Target-Oriented Fine-tuning for Zero-Resource Named Entity Recognition

Ying Zhang[1]*, **Fandong Meng[2]**, **Yufeng Chen[1]†**, **Jinan Xu[1]** and **Jie Zhou[2]**

[1]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{zhying,chenyf,jaxu}@bjtu.edu.cn
{fandongmeng,withtomzhou}@tencent.com

Abstract

Zero-resource named entity recognition (NER) severely suffers from data scarcity in a specific domain or language. Most studies on zero-resource NER transfer knowledge from various data by fine-tuning on different auxiliary tasks. However, how to properly select training data and fine-tuning tasks is still an open problem. In this paper, we tackle the problem by transferring knowledge from three aspects, i.e., domain, language and task, and strengthening connections among them. Specifically, we propose four practical guidelines to guide knowledge transfer and task fine-tuning. Based on these guidelines, we design a target-oriented fine-tuning (TOF) framework to exploit various data from three aspects in a unified training manner. Experimental results on six benchmarks show that our method yields consistent improvements over baselines in both cross-domain and cross-lingual scenarios. Particularly, we achieve new state-of-the-art performance on five benchmarks.

1 Introduction

Named Entity Recognition (NER) is one of the fundamental tasks in natural language processing. Recently, zero-resource NER draws more and more attention in recent studies (Täckström et al., 2012; Jia et al., 2019; Bari et al., 2020; Liu et al., 2020b; Wu et al., 2020a). This task describes that, in a specific domain or language, there is no labeled training data for NER. Therefore, zero-resource NER severely suffers from data scarcity.

As shown in Figure 1, the ideal training data for zero-resource NER is regarded as the *Targets*, which should satisfy two conditions at the same time: a) in the target domain or language, and b) annotated with NER labels. Thus it is intuitive to

Figure 1: The example of the *Targets* and essential knowledge from three aspects, i.e., Task, Language, and Domain. The middle rectangle denotes an Spanish NER example in news domain, which is referred as the *Targets*. The rounded rectangle above the *Targets* denotes knowledge from different tasks. The bottom left one denotes essential knowledge in Spanish and English languages. The bottom right one denotes knowledge in News and Twitter domains.

augment training data or transfer knowledge from three aspects, i.e., task, language, and domain. The aspect of domain/language can be divided into the source and the target, and the mainstream solution for zero-resource NER is transferring NER annotations from source domains/languages to target ones, e.g., from news to Twitter (Strauss et al., 2016) or from English to Spanish (Bari et al., 2020), where the former is referred as *cross-domain* and the latter as *cross-lingual*.

Based on the mainstream approach, recent researches have conducted further exploration by fine-tuning the contextualized word embeddings on different data. Their results show that only exploiting source labeled data for NER is not enough, due to the discrepancy of domain/language between the source and the target. To alleviate this problem, some studies focus on utilizing a large amount of target unlabeled data to transfer domain- or

---

* Work was done when Ying Zhang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
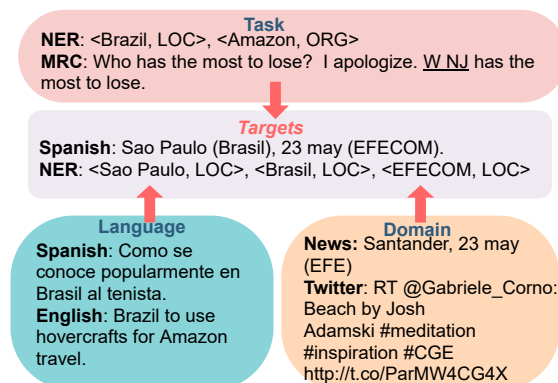† Yufeng Chen is the corresponding author.

language-specific knowledge. AdaptaBERT (Han and Eisenstein, 2019) fine-tunes the masked language model (MLM) on unlabeled data in the target domain (e.g., social media). Both Pfeiffer et al. (2020) and Vidoni et al. (2020) add extra components to learn from unlabeled data in the target language (e.g., Spanish). Besides, Phang et al. (2020) apply non-NER labeled data in the source language (i.e., English) to transfer knowledge for cross-lingual NER, which suggests that annotations for non-NER tasks (e.g., MRC) are useful for NER task. However, they only exploit non-NER annotations in the source language and ignore that in the target languages (e.g., Spanish).

Though the aforementioned studies have improved the performance of zero-resource NER in cross-domain or cross-lingual scenarios, there are two main problems in these methods: a) they conduct knowledge transferring by only considering unlabeled target data and labeled source data, which is insufficient for knowledge transfer. Particularly, they ignore the fact that labeled target data from non-NER tasks is available. b) They fine-tune contextualized word embeddings on various auxiliary tasks in a pipeline manner, where each task is performed only once. We argue that the fine-tuning procedure can not capture enough knowledge from various data when trained only once. Besides, it lacks effective strategies to approach the *Targets* closer. Target at these issues, we suggest it is necessary to exploit more diverse data and design strategies more oriented to the *Targets*.

Therefore, we propose four practical guidelines on how to fully exploit available data to alleviate data scarcity. Concretely, we highlight the necessity of transferring knowledge from three aspects, i.e., task, language, and domain (Guideline-I). Then for domain/language, we pay attention to the gap between the source and target data (Guideline-II). For task, we focus on the gap between non-NER tasks and NER (Guideline-III). Finally, we emphasize the importance of knowledge fusion between the target domain/language and NER task (Guideline-IV). According to our proposed guidelines, we design a target-oriented fine-tuning (TOF) framework for zero-resource NER to approach the *Targets*. This framework applies three tasks (i.e., MLM, MRC, and NER) to capture the knowledge from above three aspects. It enhances the training with MRC task, pseudo data, and continual learning, respectively. To validate the

effectiveness and superiority of our approaches, we conduct experiments on six popular benchmarks for zero-resource NER in cross-domain and cross-lingual scenarios.

Our contributions[1] are summarized as follows:

- We analyze the key factor of zero-resource NER and propose four practical guidelines to transfer knowledge from three aspects, i.e., Task, Language, and Domain, and strengthen connections among them.

- We design a target-oriented fine-tuning (TOF) framework based on our guidelines to exploit more diverse knowledge and approach the *Targets* closer.

- Experimental results verify the effectiveness of our method in both cross-domain and cross-lingual scenarios on six benchmarks. Particularly, our method achieves the state-of-the-art performance on five benchmarks.

2 Background

2.1 Task Definition

The goal of zero-resource NER task is to transfer NER knowledge from labeled source data to unlabeled target data. Therefore, we assume that there are three kinds of data available for training: a) NER labeled source data, b) unlabeled target data, and c) non-NER labeled target data (e.g., MRC).

2.2 Basic Framework

Our method is built on AdaptaBERT proposed by Han and Eisenstein (2019). This network is designed for unsupervised domain adaptation on sequence labeling tasks (e.g., NER). A two step fine-tuning approach is applied in AdaptaBERT, and in this section, we will describe it in detail.

Step-1: Domain Tuning. They fine-tune contextualized word embeddings by training a masked language model (MLM) to reconstruct randomly masked tokens. And this is performed on a dataset containing all available target domain data and an equal amount of unlabeled source domain data.

Step-2: Task Tuning. They fine-tune contextualized word embeddings continually and learn the prediction model for the sequence labeling task. Following (Devlin et al., 2018), they build a strong

---

NER system by simply feeding the contextualized embeddings into a linear classification layer. The log probability can be computed by the log softmax,

$$\log p(y_t|\mathbf{w}_{1:T}) = \beta_{y_t} \cdot \mathbf{x}_t - \log \sum_{y \in \mathcal{Y}} \exp(\beta_y \cdot \mathbf{x}_t),$$
(1)

where contextualized word embedding $\mathbf{x}_t$ captures information from the entire sequence $\mathbf{w}_{1:T} = w_1, w_2, ..., w_T$, and β_y is a vector of weights for each tag $y \in Y = \{PER, ORG, LOC, MISC\}$. They train the model on labeled source domain data by minimizing the negative conditional log-likelihood of labeled data.

3 Our Approach

For zero-resource NER, we firstly analyze the problem of data scarcity. Then we propose four practical guidelines to guide knowledge transfer from different data, which is adapted to both cross-domain and cross-lingual scenarios. According to these guidelines, we design a target-oriented fine-tuning (TOF) framework for zero-resource NER.

3.1 Problem Analysis

The nature of zero-resource NER task is to perform NER with no labeled target domain/language data. And to deal with this task, it is intuitive to transfer essential knowledge from other available data. Concretely, when the data satisfies the two conditions at the same time: a) in the target domain/language (e.g., Twitter/Spanish), and b) annotated for the target task (i.e., NER), we consider it as our *Targets*. While the *Targets* is unavailable under the zero-resource setting, there is abundant data meeting either condition. Therefore, we transfer knowledge from three aspects, i.e., Domain, Language, and Task, as shown in Figure 1.

Domain. It contains knowledge from specific domains (e.g., Twitter). As shown in the bottom right rectangle of Figure 1, '@Garbriele_Corno:' is the special expression that only exists in tweets and '#' is used to highlight something.

Language. It refers to linguistic knowledge in various languages. For example, the word order of 'Como se conoce popularmente en Brasil al tenista' in Spanish is different from its English expression 'As the tennis player is popularly known in Brazil'. Besides, the expressions of 'Brazil' and 'tenista' in English vary from those in Spanish.

Task. It describes hand-crafted annotations for different tasks, which is expensive and difficult to obtain (e.g., NER and MRC). For example, NER labels *LOC* and *ORG* denote names of locations and organizations, respectively. For MRC task in Figure 1, 'W NJ' is annotated as the answer to question 'Who has the most to lose?'.

Furthermore, we divide domain/language aspect into the source and target. Particularly, NER is regarded as the target task for zero-resource NER.

3.2 Four Practical Guidelines

Based on our analysis, we propose four practical guidelines on how to fully exploit available knowledge to alleviate data scarcity.

Guideline-I: It is necessary to exploit available knowledge from domain, language, and task.

Guideline-II: Bridge the gap between source domains/languages and target domains/languages.

Guideline-III: Bridge the gap between annotations for non-NER tasks and NER task.

Guideline-IV: Fuse the knowledge of both the target domain/language and NER task.

3.3 Target-Oriented Fine-tuning Framework

As shown in Figure 2, we design a Target-Oriented Fine-tuning (TOF) framework for zero-resource NER. It contains two components: a) *Knowledge Transfer*, which displays how to transfer not only domain/language but also task knowledge from various data, and b) *Fine-tuning Process*, which demonstrates a flow diagram of the complete fine-tuning process. Both components are designed based on our proposed guidelines, and their relations are illustrated in Figure 2.

3.3.1 Knowledge Transfer

As the right part of Figure 2 shows, to transfer both domain/language and task knowledge for the *Targets*, we consider six kinds of corpora: a) unlabeled NER dataset $D_{t,no}$, b) unlabeled NER dataset $D_{s,no}$, c) labeled MRC dataset $D_{t,m}$, d) labeled MRC dataset $D_{s,m}$, e) unlabeled NER dataset $D_{t,no}$, and f) labeled NER dataset $D_{s,n}$, where {a), c), e)} is in the target domain and {b), d), f)} is in the source domain. Note that e) is the *Targets* without considering labels.

According to *Guideline-I*, since there is no available data that satisfies the *Targets*, it is necessary to transfer knowledge relevant to the *Targets* from other data as much as possible. Apart from source NER labeled data, we not only exploit unlabeled
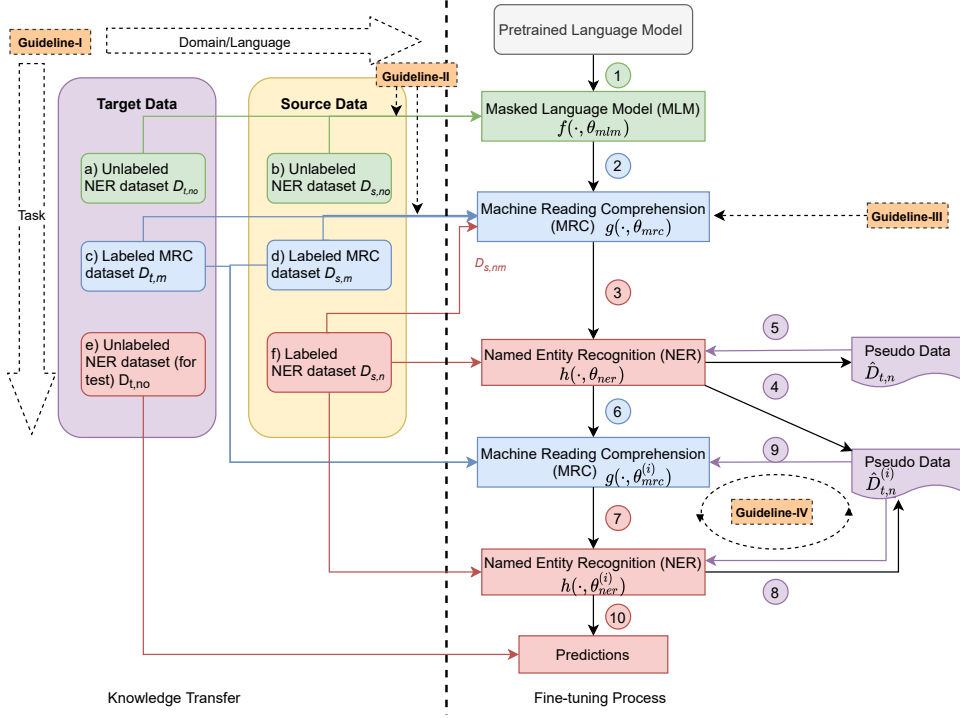
Figure 2: The overall architecture of our Target-Oriented Fine-tuning (TOF) framework. Orange rectangles denote our proposed four guidelines, and dotted lines with arrows denote how they guide *Knowledge Transfer* and *Fine-tuning Process*. Rectangles and solid lines in green, blue, and red color correspond to three fine-tuning tasks (i.e., MLM, MRC, NER), respectively. And the black solid lines with arrows denote the training steps. 'Target Data' and 'Source Data' denote data in target and source domains/languages, respectively. i in the circle denotes *Step-i*.

target data, but also utilize non-NER labeled target data. Therefore, three kinds of data are considered as shown in Figure 2: for 'Target Data', a) unlabeled NER dataset $D_{t,no}$ and c) labeled MRC dataset $D_{t,m}$; for 'Source Data', and f) labeled NER dataset $D_{s,n}$.

According to *Guideline-II*, there is discrepancy between the source and target domain/language. To deal with the gap, it is essential to apply fine-tuning tasks on the mixture of the source and target data. Besides, an effective way to bridge the gap is transforming source data into the target format, e.g., translate the source language data into target language. Therefore, we collect b) unlabeled NER dataset $D_{s,no}$ and d) labeled MRC dataset $D_{s,m}$ in the 'Source Data'.

3.3.2 Fine-tuning Process

Based on AdaptaBERT, we novelly introduce a MRC task between *domain-tuning* and *task-tuning* process. Thus, our fine-tuning process contains three fine-tuning tasks as follows.

Masked Language Model (MLM). To adapt contextualized word embeddings to both the source and target data, we use MLM (Devlin et al., 2018).

Based on *Guideline-II*, We train the model on a mixture of dataset $D_{t,no}$ and $D_{s,no}$. We use the same strategy with (Han and Eisenstein, 2019) to generate 10 random maskings for each instance.

Machine Reading Comprehension (MRC). Based on the *Guideline-III*, we add a span extraction MRC task, which has three advantages: a) MRC can enhance the ability of NER model on span extraction and help NER better capture semantic information of different entity types; b) MRC framework can be used to solve NER task (Li et al., 2020) and it becomes a bridge between NER and other tasks; and c) Recent work on framing other tasks as MRC (Wu et al., 2019; Liu et al., 2020a) provides an idea for transferring knowledge from different tasks with a unified framework. The MRC model is implemented by feeding the contextualized word embedding of each token x_t into two linear classification layers, respectively. The probability of each token being the start or the end index of a span can be computed as follows:

$$
\begin{aligned}
p_t^{start} &= softmax(W_{start} \cdot \mathbf{x}_t), \\
p_t^{end} &= softmax(W_{end} \cdot \mathbf{x}_t),
\end{aligned}
\tag{2}
$$

where W_{start} and $W_{end} \in \mathbb{R}^{d_1 \times 2}$ are learnable parameters, and d_1 denotes the dimensions of contextualized word embedding. Finally, the model is trained by optimizing the Cross-Entropy loss over p_t^{start} and p_t^{end}. According to *Guideline-II*, we train the MRC model on all available MRC data $D_{t,m}$, $D_{s,m}$ and NER data $D_{s,n}$ that is transformed into MRC format $D_{s,nm}$ following (Li et al., 2020).

Named Entity Recognition (NER). To fine-tune contextualized word embeddings continually and learn the prediction model, we feed contextualized word embeddings into a linear classification layer and maximize the probability of each token with the ground-truth entity label. Concretely, given an input token sequence $\mathbf{x} = \{x_i\}_{i=1}^N$ with N words, we firstly feed it into the feature encoder f_θ to obtain contextualized word embeddings $\mathbf{h} = \{h_i\}_{i=1}^N$ for all tokens:

$$\mathbf{h} = f_\theta(\mathbf{x}), \tag{3}$$

where h_i is the feature vector corresponding to the i-th token x_i and f_θ is based on pre-trained language model, i.e., BERT (Devlin et al., 2018), where θ denotes model parameters. Then h_i is fed into a linear classification layer with the *softmax* function to predict the probability distribution of entity labels, which is formulated as follows:

$$p(\hat{y}|x_i) = softmax(Wh_i + b), \tag{4}$$

where $\hat{y} \in Y$ with Y being one-hot vectors corresponding to different entity labels, and $\{W, b\}$ denotes learnable parameters. The loss function is defined as the cross entropy between the predicted probability distribution of each entity label and the ground-truth one for each word. We train NER model on $D_{s,n}$ and predict labels on $D_{t,no}$.

3.3.3 Training

A novel training process is proposed to narrow the gap between the knowledge from available data and the *Targets*, which contains three processes, i.e., MRC enhancing, pseudo data enhancing, and continual learning enhancing.

MRC Enhancing. We fine-tune contextualized word embeddings by sequentially training the MLM $f(\cdot, \theta_{mlm})$, MRC $g(\cdot, \theta_{mrc})$, and NER $h(\cdot, \theta_{ner})$ at *Step-1~3* in Figure 2.

Pseudo Data Enhancing. According to *Guideline-IV*, we use the trained NER model (*Step-3*) to generate pseudo labels on NER unlabeled target data $\hat{D}_{t,n}$ (*Step-4*) and then fine-tune NER model $h(\cdot, \theta_{ner}^{(0)})$ continually on generated pseudo-labeled

target data at *Step-5*.

Algorithm 1 The training process of TOF.

Input: Dataset $D_{t,no}$, $D_{s,no}$, $D_{t,m}$, $D_{s,m}$, $D_{s,n}$, and $D_{s,nm}$; MLM $f(\cdot; \theta_{mlm})$; MRC $g(\cdot; \theta_{mrc})$; NER $h(\cdot; \theta_{ner})$; pre-trained BERT $\theta^{(0)}$; Number of pseudo-data iterations T.

Output: $h(\cdot, \theta_{ner}^{(T)})$.

1: Initialize $\theta_{mlm} = \theta^{(0)}$
2: Fine-tune $f(\cdot, \theta_{mlm})$ on $\{D_{t,no}, D_{s,no}\}$
3: Initialize $\theta_{mrc} = \theta_{mlm}$
4: Fine-tune $g(\cdot, \theta_{mrc})$ on $\{D_{t,m}, D_{s,m}, D_{s,nm}\}$
5: Initialize $\theta_{ner} = \theta_{mrc}$
6: Fine-tune $h(\cdot, \theta_{ner})$ on $\{D_{s,n}\}$
7: Gen pseudo-NER $\hat{D}_{t,n} \leftarrow h(\cdot, \theta_{ner})$ on $D_{t,no}$
8: Initialize $\theta_{ner}^{(0)} = \theta_{ner}$
9: Fine-tune $h(\cdot, \theta_{ner}^{(0)})$ on $\{\hat{D}_{t,n}\}$
10: Gen pseudo-NER $D_{t,n}^{(0)} \leftarrow h(\cdot, \theta_{ner}^{(0)})$ on $\hat{D}_{t,no}$
11: Gen pseudo-MRC $\hat{D}_{t,m}^{(0)} \leftarrow \hat{D}_{t,n}^{(0)}$
12: **for** $i = 1 \rightarrow T$ **do**
13:     Initialize $\theta_{mrc}^{(i)} = \theta_{ner}^{(i-1)}$
14:     Fine-tune $g(\cdot, \theta_{mrc}^{(i)})$ on $\{D_{t,m}, \hat{D}_{t,m}^{(i-1)}\}$
15:     Initialize $\theta_{ner}^{(i)} = \theta_{mrc}^{(i)}$
16:     Fine-tune $h(\cdot, \theta_{ner}^{(i)})$ on $\{\hat{D}_{t,n}^{(i-1)}\}$
17:     Gen pseudo-NER $\hat{D}_{t,n}^{(i)} \leftarrow h(\cdot, \theta_{ner}^{(i)})$ on $D_{t,no}$
18:     Gen pseudo-MRC $\hat{D}_{t,m}^{(i)} \leftarrow \hat{D}_{t,n}^{(i)}$
19: **end for**
20: Predict $h(\cdot, \theta_{ner}^{(T)})$ on $D_{t,no}$
21: **return** $h(\cdot, \theta_{ner}^{(T)})$

Continual Learning Enhancing. We design a continual learning strategy to make full use of pseudo data and imitate the training procedure on the *Targets*. We continually perform fine-tuning between MRC and NER with considering pseudo data (*Step-6~7*), based on the following three considerations: 1) pseudo-labeled target NER data can be refined by the fine-tuned NER model after each iterations, 2) pseudo data is transformed into MRC format, which directly introduces entity type knowledge in target data through definition of MRC questions, and 3) pseudo data participates in both MRC and NER training, which can enhance knowledge connections between two tasks. At *Step-8~9*, we refine pseudo data with newly fine-tuned NER model and take it as training data. After T times iteration, we conduct predictions on unlabeled target data with NER model $h(\cdot, \theta_{ner}^{(T)})$ (*Step-10*).

The training procedure is summarized as 1. $D_{t,no}$ and $D_{s,no}$ demote unlabeled NER data in the target and source domain/language, respectively. $D_{t,m}$ and $D_{s,m}$ denote labeled MRC data in the target and source domain/language, respectively. $D_{t,n}$ and $D_{s,n}$ denote labeled NER data in the target and source domain/language, respectively. Particularly, $D_{t,n}$ is the pseudo-labeled NER data in the target domain/language generated by NER model. And we transformed it into the MRC format, as $D_{s,nm}$. $f(\cdot; \theta_{mlm})$, $g(\cdot; \theta_{mrc})$ and $h(\cdot; \theta_{ner})$ denote the model of MLM, MRC, and NER, respectively. Note that 'Gen' in Algorithm 1 denotes the generalize operation.

4 Experiments

4.1 Data Preparation

We take CoNLL03 for English (en) in the news domain as the source dataset for both cross-lingual and cross-domain tasks.

Cross-Lingual. We consider three NER datasets in target languages: CoNLL03 for German (*de*) (Tjong Kim Sang and De Meulder, 2003), CoNLL02 for Dutch (*nl*) and Spanish (*es*) (Tjong Kim Sang, 2002). All datasets are labeled with 4 entity types: PER, ORG, LOC, MISC. Each of them is split into training, validation and test sets following (Wu et al., 2020b). We use three MRC datasets in target languages: MLQA (*es*) (Lewis et al., 2019), XQuAD (*de*) (Artetxe et al., 2019), and SQuAD (*en*) (Rajpurkar et al., 2016).

Cross-Domain. We use three English datasets in target domains: CBS SciTech News dataset (Jia et al., 2019), short as CBS, in the science and technology news domain, Twitter NER (Zhang et al., 2018b) and WNUT16 (Strauss et al., 2016) in the social media domain. We use two English MRC datasets from news and twitter domains respectively: NewsQA (Trischler et al., 2016) and TweetQA (Xiong et al., 2019). The statistics of datasets are shown in 5 in Appendix A.

4.2 Data Preprocessing

NER datasets are processed in the 'BIO' scheme with four entity types, i.e., PER, LOC, ORG, and MISC except for WNUT16. We perform entity span detection task on WNUT16. Since there are ten entity types annotated in WNUT16, it is different from annotations in source domain/language. For MRC datasets, we transform all of them into a unified format following (Li et al., 2020) for MRC

training. Besides, following (Li et al., 2020), we map the labeled NER datasets to labeled MRC dataset. Concretely, we use the description of each entity for annotators as the query, and each sentence as context. The corresponding answers for each query are entity spans with the same entity type in the sentence. We delete all entity labels on the target data and only use the unlabeled data. We use training and validation sets from the source for training and evaluation, and do predictions on test sets from different target domains/languages.

4.3 Implementation Details

We use BERT-base and multilingual BERT (Devlin et al., 2018) to initialize contextualized word embeddings in cross-domain and cross-lingual scenarios, respectively. We empirically follow the hyperparameter settings of (Han and Eisenstein, 2019) and (Li et al., 2020) except for the learning rate and batch size. Due to the discrepancy between various datasets, we choose the learning rate for Adam (Zhang et al., 2018a) optimizer according to the best performance of checkpoints on the validation set. And the batch size is set to 32, 16 and 64 for MLM, MRC and NER, respectively. More hyperparameters for training procedure are listed in Appendix B.

4.4 Systems

We evaluate following systems by entity-level F1 scores (Sang and De Meulder, 2003). Moreover, we conduct each experiment 5 times and report the mean F1-score.

BERT-ML. Moon et al. (2019) apply the multilingual BERT to cross-lingual NER.

TSL. Wu et al. (2020a) propose a teacher-student learning method for cross-lingual NER.

UniTrans. Wu et al. (2020b) unify data transfer and model transfer for cross-lingual NER.

mCell LSTM. Jia and Zhang (2020) design a multi-cell compositional LSTM for cross-domain NER.

COFEE-MRC. Xue et al. (2020) inject coarse-to-fine automatically mined entity knowledge in a pre-trained language model for cross-domain NER.

AdaptaBERT. Han and Eisenstein (2019) perform *domain-tuning* and *task-tuning* as described in Section 2.2. We take the AdaptaBERT as our baseline in the cross-domain scenario.

AdaptaBERT + translation. Another baseline is set for the cross-lingual scenario. We apply trans-

Systems	cross-lingual				cross-domain			
	es	nl	de	avg	CBS	Twitter	WNUT16	avg
BERT-ML(Moon et al., 2019)[†]	76.53	**83.35**	72.44	77.44	-	-	-	-
TSL (Wu et al., 2020a)[†]	78.00	81.33	75.33	78.22	-	-	-	-
UniTrans(Wu et al., 2020b)[†]	79.31	82.90	74.82	79.01	-	-	-	-
+ ensemble	79.29	83.07	75.55	79.29	-	-	-	-
mCell LSTM(Jia and Zhang, 2020)[†]	-	-	-	-	75.19	-	-	-
COFEE-MRC(Xue et al., 2020)[†]	-	-	-	-	-	54.56	-	-
AdaptaBERT(Han and Eisenstein, 2019)[*]	75.30 (± 0.30)	78.52 (± 0.25)	70.90 (± 0.54)	75.20	75.30 (± 0.37)	65.61 (± 0.46)	63.03[‡] (± 0.23)	67.98
+ translation	76.18 (± 0.13)	80.30 (± 0.52)	72.47 (± 0.75)	76.32	-	-	-	-
TOF(*ours*)	**80.35** (± 0.29)	82.79 (± 0.17)	**76.57** (± 0.16)	**79.90**	**76.41** (± 0.5)	**67.94** (± 0.09)	**67.86** (± 0.27)	**70.74**
w/o continual learning	79.44 (± 0.08)	81.64 (± 0.17)	76.39 ± 0.17	79.16	75.95 (± 0.38)	67.13 (± 0.05)	67.70 (± 0.08)	70.26
w/o pseudo data & w/o continual learning	78.32 (± 1.11)	80.56 (± 0.35)	73.61 ± 0.61	77.50	75.34 (± 0.51)	66.18 (± 0.17)	66.45 (± 0.39)	69.32

Table 1: Results of our method and previous state-of-the-art methods for zero-resource NER in cross-lingual and cross-domain. 'avg' denotes the average of F1 scores (%) on three benchmarks. '†' denotes original results reported in their original papers. '*' denotes results re-implemented by us. Note that '‡' denotes our re-implemented result on WNUT16 and the previous state-of-the-art result on it is 62.8 reported by Han and Eisenstein (2019).

lations of source data[2] to both *domain-tuning* and *task-tuning* of AdaptaBERT.

TOF. Our method is built on two baselines for the cross-lingual and cross-domain scenario, respectively. 'w/o continual learning' denotes the framework without *Step-6~9*. 'w/o pseudo data & w/o continual learning' denotes the framework only performs MRC enhancing at *Step-1~3* of Figure 2.

5 Results and Analysis

5.1 Overall Performance

Table 1 lists main results of our method in contrast with previous state-of-the-art methods in both cross-lingual and cross-domain scenarios.

Cross-Lingual. Our *baseline* on three cross-lingual benchmarks is implemented by training AdaptaBERT with additional translations of source language data, referred as 'AdaptaBERT+translations' in Table 1. Our method achieves significant improvements over *baseline* of F1-scores 4.17, 2.49, and 4.1 for *es*, *nl*, and *de*, respectively. Compared to previous methods, our TOF framework achieves the new state-of-the-art results on two benchmarks *es* and *de*. Besides, Table 1 shows the results of our TOF after removing 'pseudo data' and 'continual learning', respectively, which demonstrates the effectiveness of these two

enhancing strategies. The improvement of our TOF on *nl* (2.49 ↑) is not as good as other two languages (*es*:4.17 ↑ and *de*: 4.1 ↑), which results from the scarcity of MRC data in *nl*. The results well demonstrate the effectiveness of our proposed framework, which benefit from our four guidelines.

Cross-Domain. We regard re-implemented results of AdaptaBERT as our *baseline*, since it not only achieves the state-of-the-art performance on WNUT16, but also outperforms the previous state-of-the-art methods on both CBS and Twitter. Our framework yields obvious improvements over the *baseline* (CBS: 1.11 ↑, Twitter: 2.33 ↑ and WNUT16: 4.83 ↑) and achieves new state-of-the-art results on three datasets. In conclusion, all these results verify the effectiveness and generalizability of our TOF in cross-domain setting.

5.2 Ablation Study

We conduct ablation studies to explore how MRC datasets make difference at step 1~3 in Figure 2. Table 2 highlights the impact of different MRC data in both cross-lingual and cross-domain scenarios.

In the cross-lingual scenario, we consider five kinds of MRC data: 1) 'w/o target MRC data' denoting training without MRC data in the target language; 2) 'w/o source MRC data' denoting training without MRC data in English; 3) 'w/o source MRC data (trans)' denoting without translating the source MRC data into the target language; 4) 'w/o

#	Methods	cross-lingual				cross-domain			
		es	nl	de	avg	CBS	Twitter	WNUT16	avg
0	TOF$_{MRC-enhancing}$	78.32	80.30	73.61	77.41	75.34	66.18	66.45	69.32
1	w/o target MRC data	0.34 ↓	-	0.72 ↓	-	0.44 ↓	1.99 ↓	0.39 ↓	0.93 ↓
2	w/o source MRC data	1.82 ↓	1.26 ↓	1.96 ↓	1.68 ↓	0	0.25 ↓	0.78 ↓	0.54 ↓
3	w/o source MRC data (trans)	2.57 ↓	0.05 ↓	1.65 ↓	1.42 ↓	-	-	-	-
4	w/o NER-MRC data	3.69 ↓	0.83 ↓	1.48 ↓	2.00 ↓	0.88 ↓	0.65 ↓	0.26 ↓	0.59 ↓
5	w/o NER-MRC data (trans)	3.02 ↓	1.01 ↓	1.30 ↓	1.77 ↓	-	-	-	-

Table 2: Ablation study for TOF$_{MRC-enhancing}$, which only performs step 1∼3 in Figure 2. Row 1∼5 list the performance changes compared with Row 0. '↓' denotes the drop of performance.

	cross-lingual				cross-domain			
	es	nl	de	avg	CBS	Twitter	WNUT16	avg
MLM → MRC → NER	78.32	80.30	73.61	77.41	75.34	66.18	66.45	69.32
MRC → MLM → NER	73.17	81.06	73.06	75.76	74.90	65.87	66.13	68.97

Table 3: Results of our TOF framework with different fine-tuning orders.

NER-MRC data' denoting without transforming the NER data into MRC format; and 5) 'w/o NER-MRC data (trans)' denoting without translating the NER-MRC data into the target language.

Results demonstrate that removing any data generally causes a performance drop. Therefore, we draw more in-depth observations as follows. For *es*, 'NER-MRC data' brings the greatest drop of the performance (Row 4). For **nl** and *de*, 'source MRC data' has the greatest impact (Row 2). Besides, 'source MRC data' affects the performance more than 'target MRC data' (Row 2 vs. Row 1). We think it is because 'source MRC data' is twice as much as the target one.

In the cross-domain scenario, since all of three target datasets are in English but in different domains, we do not consider the translated data (Row 3 and 5) in Table 2. Therefore, we conduct ablation studies on three kinds of data: 1) 'w/o target MRC data' denoting training without the target domain MRC data; 2) 'w/o source MRC data' denoting without the source domain MRC data; and 4) 'w/o NER-MRC' data denoting without transforming the NER data into MRC format.

According to the average results in Table 2, we observe that 'target MRC data', 'NER-MRC data', and 'source MRC data' are in descending order of impact. It is intuitive that on Twitter, as shown in Table 2 (Row 1 vs. Row 2 and Row 4), 'target MRC data' has the greatest impact on the performance, when the amount of the target MRC data is greater

than or equal to that of 'source MRC data' and 'NER-MRC data'. However, CBS is affected most by 'NER-MRC data' (Row 4), since its target MRC data are collected from news domain, not science and technology news. For WNUT16, both 'target MRC data' and 'source MRC data' bring more drops than 'NER-MRC data' (Row 1 and Row 2 vs. Row 4). We conjecture that since WNUT16 is an entity span detection task rather than standard NER, it is affected more by the golden MRC data than NER-MRC data.

5.3 Impact of Task Order

We explore the impact of two different sequences for MRC-enhancing, i.e, 'MLM → MRC → NER' and 'MRC → MLM → NER' as shown in Table 3. The results demonstrate that the former outperforms the latter in both cross-lingual and cross-domain scenarios. We conjecture that MLM can capture knowledge of data itself, e.g., domain-specific information and linguistic characters, and MRC captures task-specific information with annotations. Besides, MRC is more relevant to NER than MLM according to task relevance. Therefore, MRC is appropriate to be an intermediate task.

5.4 Comparison with SpanBERT

We replace the pre-trained language model BERT with SpanBERT (Joshi et al., 2020) in the Adapt-aBERT and MRC-enhancing of our TOF to compare the span-enhancing method with ours. The results are shown in Table 4. 1) SpanBERT is

#	Methods	CBS	Twitter	WNUT16	avg
1	AdaptaBERT	75.30	65.61	63.03	67.98
2	+MRC-enhancing	75.34	66.18	66.45	69.32
3	SpanBERT	75.37	67.11	64.17	68.88
4	+MRC-enhancing	75.48	67.46	67.80	70.25

Table 4: AdaptaBERT vs SpanBERT (Joshi et al., 2020) in cross-domain NER.

superior to BERT-base for NER task (Row 3 vs. Row 1). Different from BERT with masking different tokens for each instance, SpanBERT masks a span with several adjacent tokens, which is more related to NER task. 2) 'SpanBERT' underperforms 'AdaptaBERT+MRC-enhancing' on CBS and WNUT16 (Row 2 vs. Row 3), which suggests that although SpanBERT is trained on a large amount of corpus, it is not appropriate for some specific domains. Our MRC-enhancing method uses limited MRC data but achieves more improvements, which shows that MLM can not capture enough task-specific information and it is necessary to introduce other NER-related tasks. 3) Our MRC-enhancing method can make further improvements based on SpanBERT (Row 4 vs. Row 3).

6 Related Work

Zero-resource NER. Some studies (Jia and Zhang, 2020; Pfeiffer et al., 2020; Vidoni et al., 2020) focus on improving architectures of existing models, which add new components into networks to capture specific knowledge, i.e., entity types, language and task characteristics. Different from these methods, our approach only modifies the training procedure without changing model structures. Other studies introduce different auxiliary tasks to alleviate data scarcity (Han and Eisenstein, 2019; Xue et al., 2020; Phang et al., 2020). They are usually based on multi-task learning or two-phrase fine-tuning. Multi-task learning requires balance between the target task and auxiliary tasks, which needs carefully designed objectives. Although two-phrase fine-tuning is effective, it is still inadequate for available data and depends on valid data selection. Our work differs in that we not only propose four practical guidelines to guide data selection and task fine-tuning, but also design a task-oriented fine-tuning framework to exploit more diverse data and target-oriented training strategies.

Data Augmentation. Our approach is inspired by some studies on text classification. Gururangan et al. (2020) utilize unlabeled data in different domains and tasks. Ben-David et al. (2020) exploit unlabeled corpora from multiple domains. Unlike these methods, we focus on target domain/language data with annotations for other tasks, which not only transfers domain/language knowledge, but also utilizes available annotations for other tasks.

MRC for Different Tasks. Although most researches on NER focus on the sequence labeling framework (Huang et al., 2015; Ma and Hovy, 2016; Akbik et al., 2018; Liu et al., 2019), our work is also inspired by formatting other tasks as MRC, such as NER (Li et al., 2020), co-reference resolution (Wu et al., 2019), and event extraction (Liu et al., 2020a). These studies show the superiority and scalability of MRC framework and provide a reference for our work. Different from (Li et al., 2020) using MRC to build a new solution architecture for NER, we exploit MRC to improve the training procedure of NER that is based on sequence labeling. Besides, we perform continual learning between MRC and NER to enhance the impact of MRC on NER.

7 Conclusion and Future Work

In this paper, we analyze the problem of data scarcity in zero-resource NER. To alleviate this issue, we propose four practical guidelines on transferring knowledge from three aspects, i.e., domain, language, and task, and strengthening connections between the source and target data. Based on these guidelines, we design a task-oriented fine-tuning framework to enhance the training procedure with various strategies. Our approach yields significant improvements on six benchmarks and achieves the state-of-the-art on five benchmarks. In the future, we will extend our framework on different target tasks and more task-specific enhancing strategies.

Acknowledgments

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7415–7423.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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*.

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

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings 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 4238–4248, Hong Kong, China. 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, 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, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

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, Online. 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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Yijin Liu, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019. GCDT: A global context enhanced deep transition architecture for sequence labeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2431–2441, Florence, Italy. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020b. Zero-resource cross-domain named entity recognition.

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.

Taesun Moon, Parul Awasthy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with BERT. *CoRR*, abs/1912.01389.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jason Phang, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, Iacer Calixto, and Samuel R Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint arXiv:2005.13013*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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-Volume 4*, pages 142–147.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830.

Marko Vidoni, Ivan Vulić, and Goran Glavaš. 2020. Orthogonal language and task adapters in zero-shot cross-lingual transfer.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020a. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. *arXiv preprint arXiv:2007.07683*.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Coreference resolution as query-based span prediction. *arXiv preprint arXiv:1911.01746*.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A social media focused question answering dataset. *CoRR*, abs/1907.06292.

Mengge Xue, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-fine pre-training for named entity recognition. *arXiv preprint arXiv:2010.08210*.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger B. Grosse. 2018a. Three mechanisms of weight decay regularization. *CoRR*, abs/1810.12281.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018b. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*, pages 5674–5681.

A Statistics

The statistics of all datasets are listed in Table 5. We regard CoNLL03 in English as the source NER data in both cross-lingual and cross-domain scenarios. For target NER datasets, we consider cross-lingual and cross-domain scenarios, respectively.

In the cross-lingual scenario, CoNLL03 in German, CoNLL02 in Spanish, and CoNLL02 in Dutch denote the benchmark datasets in the target languages, i.e., German (*de*), Spanish (*es*), and Dutch (*nl*), respectively. In terms of MRC datasets, we apply MLQA in Spanish and XQuAD in German as labeled MRC datasets in the target languages, i.e., on *es* and *de*. Note that we use the initial validation and test splits in MLQA and XQuAD as the training and validation sets in our work. Since it is difficult to obtain labeled MRC datasets for *nl*, we consider the MRC data in the source language, i.e., English (*en*).

In the cross-domain scenario, CBS SciTech News NER datasets, short as CBS, in science and technology news domain, Tiwtter NER dataset in twitter domainm, and the shared task on entity span detection for WNUT2016 in twitter domain are considered as the target domain NER datasets. All of these three cross-domain benchmarks are in English. We use NewsQA in news domains for MRC fine-tuning on CBS, due to lack of available MRC data in science and technology news domain. TweetQA is applied to both Twitter and WNUT16 NER as the MRC data in the target domain.

B Hyperparameters

In the training procedure, we fine-tune each MLM model for 3 epochs, MRC for 6 epochs, and NER

Benchmark	Task	Language	Domain	Training	Validation	Test
CoNLL03	NER	English(en)	News	14987	3466	3684
Cross-Lingual						
CoNLL03	NER	German(en)	News	12705	3068	3160
CoNLL02	NER	Spanish(es)	News	8323	1915	1517
CoNLL02	NER	Dutch(nl)	News	15806	2895	5195
MLQA	MRC	Spanish(es)	Multi-domain	5254	500	-
XQuAD	MRC	German(de)	Multi-domain	1190	428	-
SQuAD	MRC	English(en)	Multi-domain	10000	1000	-
Cross-Domain						
CBS SciTech	NER	English(en)	Science-technoloy news	-	-	2000
TwitterNER	NER	English(en)	Social-media	4000	1000	3256
WNUT16	ESD*	English(en)	Social-media	2394	1000	3856
NewsQA	MRC	English(en)	News	92550	5167	5127
TweetQA	MRC	English(en)	Social-media	10692	1086	1979

Table 5: Dataset statistics. '*' denotes the entity span detection task.

Methods		cross-lingual			cross-domain		
		es	nl	de	CBS	Twitter	WNUT16
MRC-enhancing	MRC	2e-6	1e-6	1e-6	2e-5	2e-5	2e-5
MRC-enhancing	NER	5e-5	2e-5	2e-5	5e-5	2e-5	5e-5
Pseudo data	NER	2e-5	1e-5	1e-5	8e-5	1e-6	1e-6
Continual learning	MRC	5e-5	2e-6	2e-5	8e-6	3e-5	1e-6
Continual learning	NER	5e-5	5e-6	1e-6	3e-5	5e-6	2e-6

Table 6: Learning rate of MRC and NER model on different datasets.

for 6 epochs. Besides, we perform continual learning for one iteration and achieve the best performance. To re-implement our results easily, we set the seed to a fixed value as 2019 following (Han and Eisenstein, 2019). The learning rate of MLM on each dataset is set to 5e-5. The learning rate of MRC and NER model on different datasets is listed in Table 6. To find the proper learning rate, we perform the hyperparameter search on a set of learning rete values, i.e., 1e-6, 2e-6, 5e-6, 8e-6, 1e-5, 2e-5, 3e-5. 5e-5, 8e-5. And we choose hyperparameter values according to the best validation performance. We train our model on one NVIDIA Tesla P40 (24GB). The average runtime of our TOF framework on different datasets varies from 6 hours to 2 days, due to the data size of different datasets. Other hyperparameters are set following (Han and Eisenstein, 2019) and (Li et al., 2020).