Toward Fully Exploiting Heterogeneous Corpus: A Decoupled Named Entity Recognition Model with Two-stage Training

Yun Hu[12] [*], **Yeshuang Zhu**[3], **Jinchao Zhang**[3], **Changwen Zheng**[12], **Jie Zhou**[3]
[1] Institute of Software, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Pattern Recognition Center, WeChat AI, Tencent Inc, China
{yunhu2016,changwen}@iscas.ac.cn
{yshzhu,dayerzhang,withtomzhou}@tencent.com

Abstract

Named Entity Recognition (NER) is a fundamental and widely used task in natural language processing (NLP), which is generally trained on the human-annotated corpus. However, data annotation is costly and time-consuming, which restricts its scale and further leads to the performance bottleneck of NER models. In reality, we can conveniently collect large-scale entity dictionaries and distantly supervised data. However, the collected dictionaries are lack of semantic context and the distantly supervised training instances contain large noise, which will bring uncertain effects to NER models when directly incorporated into the high-quality training set. To address the above issue, we propose a BERT-based decoupled NER model with two-stage training to appropriately take advantage of the heterogeneous corpus, including dictionaries, distantly supervised instances, and human-annotated instances. Our decoupled model consists of a Mention-BERT and a Context-BERT to respectively learn from the context-deficient dictionaries and noised distantly supervised instances at the pre-training stage. At the unified-training stage, the two BERTs are trained together on human-annotated data to predict the correct labels for candidate regions. Empirical studies on three Chinese NER datasets demonstrate that our method achieves significant improvements against several baselines, establishing the new state-of-the-art performance.

1 Introduction

Named entity recognition is a fundamental Natural Language Processing task that labels each word in sentences with predefined types, such as Person (PER), Location (LOC), Organization (ORG), etc. The results of NER can be used in many downstream NLP tasks, e.g., relation extraction (Bunescu and Mooney, 2005), information retrieval (Chen et al., 2015), and question answering (Yao and Van Durme, 2014). Supervised methods are mainstream approaches to NER, including CRF (Lafferty et al., 2001) and neural network models (Collobert et al., 2011; Lample et al., 2016; Ma and Hovy, 2016). Recently, large-scale pre-trained language models fine-tuned upon a limited amount of annotated data achieve competitive or better performance in NER task (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019b).

Supervised NER methods require a sufficient amount of sentence-level annotated data, even for the methods using pre-trained language models. However, obtaining sentence-level annotated data is expensive and thus leads to small training data size and performance bottleneck of supervised models. In practice, entity dictionaries (or gazetteers) and unlabeled corpora can be obtained at a low cost. Furthermore, distantly supervised data can be automatically generated by matching the unlabeled data against entity dictionaries. These data can form a heterogeneous corpus, which has potential to improve the NER task. However, dictionaries contain only entity mentions without context, and distantly supervised data can be highly noisy in terms of wrong labels and wrong boundaries. As a result, it is unwise to treat dictionaries and distantly supervised data equally to human-annotated ones.

To better utilize heterogeneous corpus, we propose a BERT-based decoupled NER model with two-stage training. The decoupled model decouples mention information and context information with a *Mention-BERT* and a *Context-BERT*, which can better exploit the information in entity data and distantly supervised data respectively. In the pre-training stage, the Mention-BERT can be pre-trained using the entity dictionary with a classifica-

[*] Joint work with Pattern Recognition Center, WeChat AI, Tencent Inc. Work was done when Yun Hu was intern at WeChat AI.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1641–1652

tion task, and the Context-BERT can be pre-trained using the distantly supervised data with two auxiliary tasks (masked language modeling task and classification task). During inference, the decoupled model can utilize the mention information and context information together to make the final prediction. We evaluate our methods on three Chinese NER datasets. Experimental results show that our method outperforms baseline methods and achieves the best results, demonstrating the effectiveness of our methods. The contributions of our work can be summarized as follows [1]:

- We propose a decoupled NER model with two-stage training, which can fully exploit heterogeneous corpus consisting of dictionaries, distantly supervised instances, and human-annotated instances.
- Our model achieves the state-of-the-art results on three common Chinese NER datasets, significantly outperforming current SOTA by 1.51% on OntoNotes and 1.7% on Weibo, as well as obtaining a slight but noticeable gain on MSRA.

2 Background

2.1 Named Entity Recognition

The task of named entity recognition is to find entities in sentences with predefined types, such as PER, LOC, and so on. Given an input sentence $X = \{x_1, x_2, ..., x_n\}$ where x_i denotes the i-th token, and a predefined tag set Y, NER can be modeled as a sequence labeling or region-based classification task. In sequence labeling approaches, the model aims to assign a label $y \in Y$ to each token x_i. In region-based approaches, the model examines each candidate region $\{x_i, x_{i+1}, ..., x_{i+k}\}$ and attempts to assign a label $y \in Y$ to it, where i is the starting position of the region in sentence, and k is the length of the region. Our model follows the framework of region-based approaches.

2.2 BERT-NER Model

Recently, large-scale pre-trained language models, such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), are widely used in NLP and yield state-of-the-art performances on many tasks. Pre-trained language models follow a two-stage paradigm. They are first pre-trained on large-scale unlabeled texts via self-supervised tasks such as masked language modeling and next sentence prediction, and then fine-tuned on relatively small labeled data of downstream tasks.

BERT-NER model is easily adapted from pre-trained BERT model and can achieve competitive performance. Given a sentence X, BERT first outputs the sentence representation $H = \{h_1, h_2, ..., h_n\}$, where h_i is the representation of token x_i. Then, H is passed through a feed forward network (FNN) to obtain the label sequence $\{y_1, y_2, ..., y_n\}$:

$$y_i = softmax(W \cdot h_i + b) \quad (1)$$

where W, b are parameters of the FFN, and y_i is the predicted label of x_i.

Our model is built on top the BERT model. Compared with the BERT-NER, we propose a new decoupled architecture to better utilize heterogeneous data. Besides, different from the training tasks of BERT, our model introduces task-aware pre-training tasks into a two-stage training framework.

3 Approach

3.1 Model Architecture

Generally, an effective NER model should capture two types of information for determining an entity, i.e., mention information and context information. In traditional NER models, the mention and context information are typically coupled in annotated data. Our proposed model decouples the two types of information, making them to be more explicit and easily learned from the heterogeneous corpus.

Overview. As shown in Figure 1, our model consists of three main parts: a *Mention-BERT*, a *Context-BERT*, and a *Global-Classifier*. The input is a sentence along with a region denoting a mention candidate. The model will decouple the mention from the context and feed the two parts into the Mention-BERT and the Context-BERT respectively. Then, the outputs of the two BERTs will be concatenated and passed through the Global-Classifier to obtain the final label prediction. Additionally, the two BERT outputs are also passed through a mention-focused and a context-focused classifier respectively to provide auxiliary supervision during training, which we will elaborate later.

Mention-BERT. The Mention-BERT is used to capture the representation of the mention that to-be recognized. The input of the Mention-BERT

[1] Our code and data are available at https://github.com/huyun-cs/Decoupled_NER

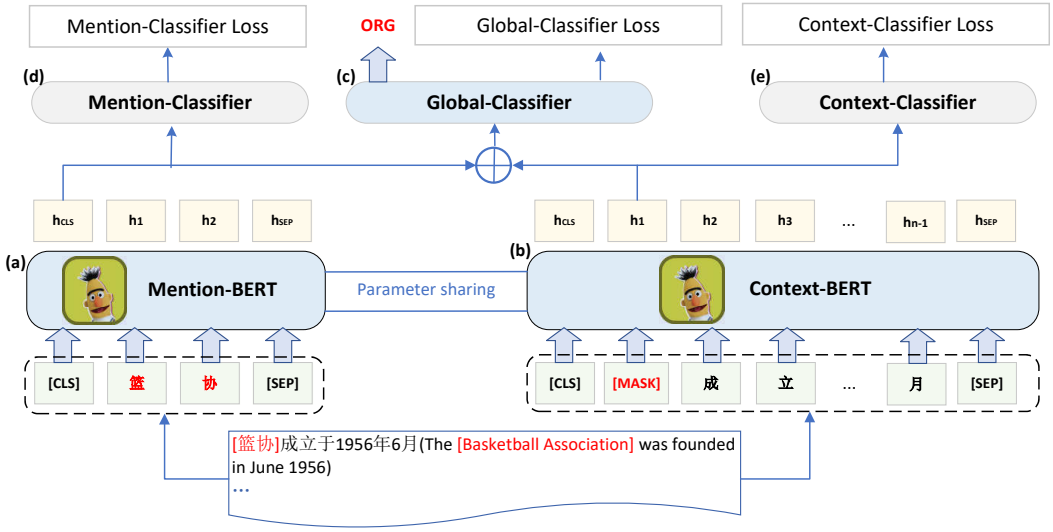

Figure 1: Model architecture. The model consists of a Mention-BERT, a Context-BERT, and a Global-Classifier. The input sentence, "篮协成立于1956年6月" (*the Basketball Association was founded in June 1956*), will first be converted into a 〈MENTION, CONTEXT〉 pair: 〈"篮协" (*Basketball Association*), "[MASK]成立于1956年6月" ([MASK] *was founded in June 1956*). Then the mention and the context act as the input of the Mention-BERT and the Context-BERT respectively. The outputs of the two BERTs will be concatenated and passed to the Global-Classifier to obtain the final tag prediction (ORG).

is an entity mention in the input sentence, and the output is the representation of the mention. The architecture of the Mention-BERT is the same as the original BERT, which is a multi-layer bidirectional Transformer encoder. As shown in Figure 1(a), given an entity mention $m = \{x_i, x_{i+1}, ..., x_{i+k}\}$, we first add two special tokens ([CLS] and [SEP]) to the beginning and the end of it, and take the output corresponding to the [CLS] token as representation h_m of the mention:

$$h_m = \text{Mention-BERT}(m) \quad (2)$$

Context-BERT. The Context-BERT aims to encode the context around an entity mention. It has the same architecture as the Mention-BERT. The input c is just the context of the candidate mention, where the mention is replaced by a special [MASK] token. The output corresponding to the [MASK] token position is used as a representation for the context, denoted as h_c:

$$h_c = \text{Context-BERT}(c) \quad (3)$$

As an example, in Figure 1(b), we have $h_c = h_1$. Note that at inference time we use only one [MASK] even for multi-token entities, as the Context-BERT are not allowed to not use any information of the mention.

Global-Classifier. The Global-Classifier determines the input mention's tag by considering both the mention representation and the context representation. In the implementation, we concatenate the output of Mention-BERT h_m and the output of Context-BERT h_c and pass them into a FFN:

$$y_g = softmax(W_g \cdot [h_m : h_c] + b_g) \quad (4)$$

where W_g, b_g are parameters of the Global-Classifier, and y_g is the final prediction.

3.2 Two-stage Training

Pre-trained language models such as BERT aim to model general patterns of language and treats entity and non-entity words indiscriminately. It is reasonable to expect that such models will not generate a perfect representation for the NER task. To better utilize external heterogeneous data for the NER task, we design a two-stage training framework: (1) pre-training the Mention-BERT and the Context-BERT on entity dictionaries and distantly supervised data, and (2) training the unified model on human-annotated data.

Heterogeneous Training Data. Despite the limited size of human-annotated data for NER, we can easily collect large-scale entity dictionaries and unlabeled text corpora, and hence generate distantly

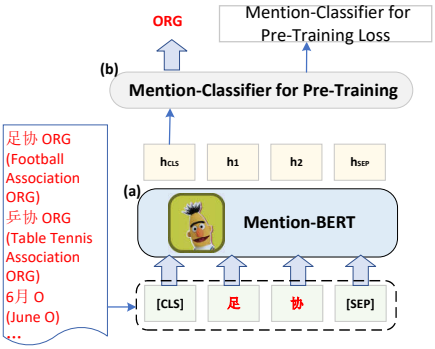

Figure 2: Pre-training process of the Mention-BERT. The Mention-BERT is pre-trained on the entity dictionary using a label classification task. For example, we try to predict that "足协" (*Football Association*) on its own is an organization.

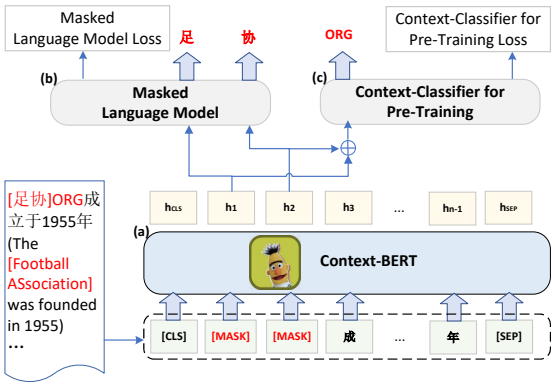

Figure 3: Pre-training process of the Context-BERT. The model is trained on a masked language modeling task and a label classification task simultaneously. For example, we train the model to predict the masked tokens as "足协" (*Football Association*) and the entity label as ORG, given only the context "[MASK]成立于1955" ([MASK] *was founded in 1955*).

supervised data. For dictionary data, the text often contains rich entity structure information. For example, a person name often consists of the First name and the Last name. For distantly supervised data, the text often contains rich context information, while has high noise. The most common mistakes are wrong labels and wrong boundaries. As a result, these data are not suitable to be directly incorporated for NER. However, they can be naturally used as data for pre-training to learn high-coverage and task-aware representations of entity mentions and contexts. On the one hand, previous research showed that further pre-training BERT to do language modeling on in-domain corpus could improve the performance of downstream tasks (Gururangan et al., 2020). On the other hand, either the entity or the context itself can be a strong indicator of entity types.

Mention-BERT Pre-Training. To better capture the regularity information of entities, the Mention-BERT is pre-trained on entity dictionaries. As shown in Figure 2, we add a feed forward classifier denoted as *Mention-Classifier for Pre-Training* on top of the Mention-BERT. The task is to classify each input term into the most probable label according to the dictionaries. For example, the output for the term "足协" (*Football Association*) should be ORG. Besides, to empower the model to learn discriminative representations for non-entity terms as well, we sample items from a common dictionary that have never been seen in any one of the entity dictionaries, and assign an O label to them.

Context-BERT Pre-Training. As shown in Figure 3, the Context-BERT is pre-trained on distantly supervised data with a hybrid task of masked language modeling and entity label prediction. For each input sentence, we pick one entity mention in it at each time and replace all tokens in it with [MASK] tokens. Given only the context with the mention masked out, the model is trained to predict both the masked tokens along with the entity label. We also randomly pick some non-entity regions and assign O labels to them. To this end, we use two classifiers, namely the *Masked Language Model* and the *Context-Classifier for Pre-Training*. The Masked Language Model is the same as in the original BERT. The Context-Classifier for Pre-Training is fed with the average pooling of the Context-BERT's outputs for all masked tokens.

Unified-Training. After pre-training, we perform the unified-training, in which the pre-trained Mention-BERT and Context-BERT are put together and further trained on human annotated data. To construct training examples, we iterate over all entity mentions in the annotated sentences and obtain pairs of 〈MENTION, CONTEXT〉 as the input of our model (see Figure 1). We also select non-entity regions as O. Given the correct label y, we define the loss of the Global-Classifier L_g as follows:

$$L_g = CE(y_g, y) \tag{5}$$

where CE is the cross-entropy loss.

Furthermore, to avoid catastrophic forgetting for

the pre-trained Mention-BERT and Context-BERT in the unified-training, we also add two auxiliary feed forward classifiers on top of Mention-BERT and Context-BERT, denoted as *Mention-Classifier* and *Context-Classifier* respectively (see (d) and (e) in Figure 1). Both of them have the same structure and objective as the Global-Classifier except input:

$$y_m = softmax(W_m \cdot h_m + b_m) \quad (6)$$
$$y_c = softmax(W_c \cdot h_c + b_c) \quad (7)$$

where W_m, b_m, W_c, b_c are parameters of the Mention-Classifier and the Context-Classifier, and y_m, y_c are the respective predictions. We define losses for the two classifiers as follows:

$$L_m = CE(y_m, y) \quad (8)$$
$$L_c = CE(y_c, y) \quad (9)$$

The final loss L of our model at the unified-training stage has three parts:

$$L = L_g + \alpha L_m + \beta L_c \quad (10)$$

where $\alpha, \beta \in [0, 1]$ are hyper-parameters.

4 Experiment

4.1 Dataset

We evaluate our methods on three Chinese NER datasets: OntoNotes 4.0 (Weischedel et al., 2013), MSRA (Levow, 2006), Weibo NER (Peng and Dredze, 2015; He and Sun, 2017). OntoNotes and MSRA are collected from newswire text, and Weibo NER is from social media text [2]. The detail of the datasets is shown in Table 1. At the unified-training stage, we treat all labeled entities in the training dataset as entity mentions. To obtain non-entity mentions, we take (1) all words and phrases labeled as noun by the LTP [3] (Che et al., 2020) lexicon tool, and (2) all words and phrases with an edit distance less than one to any of the entity mentions. During the test time, we first use LTP and SoftLexicon (Ma et al., 2020) model to obtain all regions of candidate entities. Then we use our model to predict the final label for each region.

[2]For Weibo dataset, we only focus on the subset of entities labeled as NAM, as the criteria for entity definition are the same with OntoNotes and MSRA.

[3]http://ltp.ai/

Dataset	Type	Train	Dev	Test
OntoNotes	Sentence	15.7k	4.3k	4.3k
	Char	491.9k	200.5k	208.1k
MSRA	Sentence	46.4k	-	4.4k
	Char	2169.9k	-	172.6k
Weibo	Sentence	1.4k	0.27k	0.27k
	Char	73.8k	14.5k	14.8k

Table 1: Statistics of the datasets.

4.2 Pre-training Corpora

Entity Dictionary. There are four types of entity in our experiment: PER, ORG, GPE, and LOC. We extend the dictionary used in Ding et al. (2019) with more gazetteers collected from Sougou Dictionary [4] and Baidu Dictionary [5]. Finally, our gazetteer contains 50k person names, 143k organization names, 43k geopolitical entities, and 33k location names (see Appendix 1.1).

Distantly Supervised Data. The entity dictionary above is used to match unannotated sentences to obtain distantly supervised data. For OntoNotes and MSRA dataset, we collect news documents on the People's Daily [6] published from 1949 to 2010. For the Weibo dataset, we use the Weibo unannotated data from Peng and Dredze (2015). Finally, we obtain 893k sentences of distantly supervised data for news and 837k for Weibo.

4.3 Training Setting

Some hyper-parameters for training can be found in Appendix 1.2. We set the $\alpha = 0.5, \beta = 0.5$ in unified training through experiments. To better utilize the common knowledge of the Mention-BERT and Context-BERT, and also to reduce the model size, the parameters of Mention-BERT and Context-BERT are shared. We do not share the parameters of each classifier, because the label sets and output dimensions of the classifiers may be different across the two-stage training.

4.4 Baselines

We use the following models as baselines:

BiLSTM-CRF from Lample et al. (2016), which is a classical baseline for NER.

[4]https://pinyin.sogou.com/dict/

[5]https://shurufa.baidu.com/dict

[6]http://paper.people.com.cn

Lattice LSTM from Zhang and Yang (2018), which uses a dictionary and word embedding to enhance character-based Chinese NER model.

BERT-NER from Devlin et al. (2019), which uses the outputs from the last layer of BERT model as feature representations, and does token classification to extract entity.

Incomplete-NER from Jie et al. (2019), which is based on BERT-CRF and uses cross-validation to estimate the distribution of missing labels in distant supervision [7].

MRC-NER from Li et al. (2020b), which considers NER as machine reading comprehension.

SoftLexicon from Ma et al. (2020), which proposes a simple but effective method for incorporating the word lexicon into the character representations in Chinese NER.

FLAT from Li et al. (2020a), which uses transformer to consider the relation between every character and word in the sentence.

ERNIE from Sun et al. (2019), which enhances BERT through knowledge integration by using a entity-level masked LM task and more raw text from the Web resources.

CoFEE from Xue et al. (2020), which proposes a NER-specific pre-training framework to inject coarse-to-fine automatically mined entity knowledge into pre-trained models.

4.5 Main Results

Following the evaluation metrics in previous work, entity-level (exact entity match) standard micro Precision (P), Recall (R), and F1 score are used to evaluate the results.

Table 2 presents the comparison between our model and baseline models. We can observe that our decoupled model with two-stage pre-training significantly outperforms recent models, establishing a new state-of-the-art for supervised NER. For OntoNotes, our model outperforms the SoftLexicon model by +1.51% in terms of F1. For Chinese MSRA, the proposed method outperforms the FLAT model. We also improve the F1 from 70.94% to 72.64% on Weibo dataset. We can also see that the Mention-BERT pre-trained on entity dictionary outperforms the plain decoupled model without two-stage pre-training by 0.89% in OntoNotes, 0.52% in MSRA, and 1.55% in Weibo respectively. These results show the effectiveness of Mention pre-training for the NER task. The results also show that Context-pretraining can improve performance (0.46% in OntoNote, 0.34% in MSRA, and 0.57% in Weibo). Moreover, further pre-training the Context-BERT based on Mention BERT using distantly supervised data can lead to a performance gain in F1 score(0.89% in OntoNote, 0.52% in MSRA, and 1.55% in Weibo).

OntoNotes			
	P	**R**	**F**
BiLSTM-CRF (Lample et al., 2016)	68.79	60.35	64.30
Latice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-NER (Devlin et al., 2019)	78.01	80.35	79.16
Incomplete-NER (Jie et al., 2019)	79.18	81.24	80.20
MRC (Li et al., 2020b)	82.98	81.25	82.11
SoftLexicon (Ma et al., 2020)	83.41	82.21	82.81
FLAT (Li et al., 2020a)	-	-	81.82
CoFEE (Xue et al., 2020)	82.50	82.78	82.64
Decoupled model	83.79	83.06	83.43
+ Mention Pre-train	84.34	83.54	83.93
+ Context Pre-train	84.28	83.51	83.89
+ Mention and Context Pre-train	**84.92**	**83.72**	**84.32**
MSRA			
	P	**R**	**F**
BiLSTM-CRF (Lample et al., 2016)	90.74	86.96	88.81
Latice-LSTM (Zhang and Yang, 2018)	93.57	92.79	93.18
BERT-NER (Devlin et al., 2019)	94.97	94.62	94.80
Incomplete-NER (Jie et al., 2019)	95.00	94.83	94.91
ERNIE (Sun et al., 2019)	-	-	95.0
MRC (Li et al., 2020b)	96.18	95.12	95.75
SoftLexicon (Ma et al., 2020)	95.75	95.10	95.42
FLAT (Li et al., 2020a)	-	-	96.09
Decoupled model	96.65	94.56	95.59
+ Mention Pre-train	96.67	**95.24**	95.95
+ Context Pre-train	96.67	95.20	95.93
+ Mention and Context Pre-train	**97.00**	95.23	**96.11**
Weibo			
	P	**R**	**F**
BiLSTM-CRF (Lample et al., 2016)	-	-	46.11
Latice-LSTM (Zhang and Yang, 2018)	-	-	53.04
BERT-NER (Devlin et al., 2019)	-	-	65.77
Incomplete-NER (Jie et al., 2019)	-	-	66.78
SoftLexicon (Ma et al., 2020)	-	-	70.94
Decoupled model	**72.81**	69.44	71.09
+ Mention Pre-train	70.35	**73.61**	71.94
+ Context Pre-train	71.54	71.78	71.66
+ Mention and Context Pre-train	72.14	73.14	**72.64**

Table 2: Results on the three datasets.

[7] We use the code from https://github.com/ZhuiyiTechnology/AutoIE. We combine human annotated data and distantly supervised data of equal size for training

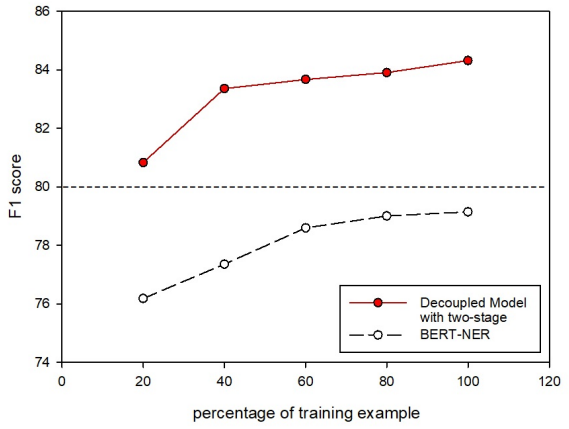

Figure 4: Model performance on OntoNotes with different proportions of hand-annotated training data size.

OntoNotes			
	P	**R**	**F**
BERT-NER	78.01	80.35	79.16
BERT-NER with mixed data	60.24	45.67	51.95
BERT-NER with two-stage	74.71	80.83	77.64
SoftLexicon	83.41	82.21	82.81
SoftLexicon with mixed data	73.38	43.41	54.55
SoftLexicon with two-stage	82.78	81.84	82.30
Decoupled model with two-stage	**84.92**	**83.72**	**84.32**
MSRA			
	P	**R**	**F**
BERT-NER	94.97	94.62	94.80
BERT-NER with mixed data	73.27	53.93	62.13
BERT-NER with two-stage	95.61	93.06	94.32
SoftLexicon	95.75	95.10	95.42
SoftLexicon with mixed data	77.64	54.08	63.75
SoftLexicon with two-stage	95.04	94.38	94.70
Decoupled model with two-stage	**97.00**	**95.23**	**96.11**
Weibo			
	P	**R**	**F**
BERT-NER	-	-	65.77
BERT-NER with mixed data	49.67	31.02	38.19
BERT-NER with two-stage	59.85	70.83	64.87
SoftLexicon	-	-	70.94
SoftLexicon with mixed data	51.34	33.12	40.26
SoftLexicon with two-stage	71.36	62.50	66.64
Decoupled model with two-stage	**72.14**	**73.14**	**72.64**

Table 3: Results of models using external data.

5 Analysis

5.1 Effect of Introducing External Data

In our experiments, it is not immediately clear which part is responsible for the final improvement: can it be the decoupled model or more additional data or both? To answer this question and show our model design can better utilize the heterogeneous corpus, we choose BERT-NER and SoftLexicon as base models to explore the effect of external data. For each base model, we experiment on two settings. First, we simply expand the training dataset by adding entity dictionary data and distantly supervised data. Second, we adopt a two-stage training strategy similar to the methods in Section 3.2, where we use the large external data to further pre-train the BERT part of BERT-NER and SoftLexicon, and then fine-tune the whole models on human-annotated data. The results are shown in Table 3. Our decoupled model achieves the best results. We can see a large performance drop when directly incorporating external training data in BERT-NER and SoftLexicon, as the distantly supervised data are noisy and its big size is unbalanced with the human-annotated data. Unexpectedly, the two base models also perform worse in the two-stage training setting. We suppose that the pre-training task of span classification is not suitable for the sequence labeling task.

5.2 Effect of Human-annotated Data Scale

To compare performances under different numbers of human-annotated training sentences, we randomly select different numbers of training sentences for training on the OntoNotes dataset.

As shown in Figure 4, our model has better performance than the BERT-NER model, which shows the effectiveness of our methods in small training data. Surprisingly, the results also show that in small data size (20% training data), our model also outperforms the BERT-NER model with full data size, which shows that our model requires less sentence-level annotated data compared with the original BERT-NER model. In addition to the model structure and external data, there are two other factors that lead to greater improvement. First, the 20% training data still contain over 3k examples in the news domain. Second, we leverage the mention boundary prediction from LTP, which provide high-quality candidates.

We also experiment on an even smaller training data size (only 1k sentences). In Table 4, we can see that our model performs better than BERT-NER on all the datasets.

5.3 Effect of Model Parameter Sharing

In practice, we share our model parameters of Mention-BERT and Context-BERT. In Table 7, we can see that the model with parameter sharing slightly outperforms the model without parameter sharing. A possible reason is that common knowl-

OntoNotes			
	P	**R**	**F**
BERT-NER	67.61	76.39	71.73
Decoupled model with two-stage	**82.82**	**68.78**	**75.15**
MSRA			
	P	**R**	**F**
BERT-NER	84.59	88.20	86.36
Decoupled model with two-stage	**93.85**	**83.04**	**88.12**
Weibo			
	P	**R**	**F**
BERT-NER	59.51	65.28	62.26
Decoupled model with two-stage	**66.04**	**65.74**	**65.89**

Table 4: Model performance on 1k human annotated training data.

	OntoNotes	**MSRA**	**Weibo**
Sharing	**84.32**	**96.11**	**72.64**
Without Sharing	84.07	96.05	72.07

Table 5: The effect of parameter sharing.

edge about NER is shared between the two BERTs.

5.4 Case Study

为改善[九江]$_{\text{GPE}}$投资环境
In order to improve the investment environment of [Jiujiang]$_{\text{GPE}}$
BERT-NER LOC
Golden / Our model GPE
在[东盟]$_{\text{ORG}}$成立30周年之际
On the 30th anniversary of the [Association of Southeast Asian Nations]$_{\text{ORG}}$
BERT-NER GPE
Golden / Our model ORG

Table 6: Case study. Our model refers to the decoupled model with two-stage training. The text in brackets is the candidate mention, followed by the golden label. Predicted labels in red denote wrong answer.

Table 6 shows two cases from OntoNotes. In the first example, the BERT-NER model misclassifies "九江"(*Jiujiang*) as LOC. We find that "九江" (*Jiujiang*) is in our dictionaries but the label is GPE. Benefited from incorporating entity dictionaries into pre-training, our model can correctly recognize "九江" (*Jiujiang*) as a city. In the second example, the BERT-NER misclassifies "东盟"(*Association of Southeast Asian Nations*) as GPE. We find that distantly supervised data contains the sentence, "在上海合作组织成立5周年大会上"(*At the 5th anniversary meeting of the Shanghai Cooperation Organization*) and the context of "上海合作组织" (*Shanghai Cooperation Organization*) is similar to "东盟" (*Association of Southeast Asian Nations*). The label of the "上海合作组织" (*Shanghai Cooperation Organization*) is GPE. With the context information from Context-BERT, our model can obtain the correct answer of "东盟" (*Association of Southeast Asian Nations*).

6 Related Work

6.1 Supervised NER Models

NER models trained on human-annotated data often achieve appropriate performance. Sequence labeling methods are widely used in NER. Traditional methods use the CRF model to solve the NER task (Lafferty et al., 2001). With the advantages of eliminating feature engineering and significant performance improvement, neural network models become prevalent in NER research, e.g., the models based on FFN (Collobert et al., 2011), CNN (Ma and Hovy, 2016), LSTM (Lample et al., 2016), and pre-trained language model (Devlin et al., 2019). Recent work also propose different ways to model the NER task other than sequence labeling, such as machine reading comprehension (Li et al., 2020b), dependency parsing (Yu et al., 2020), span classification (Sohrab and Miwa, 2018). Generally, these approaches have achieved promising results but heavily rely on human-annotated data.

6.2 Enhancing NER with External Data

Entity dictionaries or gazetteers have long been regarded as an easily-obtainable and useful resource for NER. Previous methods commonly incorporated gazetteers as additional features (Ghaddar and Langlais, 2018; Al-Olimat et al., 2018; Liu et al., 2019a; Ding et al., 2019; Lin et al., 2019; Rijhwani et al., 2020). For languages without explicit word boundaries, such as Chinese, incorporating a universal dictionary with common words besides gazetteers can be further helpful for NER (Zhang and Yang, 2018; Liu et al., 2019b; Sui et al., 2019; Gui et al., 2019b,a; Ma et al., 2020; Li et al., 2020a; Jia et al., 2020). Dictionaries can also be used to construct distantly supervised data from unlabeled corpora. Previous work on reducing the noise in distantly supervised data include new labeling schemes (Shang et al., 2018), reinforcement learning (Yang et al., 2018), cross-training (Jie et al., 2019), positive unlabeled learning (Peng et al., 2019), HMM (Lison et al., 2020), consensus network (Lan et al., 2020a). In other NLP tasks, such as relation extraction, few works have exploited using both human annotated data and dis-

tantly supervised data together (Angeli et al., 2014; Beltagy et al., 2019). Compared with previous works, our work focus on designing a new model architecture and training approaches to better exploit the heterogeneous data in NER task.

6.3 Two-stage Training Paradigm for NLP

Recently, large-scale pre-trained language models, such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), are widely used and yield state-of-the-art performances in many NLP tasks. These two-stage methods allow using large-scale unlabeled data in pre-training and small labeled data in fine-tuning. In order to adapt to specific tasks or domain, variants of BERT are proposed including small and practical BERT (Tsai et al., 2019; Lan et al., 2020b; Jiao et al., 2020), domain adaptive BERT (Yang et al., 2019a; Gururangan et al., 2020), and task adaptive BERT (Sun et al., 2019; Xue et al., 2020; Jia et al., 2020). Our work performs further pre-training on BERT and proposes task-aware training objectives to improve NER.

7 Conclusion

In this work, we focus on fully exploiting heterogeneous corpus for NER. The corpus consists of entity dictionaries, distantly supervised instances, and human-annotated instances. We propose a decoupled NER model with two-stage training. The model first learns appropriate task-aware representations in pre-training, from large-scale context-deficient dictionaries and noisy distantly supervised data. Then after unified-training, the model can predict entity labels according to both the mention and the context information. Experimental results show our method achieves better performance than previous state-of-the-art methods on three Chinese datasets. In the future, we will exploit more types of data, such as knowledge bases, and extend our approach to other languages.

References

Hussein Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. 2018. Location name extraction from targeted text streams using gazetteer-based statistical language models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1986–1997, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1556–1567. ACL.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining distant and direct supervision for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1858–1867. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: A open-source neural chinese language technology platform with pretrained models. *arXiv preprint arXiv:2009.11616*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for Chinese NER with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1462–1467, Florence, Italy. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*,

pages 1896–1907, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. IJCAI.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. 2019b. A lexicon-based graph neural network for chinese NER. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1040–1050. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, 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, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Hangfeng He and Xu Sun. 2017. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718. Association for Computational Linguistics.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced BERT pre-training for chinese NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6384–6396. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4163–4174. Association for Computational Linguistics.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.

John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Eighteenth International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Ouyu Lan, Xiao Huang, Bill Yuchen Lin, He Jiang, Liyuan Liu, and Xiang Ren. 2020a. Learning to contextually aggregate multi-source supervision for sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2134–2146. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020b. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020a. FLAT: chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6836–6842. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5849–5859. Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun, Bin Dong, and Shanshan Jiang. 2019. Gazetteer-enhanced attentive neural networks for named entity recognition. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6231–6236. Association for Computational Linguistics.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition

without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1518–1533. Association for Computational Linguistics.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019a. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307, Florence, Italy. Association for Computational Linguistics.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019b. An encoding strategy based word-character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2379–2389, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960, Online. Association for Computational Linguistics.

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

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2409–2419. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime G. Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8118–8123. Association for Computational Linguistics.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3828–3838. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical bert models 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 3623–3627.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

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, EMNLP 2020, Online, November 16-20, 2020*, pages 6345–6354. Association for Computational Linguistics.

Huiyun Yang, Shujian Huang, XIN-YU DAI, and CHEN Jiajun. 2019a. Fine-grained knowledge fusion for sequence labeling domain adaptation. 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 4188–4197.

YaoSheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2159–2169. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564. Association for Computational Linguistics.

8 Appendices

8.1 Entity Dictionary

	OntoNotes	**MSRA**	**Weibo**
Coverage rate	71.49	78.81	61.95
Conflict rat	85.35	93.13	83.63

Table 7: Coverage rate and conflict rate of entity dictionary. We use the entity dictionary to directly match the test dataset, and compute the coverage rate and conflict rate. The coverage rate is the number of entities both in the dictionary and in the test dataset divided by the number of entities in the test dataset. The conflict rate is the number of entities with inconsistent labels divided by the number of entities both in the dictionary and in test dataset.

8.2 Hyper-parameter Values

Mention-BERT pre-training			
learning rate			2e-5
batch size			128
epoch			10
Context-BERT pre-training			
learning rate			2e-5
batch size			64
epoch			10
Unified-training			
	OntoNotes	**MSRA**	**Weibo**
learning rate	5e-6	5e-6	5e-6
batch size	64	64	32
epoch	4	4	10

Table 8: Hyper-parameter values