Mulco: Recognizing Chinese Nested Named Entities through Multiple Scopes

Jiuding Yang*
University of Alberta
Edmonton, AB, Canada
jiuding@ualberta.ca

Jinwen Luo*
Platform and Content Group, Tencent
Beijing, China
jamsluo@tencent.com

Weidong Guo*
Platform and Content Group, Tencent
Beijing, China
weidongguo@tencent.com

Jerry Chen
University of Alberta
Edmonton, AB, Canada
jerry3@ualberta.ca

Di Niu
University of Alberta
Edmonton, AB, Canada
dniu@ualberta.ca

Xu Yu
Platform and Content Group, Tencent
Beijing, China
henrysxu@tencent.com

ABSTRACT

Nested Named Entity Recognition (NNER), as a subarea of Named Entity Recognition, has presented longstanding challenges to researchers. In NNER, one entity may be part of a larger entity, which can occur at multiple levels. These nested structures prevent traditional sequence labeling methods from properly recognizing all entities. While recent research has focused on designing better recognition methods for NNER in various languages, Chinese Nested Named Entity Recognition (CNNER) is still underdeveloped, largely due to a lack of freely available CNNER benchmarks. To support CNNER research, in this paper, we introduce ChiNesE, a CNNER dataset comprising 20,000 sentences from online passages in multiple domains and containing 117,284 entities that fall into 10 categories, of which 43.8% are nested named entities. Based on ChiNesE, we propose Mulco, a novel method that can recognize named entities in nested structures through multiple scopes. Each scope uses a scope-based sequence labeling method that predicts an anchor and the length of a named entity to recognize it. Experimental results show that Mulco outperforms state-of-the-art baseline methods with different recognition schemes on ChiNesE and ACE 2005 Chinese corpus.

CCS CONCEPTS

• **Computing methodologies → Information extraction**; **Language resources**.

KEYWORDS

Datasets, Nested Named Entity Recognition, Chinese Nested Named Entity Recognition, Sequence Labeling

---

*These authors contributed equally to this research.

1 INTRODUCTION

Named Entity Recognition (NER), a classic and popular topic in Natural Language Processing (NLP), has been widely studied to benefit a variety of real-world NLP applications, such as recommender systems. By treating NER as a sequence labeling task, researchers have been able to achieve excellent performance using neural-network-based architectures [14].

However, Nested Named Entity Recognition (NNER) has presented new challenges to NER research due to the complex and nested structures of entities that are often observed in most languages [34]. For example, the entity "Toronto" (a location) may be nested within the entities "the City of Toronto" (an organization) and "the political head of the City of Toronto" (a person). Such nested structures make recognition difficult. This problem is also widely present in Chinese Nested Named Entity Recognition (CNNER), which as an important subarea of NNER, plays a crucial role in many popular real-world NLP applications, such as entity linking and information retrieval widely used in recommender systems and search engines. While there are several English NNER (ENNER) datasets, there are few datasets available to support research on NNER for other languages including Chinese, which has limited the CNNER development. Existing CNNER datasets either offer restricted access [5] or contain a low ratio of nested named entities [39]. Additionally, these datasets often lack an official parsing method and test set, leading to inconsistent evaluation metrics [5, 37, 39].

In this paper, we present ChiNesE, a Chinese Nested Named Entity Recognition (CNNER) dataset, constructed to support CNNER research. ChiNesE comprises 20,000 sentences from a variety of domains and contains 117,284 entities that fall into 10 major categories. To the best of our knowledge, ChiNesE is currently the largest (in terms of the number of nested named entities) fully

open-sourced CNNER dataset with an official data split and a unified metric. In ChiNesE, 43.8% of all entities are in nested structures, with a maximum nesting level of 8.

To address the challenges posed by Nested Named Entity Recognition (NNER), various methods have been developed to extract English named entities from nested structures, which can also be adapted to solving Chinese Nested Named Entity Recognition (CNNER). Traditional NER labeling methods such as "IOB" and "BIOES" [14] fail to encode nested named entities. Thus, early studies use multiple sets of labels (e.g., IOB) to distinguish named entities of different depths. Researchers then employ multiple classifiers, each predicting a different set of labels [1], or stacked labels of different sets into joint labels, to recognize all entities at once [6]. However, most of these approaches are limited to solving NNER with a shallow nesting depth, which is insufficient for today's applications.

Recognizing the limitations of these sequence labeling methods, recent research has focused on solving the NNER task using different labeling schemes. For example, transition-based methods [22, 32] construct a binary tree from the input text and predict a sequence of actions to reconstruct the tree. However, action prediction errors can accumulate, particularly in CNNER, where entities may consist of many characters rather than just a few words in ENNER. Span-based methods [13, 30, 38] recognize named entities by scoring all available text pieces in a sentence. They must aggregate the span information of all text pieces for classification, which increases the complexity of their models. Although recent advancements in NNER have shown improvements, most of them have not been extensively tested on CNNER. Additionally, these methods have primarily explored new labeling approaches for NNER rather than leveraging the potential of applying the sequence labeling approach to NNER. However, in real-world NNER applications, ensembles of multiple models using different techniques are commonly employed, where the sequence-labeling-based approach can serve as a valuable complement to enhance performance.

To bridge this gap and address the lack of sequence labeling methods in CNNER development, we propose Mulco, a novel method for recognizing Chinese nested named entities using four different scopes. Based on the ChiNesE dataset, our method utilizes a scope-based sequence labeling scheme that uses an anchor and the length of an entity to encode its position and category. This sequence labeling method encodes entities at the character level, making it easier for models to learn. By combining multiple scopes using simple model structures, Mulco is able to overcome the limitations of traditional sequence labeling methods on CNNER tasks and effectively recognize named entities from nested structures.

To evaluate the performance of Mulco, we also parse the Chinese corpus of ACE 2005 using the existing parsing method for ACE 2005 English NNER [1] and reproduce a number of single-model baseline methods with different labeling schemes on both ChiNesE and ACE 2005 for comparison, including Pyramid [33] and Biaffine [38]. The experimental results demonstrate the superior performance of Mulco compared to all single-model baselines on both datasets. Further experiment confirms its effectiveness as a complementary method for recognizing entities from heavily nested structures when employed in ensemble techniques. Our code and the full

---

[1] https://github.com/nlpcl-lab/ace2005-preprocessing

ChiNesE dataset, along with the parsing script for the ACE 2005 Chinese corpus, are available at https://github.com/XpastaX/Mulco to facilitate future research on CNNER.

2 RELATED WORKS

NER. Traditional Named Entity Recognition (NER) methods primarily focus on flat NER, where no named entity overlaps with others. Early research was mostly rule-based [23, 26], but the development of neural networks led to the incorporation of feature engineering with machine learning [25, 35], treating NER as a sequence labeling task. The availability of more computing resources has further driven the use of deep learning models [14], which often utilize pre-trained word (character) vectors [24] and pre-trained language models [4] to improve performance.

NNER. Nested Named Entity Recognition (NNER) is a subarea of NER in which an entity may contain other entities or be part of larger entities [34]. Like NER, early methods for NNER used rule-based approaches [8, 12]. More recent works on NNER have combined pre-trained language models and word vectors with deep learning methods for improved performance, which can be broadly classified into four categories: layer-based [19, 29, 33], transition-based [22, 32], hypergraph-based [11, 31] and span-based [2, 3, 28, 36, 38]. For example, Wang et al. [33] proposed a layer-based method called Pyramid that stacks flat NER layers, each of which is used to recognize text pieces of varying lengths. Wang et al. [32] designed a transition-based method that labels each sentence into a tree structure and tackles NNER by predicting a sequence of actions to reconstruct the trees. Katiyar and Cardie [11] constructed a hypergraph structure based on "BIOES" sequence labeling tags and used a standard LSTM-based sequence labeling model to learn the nested entity hypergraph structure. Yu et al. [38] treated NER as a parsing problem and developed a span-based method using biaffine attention to consider the possibility of each text span being a mention. In addition, there are also other NNER research which has employed generative approaches [16] or used head words as anchors to facilitate named entity classification [17].

CNNER. Compared to English NNER, Chinese NNER (CNNER) has received relatively little attention in research [7, 15, 41]. Some recent works on CNNER include: Fu and Fu [6] developed a method based on the assumption that the nested depth of most Chinese named entities is no deeper than two. They reformulated CNNER as a cascaded chunking problem on a sequence of words, but their method requires pre-segmentation of Chinese sentences, which is not always provided. Additionally, their method cannot recognize entities with a depth greater than two. Chen et al. [2, 3] proposed region-based methods that first detect the boundaries of named entities and then assemble them into candidates for further recognition. These methods use multiple steps to recognize named entities, which increases inference time. Li et al. [13] proposed a unified NER model that considers CNNER. It is a span-based method that encodes entities into word-to-word relations and predicts the relations between all words in the text to extract nested named entities. It combines the predictions of an MLP predictor and a Biaffine [38] predictor to form an ensemble model that requires more resources to train than other single-model methods.

Table 1: The statics of ACE 2005 and ChiNesE. 43.8% of the entities in ChiNesE are nested , and 45.3% of entities in ACE 2005 are nested. "avg. char." represents the average characters of the sentences in each dataset.

	ACE 2005				ChiNesE			
	train	valid	test	all	train	valid	test	all
#sentence	5,999	727	734	7,460	175,00	1,000	1,500	20,000
#nested	3,033	387	430	3,850	7,500	990	1,484	9,974
#entity	27,590	3,316	3,812	34,727	95,151	8,833	13,300	117,284
#nested	12,464	1,480	1,801	15,745	36,289	6,075	9,045	51,409
avg. char.	42.9	40.8	45.3	42.9	60.8	65.9	66.7	61.5
max depth	7				8			
#category	7				10			

Table 2: The example entities of category in ChiNesE.

ChiNesE	Example
Person	Jackie Chan
Location	Toronto
Organization	WTO
Time	01/20/2023
Work	Forrest Gump
Food	Banana
Product	iPhone
Medicine	Aspirin
Event	FIFA World Cup
Organisms	Cat, Rose

Table 3: The distribution of the entities in ACE 2005 and ChiNesE.

ACE 2005	Amount	ChiNesE	Amount
Person	14,539	Person	20,510
Location	1,550	Location	32,924
Organization	6,886	Organization	20,194
Geo-Political	9,033	Time	11,583
Facilities	1,662	Work	11,014
Vehicle	683	Food	5,647
Weapon	374	Product	4,352
-		Medicine	3,927
-		Event	3,807
-		Organisms	3,326

CNNER Datasets. Many early CNNER research [6, 40] is developed based on the Chinese corpus provided by People's Daily 1998 [39] and Chinese Tree Bank (CTB) [37]. However, People's Daily only has three categories for named entities, and the number of the nested named entities and their depth is much lower than existing English NNER datasets [5, 27]. CTB is mainly used for Chinese parsing, thus research developed based on it usually requires parsing information to recognize named entities [40]. Recent works [2, 3, 13] parse ACE 2004 and ACE 2005 Chinese corpus [5] and collects a CNNER dataset to develop their methods. However, the detailed parsing and the data split methods are implicit, which makes later researchers hard to follow their works. Moreover, the ACE corpus is not free for access, which could cause difficulties for researchers to study CNNER.

3 DATASET

We have created ChiNesE to support research on **Chi**nese **Nes**ted Named **E**ntity Recognition (CNNER). The goal of CNNER is to recognize all named entities within nested structures in a given sentence. An example of a nested structure containing multiple nested named entities is shown in Figure 1. More information about NNER can be found in Wang et al. [34]. Detailed statics of ChiNesE can be found in Table 1.

Construction of categories. The categories in ChiNesE were chosen from Chinese Wikipedia[2]. We collected the tags of all entities and kept the ones that occurred most frequently. We then manually filtered out labels that had redundant meanings and removed labels that were not well-known to most people. The remaining labels were then grouped into 10 major categories. Table 2 gives an example named entity of each category.

Data collection and annotation. The annotation of nested named entities is difficult and costly. To minimize cost and maintain quality, we constructed ChiNesE in three steps.

First, we collected 5,000 online passages from QQ Browser, a social media application with over 300 million general audience, and used TextRank4ZH[3] to extract key sentences. We then hired three annotators with professional knowledge backgrounds to label all named entities in the sentences. All annotators had received their Bachelor's degrees and had at least one year of experience in NLP data annotation. The agreement between the three annotators was 90.26%. To ensure quality, only annotations that all three annotators agreed upon were kept.

Then, to collect additional samples, we trained a Biaffine NER model [38] with a BERT encoder [4] on the annotated dataset. The model was run on an additional 45,000 sentences extracted

---

[2]https://dumps.wikimedia.org/zhwiki/
[3]https://github.com/letiantian/TextRank4ZH

from different passages. To maintain consistency, the same three annotators labeled these sentences, using the predictions of the trained NNER model as a starting point to improve accuracy. Two annotators each labeled 22,500 sentences and a third annotator checked their work. The third annotator agreed with 93.40% of the labeled entities, and the remaining were removed to ensure quality. Out of the 50,000 total sentences, we kept 10,000 sentences with nested entities to construct ChiNesE, and an additional 10,000 sentences with the largest number of flat named entities to offer more training samples for ChiNesE.

Finally, to ensure accurate metrics, we selected 1,000 and 1,500 sentences with nested named entities from the 10,000 nested sentences to use as our validation set and test set, respectively. We re-trained the Biaffine model using the remaining 17,500 sentences and generated predictions for the 2,500 samples in the validation and test sets. The same three annotators then carefully refined the annotation of these 2,500 sentences, using the predictions of the Biaffine model only to speed up the process. 94.41% labeled entities are agreed by all annotators, and the rest are removed. Notice that, though the validation set and the test set are annotated supported by the predictions from Biaffine, this would not give any advantage to the Biaffine method in the final experiments. The predictions from the Biaffine model were only employed to accelerate the annotation process. Annotators were instructed to add any missing entities and remove any incorrect entities based on the Biaffine model's predictions. It is important to emphasize that the training samples remained unchanged throughout this process.

Table 1 gives the statistics of ChiNesE and the parsed ACE 2005, and Table 3 gives the distribution of their entities. Our dataset is much larger than ACE 2005, with deeper depth and more categories. We sampled the online passages from 27 domains (e.g., Medicine, Sports, E-Sports) based on their click rates. The nested named entities in ChiNesE are more diverse compared to those in ACE 2005: 31% of nested entities in ChiNesE share the last character with another entity, while only 0.05% of nested entities in the parsed ACE 2005 have the same last character with another entity. Additionally, ChiNesE is specially designed for CNNER. Most of the sentences in the validation and test sets have nested named entities, making it an ideal dataset for testing the performance of models on CNNER tasks.

4 METHODOLOGY

In this section, we present Mulco, our proposed approach for addressing the CNNER problem. Mulco utilizes multiple scopes to accurately identify entities from nested structures.

4.1 Recognition through a Scope

The NNER task involves recognizing all entities from a given sentence. Inspired by modern computer vision methods [9], we use scopes to locate named entities in a sentence under a sequence labeling scheme, which recognizes entities by identifying their anchor and predicting its length. For example, we define P as the scope which uses the first token of the named entity to locate it. To use P to locate "Beijing Tiananmen" (entity) in "I am going to Beijing Tiananmen" (sentence), we first find the position of the word "Beijing" (the first token of "Beijing Tiananmen"), which is the 5$^{\text{th}}$

token of the sentence. Next, by knowing the length of "Beijing Tiananmen' is 2, we can locate the last token "Tiananmen", which is the 6$^{\text{th}}$ token of the sentence. The entity "Beijing Tiananmen" can then be extracted from the sentence.

We employ the sequence labeling method to enable the recognition method described above. Specifically, for a given sentence $T = \{t_i\}_{1 \leq i \leq N}$, where t_i represent the ith character of the sentence, and N is the total number of characters. In Chinese, a character acts as a token. We use two sequences of labels to locate entities. The first sequence of labels are the anchor labels $C^{\text{P}} = \{c_i^{\text{P}}\}_{1 \leq i \leq N}$, where c_i^{P} is the anchor label of character t_i for the scope P. We have

$$c_i^{\text{P}} = \begin{cases} \texttt{NA}, & t_i \text{ is not an anchor} \\ \texttt{cate}, & \text{otherwise} \end{cases},$$

where $c_i^{\text{P}} = \texttt{cate}$ if t_i is the first token of an named entity, and $\texttt{cate}$ is the category of the entity with the anchor t_i. $\texttt{NA}$ means "not an anchor".

By predicting the anchor labels, we can locate the starting character of the entities, and determine their categories. However, to fully recognize the named entity, we also need a sequence of the length labels $L^{\text{P}} = \{l_i^{\text{P}}\}_{1 \leq i \leq N}$ to obtain the lengths of the entities, where l_i^{P} is the length label of t_i. We define

$$l_i^{\text{P}} = \begin{cases} 0, & t_i \text{ is not an anchor} \\ z, & \text{otherwise} \end{cases},$$

where $z \in \mathbb{Z}^+$ is the length of the entity which has t_i as its anchor for the scope P.

It is obvious that in flat NER, where no nested named entity is considered, we can recognize all named entities by locating anchors and their lengths with a scope P. However, when recognizing named entities from NNER, using single scope is no longer sufficient. For example, if we also define "Beijing" as a named entity of location while having "Beijing Tiananmen", when we use P as the scope, "Beijing" will then be the anchor of both "Beijing" and "Beijing Tiananmen", and that will cause critical problems on the sequence labeling.

4.2 Recognition through Multiple Scopes

An intuitive way to recognize nested named entities is to employ multiple scopes to handle the nested structures between named entities. Thus, in Mulco, we use four different scopes to solve CNNER problem.

Figure 1 gives an example of how the four scopes work. For convenience, we use B-min, B-max, E-min and E-max to denote the four scopes used in Mulco. B-min, B-max detect whether the current token is the first token of the respective shortest entity and longest entity. Similarly, E-min and E-max determine if a token is the last token in respective shortest entity and longest entity. The example gives a heavily nested structure of name entities. None of the four scopes can find all named entities only by itself. For example, B-min is unable to recognize "Haidian District People's Government", "Haidian District, Beijing Municipality" or "Haidian District People's Government of Beijing Municipality". On the other hand, B-max failed to recognize "Beijing Municipality", "Haidian District" and "Haidian District, Beijing Municipality". However, if we combine

Figure 1: An example of recognizing nested named entities in "Haidian District People's Government of Beijing Municipality".

the two scopes, we can cover most entities in the example except "Haidian District, Beijing Municipality", which is later covered by E-max. By aggregating the results of all the four scopes, Mulco is able to decode the name entity in deeply nested structures (100% of entities in ChiNesE).

4.3 Mulco

With the effective scope-based recognizing method designed above, we are able to solve the CNNER problem in sequence labeling approach. Figure 2 illustrates the structure of Mulco. Given a sentence $T = \{t_i\}_{1 \leq i \leq N}$, we first encode the sentence with

$$E = \{e_i\}_{1 \leq i \leq N} = \text{BERT}(T),$$

where E is the sequence of the character embeddings of T, $e_i \in \mathbb{R}^d$, d is the dimension of the embedding and $\text{BERT}(\cdot)$ extract the BERT embedding of the input sentence.

Since we use a single character to represent the anchor of an entity, the contextual information will be very important for better recognition. Although BERT can aggregate the contextual information of a character using its self-attention mechanic, a Bi-LSTM structure may further improve it. Thus, we employ a multi-layer Bi-LSTM to calculate the contextualized embedding H by

$$H = \{h_i\}_{1 \leq i \leq N} = \text{Bi-LSTM}(E),$$

where $h_i \in \mathbb{R}^m$, m is the dimension of the contextualized embedding and $\text{Bi-LSTM}(\cdot)$ extract the embeddings from the inputs.

Once obtained the contextualized embedding, in each scope, we employ two linear classifiers to predict the anchor label and the length label for t_i. Here we take B-min as an example:

$$c_i^{\text{B-min}} = W_C^{\text{B-min}} h_i + b_C^{\text{B-min}},$$
$$l_i^{\text{B-min}} = W_L^{\text{B-min}} h_i + b_L^{\text{B-min}}, \tag{1}$$

where $c_i^{\text{B-min}} \in \mathbb{R}^{N_c}$ and $l_i^{\text{B-min}} \in \mathbb{R}^{N_l}$ are the ith (one-hot) anchor label and the (one-hot) length label of B-min respectively. $W_C^{\text{B-min}} \in \mathbb{R}^{m \times N_c}$, $b_C^{\text{B-min}} \in \mathbb{R}^{N_c}$, $W_L^{\text{B-min}} \in \mathbb{R}^{m \times N_l}$ and $b_L^{\text{B-min}} \in \mathbb{R}^{N_l}$ are the trainable parameters of the linear classifiers of B-min. N_c represents the number of categories and N_l is a hyper-parameter represents the max length of all entities.

For B-min, denote the predicted sequence of anchor labels as $C^{\text{B-min}} = \{c_i^{\text{B-min}}\}_{1 \leq i \leq N}$, and the predicted sequence of length

labels as $L^{\text{B-min}} = \{l_i^{\text{B-min}}\}_{1 \leq i \leq N}$. By employing the cross-entropy function, the training object for B-min is then

$$\mathcal{L}^{\text{B-min}} = \sum_{i=1}^{N} \ell_i^{\text{B-min}}, \tag{2}$$

where

$$\ell_i^{\text{B-min}} = -\sum_{j=1}^{N_c} \bar{c}_{i,j}^{\text{B-min}} \log c_{i,j}^{\text{B-min}}$$
$$-\sum_{k=1}^{N_l} \bar{l}_{i,k}^{\text{B-min}} \log l_{i,k}^{\text{B-min}}. \tag{3}$$

Here $c_{i,j}^{\text{B-min}}$ and $l_{i,k}^{\text{B-min}}$ are the jth and the kth elements of $c_i^{\text{B-min}}$ and $l_i^{\text{B-min}}$ respectively, while $\bar{c}_{i,j}^{\text{B-min}}$ and $\bar{l}_{i,k}^{\text{B-min}}$ are their corresponding true labels.

The overall training object is then

$$\mathcal{L} = \mathcal{L}^{\text{B-min}} + \mathcal{L}^{\text{B-max}} + \mathcal{L}^{\text{E-min}} + \mathcal{L}^{\text{E-max}},$$

where $\mathcal{L}^{\text{B-max}}$, $\mathcal{L}^{\text{E-min}}$ and $\mathcal{L}^{\text{E-max}}$ are the training objects of the other scopes, obtained with the same method following Equation (1)-(3).

As shown in Figure 2, we extract the predicted named entities of all scopes as the results. We will keep the category with the highest confidence if an entity is recognized by multiple scopes with different categories. Confidence is how the scope believes the entity belongs to a category, which can be found in the prediction of the anchor label.

5 EXPERIMENTS

To assess the effectiveness of Mulco, we conducted evaluations on both the ChiNesE and ACE 2005 datasets. We compared our proposed method with several state-of-the-art non-ensemble methods that employ different approaches for recognizing nested named entities. Furthermore, we performed experiments on an ensemble of Mulco and a span-based method to demonstrate the complementary nature of Mulco, as a sequence labeling method, when combined with a span-based approach for recognizing entities from heavily nested structures. In line with previous research on NER, we report standard precision (P), recall (R), and micro F1-score (F1) to evaluate performance.
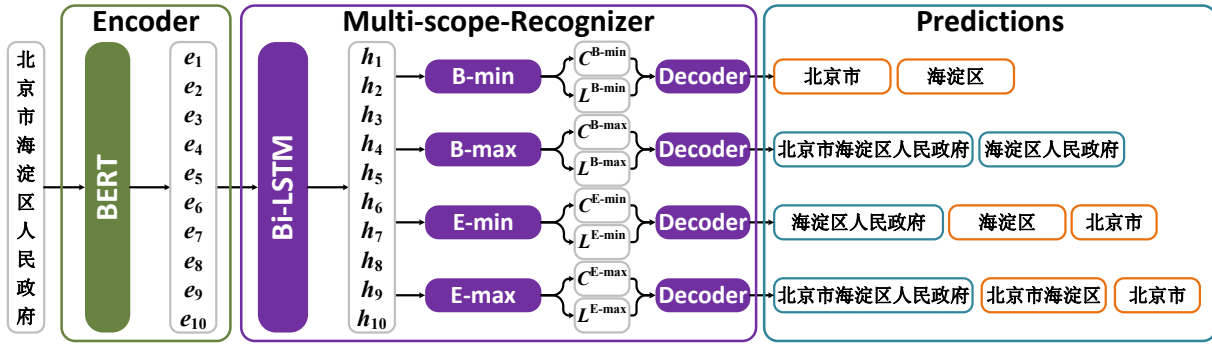
Figure 2: The structure of Mulco. The example is from Figure 1

5.1 Datasets

In addition to ChiNesE, we also use ACE 2005 [5] to further evaluate the performance of all baselines. ACE 2005 is a corpus consisting of broadcasts, newswires, and weblogs in Arabic, Chinese, and English, and is widely used to develop NER methods. However, while some research has been conducted using the Chinese portion of ACE 2005 for CNNER, there is no commonly used division of the data into training, validation, and test sets. Therefore, we follow the approach of [10, 21] and parse the Chinese corpus ourselves, dividing it into training, validation, and test sets in an 8:1:1 ratio. We then split the documents into sentences to form the samples. Detailed statistics are shown in Table 1 and Table 3.

5.2 Baseline Methods

We have used Innermost and Outermost as baselines and applied the traditional "BIOES" sequence labeling method [14, 18] which only recognizes the shortest and longest named entities respectively. Additionally, we have evaluated four state-of-the-art models that are based on different labeling methods. To ensure a fair comparison, we have only reproduced the methods for which valid official codes were available. Furthermore, we have implemented a simple vote-based ensemble method called "Mulco+Biaffine', which combines the predictions of two best performed models: Mulco and Biaffine. This ensemble technique allows us to gain insights into how the performance can be influenced by combining predictions made by different labeling approaches. Below, we provide a detailed introduction to each baseline method.

Pyramid [33] is a layer-based method that stacks interconnected layers to predict whether a text segment of a certain length belongs to an entity. The length of the text segment (n-gram) detected by each layer increases with the depth of the layer, resulting in a pyramid-shaped structure in the model.

Wang et al. [32] propose a scalable transition-based method, in which each sentence is transformed into a tree consisting of its words. They design actions to represent the nested structures of entities, treat the words in a sentence as buffers, and attempt to reconstruct the tree by predicting a sequence of actions.

Biaffine [38] is a span-based method that predicts all possible regions in a sentence to find entities. For each region, a representation of the start and end tokens is generated to represent the information of the entire text piece within the region when predicting.

Shen et al. [28] also proposed a span-based method. They were inspired by object detection techniques in computer vision. They began by generating span seeds and then used regression to adjust the boundaries of the spans. In a second step, they classified the recognized entities.

W^2NER [13] is a span based method. The authors propose a word to word relation classification architecture which can encode all named entities with a matrix of word to word relations. Two separate predictors (Biaffine and MLP) are employed to learn to predict the relation matrix. The results are summed up to calculate the final prediction.

5.3 Experimental Setting

All experiments were conducted on a single NVIDIA RTX 3090 GPU. For a fair comparison, we fine-tuned BERT based on "bert-base-chinese"[4] and only used the output BERT embedding as the input for all models.

For Mulco, we used 4-layer and 2-layer Bi-LSTM for ACE 2005 and ChiNesE, respectively. The dimension of the contextualized embedding was set to 1536. We used AdamW [20] as the optimizer, and set the maximum entity length to 512, the same as the maximum length of the input text. The number of parameters for all settings are less than 133 millions.

Table 4: The experimental settings of Mulco.

Parameters	Region	ACE 2005	ChiNEsE
batch	16	16	16
BERT lr	1e-5, 2e-5	2e-5	1e-5
other lr	1e-4, 2e-4	2e-4	1e-4
weight decay	5e-2, 1e-2, 5e-3	5e-2	5e-2
dropout	0.1, 0.3, 0.5	0.5	0.3

Following the literature, we perform grid-search and report the evaluation result with the best validation performance. Details are provided in Table 4. Specifically, we trained Mulco for 50 epochs with a batch size of 16 on both datasets. The learning rate for BERT was 0.00002 for ACE 2005 and 0.00001 for ChiNesE, while the learning rate for the other trainable parameters was 0.0002 for ACE 2005 and 0.0001 for ChiNesE. The dropout rates were 0.5 for

---

[4]https://huggingface.co/bert-base-chinese

ACE 2005 and 0.3 for ChiNesE, and the weight decay factors were 0.05 for both datasets.

For Innermost and Outermost, we trained them for 50 epochs with a batch size of 16 on both datasets. The learning rate was set to 2e-5, with a dropout rate of 0.5 and a weight decay factor of 0.05. For the other baseline methods, we conducted fine-tuning of BERT with a learning rate of 0.00001, unless specified otherwise in the original experimental settings of the baselines. The remaining parameters were kept the same as their original experimental configurations.

5.4 Results

We mainly compare the performance between non-ensemble methods. Table 5 shows the results of all experiments. As expected, the Innermost and Outermost methods achieved the highest precision on both datasets with their traditional sequence labeling approach. However, this labeling scheme cannot handle nested named entities, resulting in much lower recall rates compared to the other baselines. The recall rates of Innermost are much higher than those of Outermost because when nested, there may be multiple short named entities within a long named entity.

The transition-based method [32] did not perform well on either dataset. It generates actions to decode nested named entities and uses special labels to link two adjacent words belonging to the same entity. This works well in English NER, as named entities typically consist of a few words. However, in CNNER, named entities usually consist of many characters, which presents additional challenges for the model.

Pyramid [33] and Biaffine [38] showed similar performance on both datasets. Pyramid performed better when the ratio of nested named entities was not very high (in the test set of ACE 2005), while Biaffine, as a span-based method, performed better when the entities were heavily nested (in the test set of ChiNesE). The span-based method proposed by Shen et al. [28] performed better than all other baselines on ACE 2005, but not as well as Biaffine and Pyramid on ChiNesE, which has a much higher ratio of nested entities in its test set.

Our proposed method achieved the best overall performance (F1) on both datasets, outperforming the other models by at least 1.12 on ACE 2005 and 0.93 on ChiNesE. By using multiple scopes, Mulco achieved higher recall rates than all other baseline methods on both datasets, which were 1.54% and 2.15% higher than all other models. This is because each scope is specifically designed for a certain type of named entity, and by aggregating the predictions from all scopes, more potential named entities can be retrieved from the given sentence. Although this results in lower precision, the overall performance is greatly improved. Additionally, since NER is one of the first steps in information retrieval, a higher recall rate may benefit downstream tasks.

While our primary focus is on addressing the lack of sequence labeling methods on CNNER through the development of Mulco, we also conducted experiments to gain insights into how ensemble techniques can affect performance. Ensemble techniques are commonly employed to enhance task performance. W^{2}NER leverages these advantages and outperforms all non-ensemble baselines. However, it is noteworthy that Mulco, when used alone, achieves comparable performance on the ACE 2005 dataset and demonstrates competitive performance on the ChiNesE dataset, even when compared to the ensemble method W^{2}NER. To further evaluate the impact of Mulco, as a sequence labeling method, on real-world CN-NER applications, we introduce an ensemble version of Mulco and Biaffine utilizing a voting mechanism. In this ensemble approach, a named entity is considered a prediction if it receives a vote from either the Mulco or Biaffine model.

For the ACE 2005 dataset, the ensemble model "Mulco+Biaffine" performs worse than "Mulco" alone. This is primarily due to the Biaffine model exhibiting low precision on the ACE 2005 dataset. Consequently, during the voting process, it produces many negative predictions, which negatively impact the overall performance. In contrast, the Biaffine method demonstrates significantly higher precision on the ChiNesE dataset, thereby benefiting the performance of the ensemble model.

The discrepancy in precision between the two datasets can be attributed to the fact that Biaffine, as a span-based method, performs better in predicting nested named entities compared to flat named entities. This is particularly evident in the case of ChiNesE, which exhibits more complex nested structures compared to ACE 2005. As detailed in Section 3, ChiNesE has been specifically designed to cater to the needs of CNNER. Almost all sentences in the validation and test sets of ChiNesE contain nested structures, while ACE 2005 does not. This also emphasizes the significance of constructing the ChiNesE dataset, as it provides high-quality data for CNNER development and enables researchers to tackle the challenges associated with nested named entities effectively, advancing the field of CNNER.

Furthermore, the distinctive recognition approaches employed by Biaffine and Mulco make them complementary to each other, particularly in scenarios with heavily nested entities. As a result, their ensemble model surpasses the performance of W^{2}NER on ChiNesE, where the heavily nested entities is in common. This highlights the importance of developing Mulco, as it provides complementary predictions from the perspective of sequence labeling methods, thereby facilitating real-world applications of CNNER that often involve ensembles of multiple models. Moreover, the high recall rate achieved by the "Mulco+Biaffine" ensemble on ChiNesE could be a significant contribution to real-world applications, such as recommendation systems. In other words, the ability to retrieve a greater number of nested entities expands the upper limit of the downstream tasks.

5.5 Ablation Studies and Analysis

We conduct the ablation studies to verify the effectiveness of the components of Mulco and report the experimental results on both datasets in Table 6.

Scope v.s. traditional sequence labeling. In comparing the performance of each scope of Mulco, we found that B-min and B-max had higher recall rates but lower precision than the Innermost and Outermost methods. Although they both use sequence labeling and aim to find the longest and the shortest named entities, they use different methods to recognize named entities. The Innermost and Outermost methods use the "BIOES" scheme to label each token in a sentence. For an entity to be recognized, each token within it must be correctly classified. This results in a high precision for these

Table 5: The experimental results on ChiNesE and ACE 2005. Results with * are from ensemble models. We focus on the comparison between non-ensemble methods.

	Approach	ACE 2005			ChiNesE		
		P	R	F1	P	R	F1
Innermost	Sequence Labeling	89.27	65.74	75.72	93.79	54.42	68.88
Outermost	Sequence Labeling	**89.38**	61.54	72.83	**94.21**	43.94	59.93
Wang et al. [32]	Transition-based	77.62	81.60	79.56	86.56	76.34	81.13
Pyramid [33]	Layer-based	83.22	76.29	84.72	90.79	77.71	84.02
Biaffine [38]	Span-based	81.57	86.63	84.02	92.14	77.69	84.30
Shen et al. [28]	Span-based	84.43	85.29	84.86	90.03	78.32	83.77
W^{2}NER [13]	Ensemble	83.72*	88.30*	85.95*	86.08*	81.09*	86.11*
Mulco	Sequence Labeling	83.89	**88.17**	**85.98**	90.57	**80.47**	**85.23**
Mulco+Biaffine	Ensemble	80.26*	90.24*	84.96*	86.69*	86.32*	86.50*

Table 6: The results of the ablation studies.

	ACE 2005			ChiNesE		
	P	R	F1	P	R	F1
Mulco	83.89	**88.17**	85.98	90.57	**80.47**	**85.23**
B-min	**86.87**	69.41	77.16	92.11	57.44	70.75
B-max	84.66	67.60	75.17	91.12	56.85	70.02
E-min	86.19	87.73	**86.95**	93.14	72.13	81.30
E-max	86.20	87.59	86.89	**93.19**	72.05	81.27
-LSTM	82.20	87.15	84.60	89.71	80.71	84.98

Table 7: The agreements of predictions. The value represents the percentage of entities that are agreed by $k \in \{1, 2, 3, 4\}$ scopes.

		Number of Scopes			
	k	1	2	3	4
ACE 2005	Ground Truth	0.03	3.03	35.81	61.13
	Prediction	3.08	7.10	34.46	55.35
ChiNesE	Ground Truth	3.21	14.96	47.88	33.95
	Prediction	7.45	12.99	40.44	39.12

methods, as any wrong prediction will fail to retrieve the entity. On the other hand, B-min and B-max predict an entity by recognizing its first token and the length of the entity. This approach uses two classifiers to recognize entities, which lowers precision. However, it also allows for the recall of more candidates by retrieving entities that the traditional "BIOES" labeling method cannot recognize, resulting in a higher overall performance in terms of recall rate.

As an example, consider the performance of B-max in Figure 1, it can properly recognize both "Haidian District People's Government of Beijing Municipality" and "Haidian District People's Government" with a single scope. B-max recognizes an entity by considering only the anchor label and the length label of its first token. As a result, the two entities can be distinguished by B-max due to their different starting characters. On the other hand, the traditional sequence labeling method used in the Outermost method, considers all characters in an entity when recognizing, thus only allowing one character to have one label, even if it is an element of more than one entities, which results in the recognition of only "Haidian District People's Government of Beijing Municipality".

Diversity of nested named entities. We also found some interesting observations in the experiment results of E-min and E-max. As shown in Table 6, E-min and E-max have better performance than B-min and B-max on both datasets, and even achieve better performance than the default Mulco scope setting on ACE 2005 with only a single scope. This is because in Chinese, nested named entities are more likely to share the same starting characters with

other named entities. Out of the 1801 nested named entities in the test set of ACE 2005, 1479 of them are in such a situation, while there are only 4 named entities that have the same ending characters as other entities. We can see that the ratio (1479:4) is extremely unbalanced. This is why a single scope using the end character as the anchor can achieve even higher performance than the combination of four scopes. However, ChiNesE has a more balanced ratio, which is 7371:3218. Our dataset is collected from online passages of multiple domains, making it closer to the real-world CNNER application. Although E-min and E-max still outperform B-min and B-max on ChiNesE, the portion of named entities they can cover is much smaller than on ACE 2005. By combining the four scopes, Mulco reaches a much higher performance than the performance of any single scope on ChiNesE.

Necessity of contextual information. In Table 6, "-LSTM" reports the performance of Mulco without the Bi-LSTM encoder on ACE 2005 and ChiNesE. The F1 score drops by 1.38 on ACE 2005 and 0.25 on ChiNesE. The drop on ChiNesE is much lower than it is on ACE 2005. This is because BERT, a transformer-based pre-trained language model, can also grab contextual information for an input character. With a large enough dataset, BERT can better understand how to pay attention to contextual information for the CNNER task. Thus, Mulco can have enough contextual information for prediction even without LSTM on ChiNesE. However, on ACE 2005, the data size does not sufficiently support BERT to learn how to extract contextual information, and a Bi-LSTM model is required to enhance this ability. Additionally, the recognition of scope depends

Table 8: Two real case predictions made by Mulco on the ChiNesE dataset. The first sentence is "Lijin County Yanwo Town Middle School hosts the second Annual Table Tennis Competition for Faculty and Staff". The second sentence is "call for submissions for the 9th China Swan Poetry Award". Predictions marked with * indicate false predictions.

Sentence: 利津县盐窝镇中学举办第二届教职工乒乓球比赛		Ground Truth				Prediction			
Entity	English	B-min	B-max	E-min	E-max	B-min	B-max	E-min	E-max
利津县	Lijin County	Location	-	Location	Location	Location	-	Location	Location
利津县盐窝镇	Lijin County Yanwo Town	-	-	-	Location	-	-	-	Location
利津县盐窝镇中学	Lijin County Yanwo Town Middle School	-	Organization	-	Organization	-	Organization	-	Organization
盐窝镇	Yanwo Town	Location	-	Location	-	Location	-	Location	-
盐窝镇中学	Yanwo Town Middle School	-	Organization	Organization	-	-	Organization	Organization	-
第二届教职工乒乓球比赛	the second Annual Table Tennis Competition for Faculty and Staff	Event	Event	Event	Event	Event	Event	Event	Event
Sentence: 第九届中国白天鹅诗歌奖征稿启事		Ground Truth				Prediction			
Entity	English	B-min	B-max	E-min	E-max	B-min	B-max	E-min	E-max
第九届中国白天鹅诗歌奖	the 9th China Swan Poetry Award	Event	Event	-	Event	Event	Event	-	Event
中国	China	Location	-	Location	Location	Location	-	Location	Location
天鹅	Swan	Organisms	Organisms	Organisms	Organisms	-*	-*	-*	-*
中国白天鹅诗歌奖	China Swan Poetry Award	-	Event	-	-	-	Event	Event*	-
白天鹅诗歌奖	Swan Poetry Award	Event	Event	Event	-	Event	Event	-*	-

on the embedding of a single character, thus including contextual information into the embeddings is critical to the performance of our model. If we use a pre-trained word vector instead of BERT embedding as input, a contextualization model such as LSTM would then be necessary.

Agreement between scopes. We conducted further tests to analyze the agreement between the four scopes, and the results are presented in Table 7. Each value in the table represents the percentage of entities that are agreed upon by different numbers of scopes. For instance, using the ground truth label, we found that 47.88% of entities in the ChiNesE dataset can be detected by three scopes. However, in the actual predictions, only 40.44% of entities are recognized by three scopes. The discrepancy between the 'Ground Truth' and "Prediction" values is caused by the false predictions.

From the table, we observe that ChiNesE consists of more challenging entities compared to ACE 2005. Specifically, there are more entities in ChiNesE that can only be recognized by fewer scopes. Taking $k = 1$ as an example, in the ground truth, 3.21% of entities in ChiNesE can be recognized by a single scope, whereas in ACE 2005, only 0.03% of entities can be recognized by a single scope. This indicates that ACE 2005 is relatively easier than ChiNesE, with a majority of its entities being able to be recognized by at least two scopes.

5.6 Case Study

To further demonstrate the performance of Mulco, we conducted a case study on the ChiNesE dataset. Table 8 presents two real case predictions made by Mulco. In the table, each scope assigns a category to the entity if it recognizes it as a named entity; otherwise, a "-' is given.

For the first sentence, as indicated in the "Ground Truth" column, no single scope can recognize all entities by itself. However, by combining the predictions of the four scopes, Mulco successfully identifies all named entities from the complex nested structure, including the challenging entity "Lijin County Yanwo Town", which can only be located by the E-max scope.

In the case of the second sentence, Mulco misses the named entity "Swan". This may be attributed to the fact that "Swan" itself represents an animal, which significantly differs in semantic

context compared to the entities it is part of, such as "China Swan Poetry Award", which represents an event. As each scope in Mulco recognizes an entity based on the embedding of a single token, the semantics of "Swan" can be easily diluted by its contextual meaning as an event. To mitigate this issue, we can aggregate the pretrained word vectors of each token into the inputs of each scope to emphasize their original meaning.

Furthermore, there is also a false prediction made by E-min, which recognizes "China Swan Poetry Award" as the shortest named entity, with the end token identified as "Award" rather than "Swan Poetry Award". However, since Mulco incorporates four scopes, such an error made by a single scope may not result in an overall false prediction of the two named entities, as both of them can be recognized by multiple scopes.

6 CONCLUSION

This paper presents a new dataset, ChiNesE, to facilitate Chinese Nested Named Entity Recognition (CNNER) research. The dataset contains 20,000 high-quality samples, with 117,284 named entities falling into 10 categories, with 43.8% of them being nested. The samples are taken from online passages of multiple domains, making it more diverse and closer to real-world CNNER applications today. To address the limitations of sequence labeling and CNNER, we propose a novel scope-based labeling method Mulco that extracts named entities by four scopes, which predict the anchors and the lengths of the entities to recognize them and avoid the limitations of traditional sequence labeling schemes. Experimental results show that Mulco performs best on ChiNesE and ACE 2005, and is a good complementary method for real-world CNNER applications when employed in ensemble techniques.

7 ETHICS STATEMENT

The creation of ChiNesE dataset incurred a cost of over 19,000 CNY. We hired three experts in NLP data annotation who each had at least one year of working experience. The annotation process took 504 work hours to complete, with an hourly rate of about 37.69 CNY, which is higher than the standard local rate of 25.3 CNY per hour. The online passages used in the dataset were sourced from QQ Browser, which has already been censored by the platform.

REFERENCES

[1] Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*. 65–72.

[2] Yanping Chen, Guorong Wang, Qinghua Zheng, Yongbin Qin, Ruizhang Huang, and Ping Chen. 2019. A set space model to capture structural information of a sentence. *IEEE Access* 7 (2019), 142515–142530.

[3] Yanping Chen, Yuefei Wu, Yongbin Qin, Ying Hu, Zeyu Wang, Ruizhang Huang, Xinyu Cheng, and Ping Chen. 2019. Recognizing nested named entity based on the neural network boundary assembling model. *IEEE Intelligent Systems* 35, 1 (2019), 74–81.

[4] 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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

[6] Chunyuan Fu and Guohong Fu. 2012. A dual-layer CRFs based method for Chinese nested named entity recognition. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. 2546–2550. https://doi.org/10.1109/FSKD.2012.6234172

[7] Zhifeng Hao, Hongfei Wang, Ruichu Cai, and Wen Wen. 2013. Product named entity recognition for Chinese query questions based on a skip-chain CRF model. *Neural Computing and Applications* 23, 2 (2013), 371–379.

[8] Wang Houfeng and Shi Wuguang. 2005. A simple rule-based approach to organization name recognition in chinese text. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 769–772.

[9] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. 2022. A Review of Yolo algorithm developments. *Procedia Computer Science* 199 (2022), 1066–1073.

[10] Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1446–1459.

[11] Arzoo Katiyar and Claire Cardie. 2018. Nested Named Entity Recognition Revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 861–871. https://doi.org/10.18653/v1/N18-1079

[12] Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation* 24, 2 (2014), 473–489. https://doi.org/10.1093/logcom/exs079

[13] Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10965–10973.

[14] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. on Knowl. and Data Eng.* 34, 1 (jan 2022), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

[15] Ren Li, Tianjin Mo, Jianxi Yang, Dong Li, Shixin Jiang, and Di Wang. 2021. Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. *Advanced Engineering Informatics* 50 (2021), 101416.

[16] 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*. Association for Computational Linguistics, Online, 5849–5859. https://doi.org/10.18653/v1/2020.acl-main.519

[17] Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5182–5192. https://doi.org/10.18653/v1/P19-1511

[18] Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing* 473 (2022), 37–53.

[19] Xinwei Long, Shuzi Niu, and Yucheng Li. 2020. Hierarchical Region Learning for Nested Named Entity Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4788–4793. https://doi.org/10.18653/v1/2020.findings-emnlp.430

[20] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[21] Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 857–867.

[22] Zita Marinho, Alfonso Mendes, Sebastiao Miranda, and David Nogueira. 2019. Hierarchical nested named entity recognition. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 28–34.

[23] Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. 1–8.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[25] Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D Spyropoulos. 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 426–433.

[26] Lisa F Rau. 1991. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*. IEEE Computer Society, 29–30.

[27] Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A Dataset for Nested Named Entity Recognition in English Newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5176–5181.

[28] Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2782–2794.

[29] Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics* 8 (2020), 605–620.

[30] Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. Boundary enhanced neural span classification for nested named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9016–9023.

[31] Bailin Wang and Wei Lu. 2018. Neural Segmental Hypergraphs for Overlapping Mention Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 204–214.

[32] Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018. A Neural Transition-based Model for Nested Mention Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1011–1017. https://doi.org/10.18653/v1/D18-1124

[33] Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A Layered Model for Nested Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5918–5928. https://doi.org/10.18653/v1/2020.acl-main.525

[34] Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested Named Entity Recognition: A Survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2022).

[35] Casey Whitelaw and Jon Patrick. 2003. Evaluating corpora for named entity recognition using character-level features. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 910–921.

[36] Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained Named Entity Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1430–1440. https://doi.org/10.18653/v1/P19-1138

[37] Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

[38] Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6470–6476. https://doi.org/10.18653/v1/2020.acl-main.577

[39] S Yu, H Duan, and Y Wu. 2018. Corpus of multi-level processing for modern Chinese. *Available at: opendata. pku. edu. cn/dataset. xhtml* (2018).

[40] Xiantao Zhang, Dongchen Li, and Xihong Wu. 2014. Parsing named entity as syntactic structure. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[41] Yuejie Zhang, Zhiting Xu, and Tao Zhang. 2008. Fusion of multiple features for Chinese named entity recognition based on CRF model. In *Asia Information Retrieval Symposium*. Springer, 95–106.