical results suggest that each component in DoSEA is beneficial for cross-domain NER.

5.3 Fined-grained comparisons

To understand the performance of DoSEA at the entity type level, we make fine-grained comparisons on the Politics domain dataset. As mentioned above, we consider three hierarchical entity type pairs in the Politics domain. "Event" to "Election" entity type pair is not considered, because the source domain does not contain the "Event" entity type.

As shown in Table 4, the most interesting result is that the BERT-Tagger method has difficulty in identifying "Person" entities, although there are huge annotated samples for "Person" entities in the source domain. We speculate that the model structure with independent CRF layers seriously hinders the transfer of knowledge from the source domain to the target domain. However, the MRC-NER method achieves relatively high performance on "Person" entity type, which proves the advantages of MRC-NER which shares all model parameters across domains.

MRC-NER method achieves relatively higher performance on "Event" entity type than other methods. We found that the performance on "Event" entity fluctuated greatly during the whole training process. After training data statistic, we think the reason for the instability performance may be that the annotated samples for "Event" entity are very small, only 22 samples. However,

Sentence	In **India**, Prime Ministers **Indira Gandhi** and her son **Rajiv Gandhi** (neither of whom were related to **Mahatma Gandhi**, who was assassinated in 1948), were assassinated in 1984 and 1991 respectively.
Golden labels	**India**: *Country*; **Indira Gandhi**: *Politician*; **Rajiv Gandhi**: *Politician*; **Mahatma Gandhi**: *Politician*
BERT-Tagger-joint	**India**:[*B-location*]; **Indira Gandhi**:[*B-person I-politician*]; **Rajiv Gandhi**:[*B-person I-politician*]; **Mahatma Gandhi**: [*B-person I-person*]
MRC-NER-joint	**India**:[*Location*; *Country*]; **Indira Gandhi**:[*Person*; *Politician*]; **Rajiv Gandhi**:[*Person*; *Politician*] **Mahatma Gandhi**: *Person*
DoSEA(Ours)	**India**: *Country*; **Indira Gandhi**: *Politician*; **Rajiv Gandhi**: *Politician*; **Mahatma Gandhi**: *Politician*

Table 5: The results of an example from Politics domain test dataset.

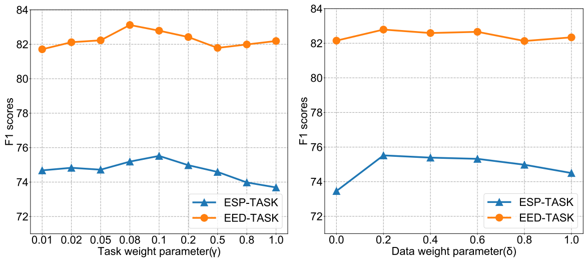


Figure 3: The impact of weight parameters γ and δ on the performance of Politics domain dataset.

our DoSEA accomplishes significant F1-score improvements over all the three hierarchical entity type pairs, demonstrating the effectiveness in reducing the subtype conflicts between domains and providing additional reference about the entity subtypes which are specific to the target domain.

5.4 Case study

Table 5 shows a case study comparing DoSEA with two baseline methods, which is more representative than the others. As we can observe, BERT-Tagger misidentifies `India` as a "Location" entity and also fails to recognize all "Politician" entities. These prediction results hurt the performance of both entity supertype and subtype, which shows a phenomenon of subtype conflict. Meanwhile, the results of the BERT-Tagger method also show label-level mistakes, which presents a challenge to completely identify the correct entities. Since the MRC-NER method can handle the nested entities, it identifies `India` as both a "Location" and "Country" entity, which causes performance degradation on "Location" entity. However, DoSEA correctly identifies all entities in the sentence, which well demonstrates how it can mitigate the subtype conflicts between entity types among different domains.

5.5 Hyperparameter Sensibility

We explore the impact of weight parameter γ in Eq.13 and δ in Eq.14 on Politics domain dataset.

Auxiliary Task Weight. Task weight γ affects the training process for the multi-task inference layer. From Figure 3, we can see that DoSEA keeps a stable F1-scores performance on both entity existence discrimination task and entity span prediction task when $\gamma > 0.01$ and $\gamma < 0.2$, suggesting the stability of the DoSEA. As γ continues to increase, the performance of entity prediction task began to decline, and the best γ parameter is 0.1.

Data Weight. Data weight δ controls how much transfer knowledge the DoSEA model should learn from the transformed subtype entities in the source domain. As we observed, δ have a relatively higher influence on F1 scores of the entity prediction task when $\delta \leq 0.2$ and $\delta \geq 0.6$. Therefore, the reasonable value range for the δ parameter is $\delta \geq 0.2$ and $\delta \leq 0.6$.

6 Conclusion

In this paper, we propose a novel framework named Domain-specific Entity-aware(DoSEA) for cross-domain NER and focus on the issues raised by the domain-specific entity types. Our framework is built on the MRC-NER task, which shares all model parameters across domains. Then, we introduce an entity existence discrimination task and an entity-aware training setting to alleviate the subtypes conflicts, which learns to transform the entities with inconsistent annotations into target domain-specific entities in source domain samples. Experiments show that DoSEA achieves new state-of-the-art performance over six cross-domain benchmarks under the jointly training strategy.

References

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 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*, pages 1923–1933. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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

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. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 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 of the Asian Federation of Natural Language Processing*, pages 473–482. Association for Computer Linguistics.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*.

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

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1340–1350. Association for Computational Linguistics.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020a. Event extraction as machine reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2725. Association for Computational Linguistics.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 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*, pages 19–25. Association for Computational Linguistics.

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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence*, pages 13452–13460.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, and Timothy Baldwin. 2016. Named entity recognition for novel types by transfer learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 899–905. Association for Computational Linguistics.

Juan Diego Rodríguez, Adam Caldwell, and Alexander Liu. 2018. Transfer learning for entity recognition of novel classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1974–1985. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382. Association for Computational Linguistics.

Erik F. 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. Association for Computational Linguistics.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

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

Rui Wang, Masao Utiyama, Andrew M. Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2018a. Sentence selection and weighting for neural machine translation domain adaptation. *ACM*, 26(10):1727–1741.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018b. 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*, pages 1–15. Association for Computational Linguistics.

Mengge Xue, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-fine pre-training for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6345–6354. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Target-oriented fine-tuning for zero-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 1603–1615. Association for Computational Linguistics.