

A Boundary Regression Model for Nested Named Entity Recognition

Yanping Chen*, Lefei Wu^{†*}, Liyuan Deng*, Yongbin Qing*,
Ruizhang Huang*, Qinghua Zheng[†], Ping Chen[‡]

* Guizhou University, Guiyan, China

[†] Xian'an Jiaotong University, Xi'an, China

[‡] University of Massachusetts Boston, Boston, USA

Abstract—Recognizing named entities (NEs) is commonly conducted as a classification problem that predicts a class tag for an NE candidate in a sentence. In shallow structures, categorized features are weighted to support the prediction. Recent developments in neural networks have adopted deep structures that map categorized features into continuous representations. This approach unfolds a dense space saturated with high-order abstract semantic information, where the prediction is based on distributed feature representations. In this paper, the regression operation is introduced to locate NEs in a sentence. In this approach, a deep network is first designed to transform an input sentence into recurrent feature maps. Bounding boxes are generated from the feature maps, where a box is an abstract representation of an NE candidate. In addition to the class tag, each bounding box has two parameters denoting the start position and the length of an NE candidate. In the training process, the location offset between a bounding box and a true NE are learned to minimize the location loss. Based on this motivation, a multiobjective learning framework is designed to simultaneously locate entities and predict the class probability. By sharing parameters for locating and predicting, the framework can take full advantage of annotated data and enable more potent non-linear function approximators to enhance model discriminability. Experiments demonstrate state-of-the-art performance for nested named entities¹.

I. INTRODUCTION

Named entity (NE) recognition is often modeled as a sequence labeling task, where a sequence model (e.g., conditional random fields (CRF) [21] or long short-term memory (LSTM) [11]) is adopted to output a maximized label sequence. However, NEs in a sentence have more complex structures, and nested NEs are widely used to represent semantic relationships between entities (e.g., affiliation, ownership, hyponymy). The nesting ratio in the GENIA corpus and ACE corpus is 35.27% and 33.90%, respectively [22], [5]. Because sequence models usually assume a flattened structure for an input sentence, it is not effective to find nested NEs. For example, “University of Washington” is an organization NE, where “Washington” is a nested NE indicating the location of the university. A label sequence (e.g., “S-I-S”, or “S-I-I”) fails to decode this phenomenon.

Sequence models can be revised to support nested NE recognition, e.g., the cascading strategy, layering strategy

or joint strategy [1]. In the cascading or layering strategy, each type or each layer of nested NEs is processed by an independent sequence model. In the joint strategy, nested labels are transformed into structured labels. For example, “S_GPE+S_ORG” represents that a unit is the start of a geographical name and an organization name at the same time. These strategies have three shortcomings: 1) Nested NEs in the same type or layer still cannot be recognized. 2) They cannot make full use of annotated data. 3) Some units are forced to change their labels in falseness to obtain flattened label sequences. 4) The joint strategy substantially increases the number of entity types, which worsens the performance of NE recognition.

In recent years, many nesting-oriented models have been proposed to handle the nesting problem. They usually transform the nesting structure into a nonsequential structure, e.g., a hypergraph [20] or a tree representation [31]. There are models dividing the recognition process in multistage pipelines, e.g., verifying every fragment [28] or assembling NE boundaries [2]. Others adopt a sequence-to-sequence [29] or an end-to-end framework [36]. Regardless of the approach, they may suffer from several problems, such as a greater dependence on external toolkits (e.g., parsing), an inability to guarantee global optimization (in multistage pipelines) or the lack of fully using annotated information.

This paper proposes an end-to-end architecture to recognize nested NEs, named the boundary regression (BR) model. In this model, each input sentence is first mapped into *recu* (recurrent) feature maps by a basic network, which can be truncated from a standard neural network for NE recognition. Then, each feature map location is combined with others to generate default bounding boxes at each spatial location and scale. A bounding box is a high-order abstract representation of an NE candidate that mixes information about its semantic dependency and context. In addition to the class category, each bounding box is also marked with its location parameters, which denotes the start position and the length of an NE candidate in a sentence. In the training process, in addition to maximizing bounding box confidence scores to be an entity, a regression layer is added to minimize its location offset relative to a true NE, which enables every bounding box to approach a neighboring true NE. This model is an end-to-end multiobjective learning framework that simultaneously

¹Our codes will be available at: <https://github.com/wuyuefei3/BR>

²In the “S-I-O” encoding, tags “S”, “I” and “O” indicate that a unit is a *Start*, *Inside* or *Outside* of an NE.

locates entities and predicts class probabilities. By shared parameters of multitasks, the BR model can benefit from both boundary regression and global optimization. It enables more potent nonlinear function approximators to enhance model discriminability.

The contributions of this paper include the following.

- 1) Bounding boxes are proposed to represent abstract NE representations. Nested NEs are distinguished by overlapped bounding boxes marked with location parameters. In addition to predicting classification confidence scores, nested NEs can be located from a sentence by the regression operation.
- 2) An end-to-end multiobjective learning framework is designed. Under this framework, entity locations and class probabilities are simultaneously predicted. By sharing the same network parameters for locating and predicting, it enables more potent nonlinear function approximators for the NE recognition.

The structure of this paper is organized as follows. Before discussing the details of this model, our motivation is first discussed in Section II. Section III presents the definition about boundary regression and the architecture of the BR model. Experiments are conducted in Section IV. In this section, several issues about the BR model are discussed. Finally, the BR model is compared with several state-of-the-art systems. Section V and Section VI introduce related works and the conclusion.

II. MOTIVATION

The BR model is motivated by techniques developed in object detection in computer vision. At first sight, a sentence is a one-dimensional linear textual stream, and an image is a two-dimensional pixel patch. They are totally different in external representation. However, as shown in Figure 1, spatial distribution patterns of entities and objects have similar structures.

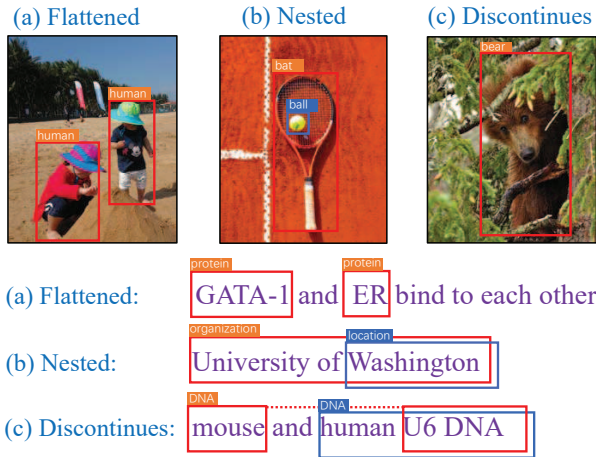


Fig. 1: Similarity between Entities and Objects

The structures can be roughly divided into three categories: *flattened*, *nested* and *discontinues*. Flattened entities

(or objects) are spatially separated from each other. In nested structure, two or more entities or objects can be overlapped with each other. The discontinuous structure can be transformed into the nested structure, where discontinuous entities or objects are handled as a whole element. In addition to the structure similarity, there are two differences between them. First, the detection of objects is mainly based on internal features of objects. Therefore, an object can move in an image. It has less influence on the object detection. However, in a sentence, entities have a strong semantic dependency in a sentence. Second, features of the nested object are covered by the upper object. It increases difficulties in detecting the blocked object. On the other hand, in NE recognition, overlapped features are shared by nested entities.

In object detection, a deep neural network is usually adopted to map an input image into abstract representations known as *conv* feature maps. Then, bounding boxes are generated from feature maps. Bounding boxes are abstract representations of objects. Each box has four parameters denoting its location (x, y) and shape ($\Delta x, \Delta y$) in an image. These parameters are the key to distinguish nested objects. Finally, in an end-to-end object detection model, a multiobjective learning architecture is designed to simultaneously locate objects and predict class probability.

In recent years, based on deep neural networks, language and vision can be embedded into a distributed representation and mapped into an abstract semantic space. Therefore, the combination of natural language processing and computer vision has become popular in research communities, e.g., the text retrieval approach in videos [27] and multimodal deep learning [34]. Motivated by the success achieved in object detection of computer vision, we present an end-to-end multiobjective learning framework to support NE recognition. The core innovation of the BR model is the introduction of the “bounding box” that is an abstract representation of NEs. Every box has two parameters denoting its location in a sentence, which can distinguish nested NEs. The location parameters can also support the regression operation to locate NEs in a sentence. This concept is visualized in Figure 2.

As shown in Figure 2, an input sentence is first mapped into *recu* feature maps by a deep neural network. The feature maps can be seen as an abstract representation of an input sentence. Every feature map denotes a representation of an NE boundary, which can be bounded with others to generate bounding boxes (an example d_i is shown in Figure 2). A bounding box that correctly matches a true NE is referred to as a “true bounding box” (or true box). Every box has two parameters to indicate its position (s_i) and shape (or length) (l_i)³. The regression operation predicts the position offset and shape offset (Δs_i and Δl_i) respectively relative to a true box, (e.g., g_j in Figure 2). Finally, in the output, locations of the recognized NEs are updated as $\tilde{d}_i = \{s_i + \Delta s_i, l_i + \Delta l_i\}$. Because the outputs of a regression operation are continuous values, they are rounded

³In this paper, the position parameter (s_i) and shape parameter (l_i) of an NE are also referred to as location parameters.

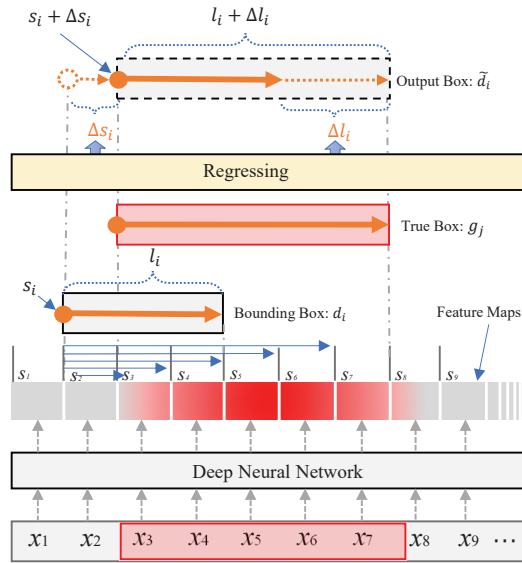


Fig. 2: Regression Operation

to the nearest word boundary locations.

To design an end-to-end multiobjective learning architecture for NE recognition, we should carefully take the following four issues under consideration:

1) **Representation:** Object detection usually uses stacked convolutional layers to map an image into *conv* feature maps. In language processing, the recurrent (or attention) neural network is more effective in capturing the semantic dependency in a sentence. In this paper, the abstract representations of input sentences are referred to as *recu* feature maps.

2) **Region proposal:** In the *recu* feature map, a feature map location can be seen as an abstract representation of a possible entity boundary. It can be bounded with other feature maps to generate default bounding boxes. Every bounding box is an abstract representation of an NE candidate labeled with its location information and class category.

3) **Multiobjective learning:** Because the recurrent neural network can learn the semantic dependency, a bounding box contains semantic information about the whole sentence. In addition to predicting conditional class probabilities on a bounding box, a regression layer can be stacked to predict its location in a sentence.

4) **Maximum of overlapping neighborhoods:** In the prediction process, every bounding box will approach a true bounding box. They are overlapped in the neighborhood of a true bounding box. It is necessary to collect the most likely matched bounding boxes from overlapped bounding boxes.

According to the above discussion, the architecture of the BR model is given in the following section.

III. MODEL

In this paper, instead of modeling the NE recognition task as a classification problem, we frame the task as a multiobjective optimization process. In addition to outputting discretized entity categories, the regression operation is integrated to

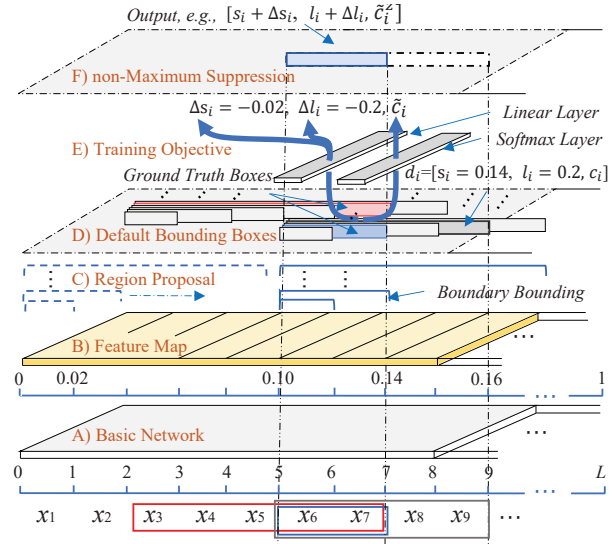


Fig. 3: An input sentence is input into a *basic network*, which maps the sentence into a *recu* feature map. *Region proposal* generates *default bounding boxes* from the *recu* feature map. Then, the model is trained to satisfy the *training objective*. *Non-maximum suppression* is adopted to produce a final decision.

compute the location offset of an NE candidate relative to a true NE in a sentence. The structure of the BR model is shown in Figure 3.

As Figure 3 shows, six specific issues (referred to as A to F) are highlighted in the BR model. They are discussed as follows.

A. Basic Network

The basic network is adopted to map a sentence into a distributed representation, where words (or characters) are embedded into vectors by a lookup table initialized randomly or pretrained with external resources. In the task of NE recognition, the semantic meaning of a word is heavily influenced by its context. Therefore, a recurrent neural network or the attention mechanism can be used to capture the semantic dependency between words.

The basic network can be truncated from a standard architecture of a high-quality sentence labeling task (e.g., an NE recognition or a POS tagging task). In the current BR model, we adopted a basic network consisted of an embedding layer and a Bi-LSTM layer. To simplify the region proposal step, we set the length of the input sentence as a fixed number, denoted as L . Longer or shorter sentences are trimmed or padded, respectively.

B. Feature Map

The output of the basic layer is denoted as *recu* feature maps, which represents high-order abstract features of a sentence integrated with semantic dependency information between words. Because images have an invariance property for a zooming operation, object detection can benefit from

multigranularity representations, where the region proposal can be implemented on multiscale feature maps to generate bounding boxes with different scales.

In natural language processing, it is difficult to condense a textual sequence into multigranularity representations. At present, for an input sentence, we only generate a single *recu* feature map layer. It is a high-order abstract representation of an input sentence. Each feature map location corresponds to a possible entity boundary. To support the regression operation, the lengths of feature map locations are normalized into interval $[0, 1]$ to smooth the learning process.

C. Region Proposal

Each feature map location corresponds to an abstract representation of a possible NE boundary in a sentence. It can be set as a start boundary of an NE and then assembled with the other feature maps to generate default bounding boxes with different lengths. In this paper, a feature map is bounded with K left feature maps, where the value K is a predefined parameter corresponding to the longest NE candidate. It outputs K bounding boxes per feature map location.

The current region proposal is similar to an exhaustive enumeration method, which verifies every possible NE candidate up to a certain length (e.g., Sohrab et al. [28]). However, they are different. Bounding boxes are abstract representations of a possible NE. It has a receptive field⁴ across the whole sentence. It is labeled with location information. Therefore, bounding boxes can be used directly to support the classification and boundary regression. Furthermore, in the training process, location parameters of bounding boxes can be used to control the data imbalance problem. For example, the overlapping rate between a bounding box and a true bounding box can be adopted to filter insignificant bounding boxes (discussed in Section III-D). It can be used to reduce computational complexity and decrease the influence caused by negative examples.

D. Default Bounding Boxes

Let \mathbf{D} denote a bounding box set generated from an input sentence S . Each bounding box $d_i \in \mathbf{D}$ has 3 parameters d_i^s , d_i^l and \mathbf{c}_i . Parameters d_i^s and d_i^l are two real numbers denoting the start position and length of d_i in a sentence, respectively⁵. Parameter $\mathbf{c}_i = (c_i^1, c_i^2, \dots, c_i^Z)$ is a one-hot vector representing the entity type of d_i , where Z is the number of entity types. A bounding box d_i can be referred to as a three-tuple $d_i = \langle d_i^s, d_i^l, \mathbf{c}_i \rangle$. If a bounding box corresponds to a true NE, it is referred to as a ground truth box and represented as $g_j = \langle g_j^s, g_j^l, \mathbf{c}_j \rangle$.

Bounding boxes are marked with the position and shape parameters. This information is useful for selecting specific bounding boxes. Given two bounding boxes (e.g., A and B),

the Intersection over Union (IoU) between them is formalized as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \quad (1)$$

Let \mathbf{D}_G represent the set of all ground truth boxes in \mathbf{D} . We define two sets as follows:

$$\begin{aligned} \mathbf{D}_p &= \{d_i | d_i \in \mathbf{D}, \exists g_j \in \mathbf{D}_G (IoU(d_i, g_j) \geq \gamma)\} \\ \mathbf{D}_n &= \{d_i | d_i \in \mathbf{D}, \forall g_j \in \mathbf{D}_G (IoU(d_i, g_j) < \gamma)\} \end{aligned} \quad (2)$$

In this paper, \mathbf{D}_p is referred as the positive bounding box set, and \mathbf{D}_n is referred to as the negative bounding box set. In the region proposal process, a large number of negative bounding boxes will be generated, which lead to a significant data imbalance problem. It is also expensive in computing. In the training process, we collect \mathbf{D}_p and \mathbf{D}_n with a ratio of 1:3 for balancing positive and negative samples. It guarantees a faster optimization and a stable training process.

Given a bounding box d_i , the relative ground truth box is identified as:

$$g_j = \arg \max_{g_i \in \mathbf{D}_G} IoU(d_i, g_j) \quad (3)$$

Given a ground truth box g_j , all bounding boxes of \mathbf{D}_p satisfying Equation 3 are referred to as \mathbf{D}_{g_j} . They are the neighborhoods of g_j .

For convenience, Table I lists the definitions about bounding box sets.

TABLE I: Symbol Table

Symbol	Meaning
\mathbf{D}	A bounding box set generated from a sentence S ;
\mathbf{D}_G	The set of all ground truth boxes in \mathbf{D} ;
\mathbf{D}_p	The positive bounding box set;
\mathbf{D}_n	The negative bounding box set;
\mathbf{D}_{g_j}	The neighborhoods of g_j ;

E. Training Objective

For every $d_i \in \mathbf{D}_p$, the location offset of d_i relates to a ground truth box that will be predicted by the linear layer.

Let $\hat{d}_{ij}^s = (g_j^s - d_i^s)/g_j^l$ and $\hat{d}_{ij}^l = \log(g_j^l/d_i^l)$ be the normalized position offset and shape offset between d_i and g_j . Given a bounding box d_i , the BR model outputs 3 parameters: Δd_i^s , Δd_i^l and $\tilde{\mathbf{c}}_i$. Δd_i^s and Δd_i^l denote the predicted position offset and shape offset of d_i relative to g_j . $\tilde{\mathbf{c}}_i$ is a confidence score that reflects the confidence that a box contains a true NE. As Figure 3 shows, Δd_i^s and Δd_i^l are regressed by a linear layer, while the classification confidence score $\tilde{\mathbf{c}}_i = (\tilde{c}_i^0, \tilde{c}_i^1, \dots, \tilde{c}_i^Z)$ is predicted by a softmax layer.

We define a characteristic function $E_{ij}^z = \{0, 1\}$ indicating that a default box d_i is matched to a relative ground truth box g_j selected by Equation 3. In the training process, the regression operation updates Δd_i^s and Δd_i^l for the purpose of approaching \hat{d}_{ij}^s and \hat{d}_{ij}^l , respectively. The location loss can be computed as:

$$L_{loc}(x, s, l) = \sum_{g_j \in \mathbf{D}_G} \frac{1}{N} \left(\sum_{d_i \in \mathbf{D}_{g_j}} \sum_{h \in \{s, l\}} E_{ij}^h \text{Smooth}_{L_1}(\Delta d_i^h - \hat{d}_{ij}^h) \right) \quad (4)$$

⁴The receptive field is the range in an input sentence, and the bounding box contains its information.

⁵The end position of d_i can be computed as $d_i^s + d_i^l$

IV. EXPERIMENTS

In the above equation, $N = |\mathbf{D}_{g_j}|$ is the cardinality of \mathbf{D}_{g_j} . It is used to normalize the weight between $g_j \in \mathbf{D}_G$. Smooth_{L_1} is a robust L_1 loss that quantifies the dissimilarity between d_i and g_j . It is less sensitive to outliers [10].

Confidence loss is a softmax loss over multiple class confidences. It is given as follows:

$$L_{con}(x, c) = - \sum_{d_i \in \mathbf{D}_p} E_{ij}^z \log(\tilde{c}_i^z) - \sum_{d_i \in \mathbf{D}_n} \log(\tilde{c}_i^0) \quad (5)$$

where $\tilde{c}_i^z = \exp(\tilde{c}_i^z) / \sum_{z=0}^Z \exp(\tilde{c}_i^z)$, \tilde{c}_i^0 is the confidence score indicating that an example is negative. The total loss function combines the location loss and confidence loss.

$$L(x, s, l, c) = L_{loc}(x, s, l) + \alpha L_{con}(x, c) \quad (6)$$

where α is a predefined parameter balancing the weight between the location loss and confidence loss. The *training objective* is to reduce the total loss of the location offset and class prediction. In the training process, we optimize their locations to improve their matching degree and maximize their confidences.

F. Non-Maximum Suppression

In the prediction process, the BR model outputs a set of bounding boxes for each input sentence, referred to as $\mathbf{D} = \{d_1, \dots, d_{|\mathbf{D}|}\}$. Every box $d_i \in \mathbf{D}$ has 3 outputs: Δd_i^s , Δd_i^l and \tilde{c}_i indicating the position offset, shape offset and class probability of d_i relative to a truth bounding box, respectively. After Δd_i^s and Δd_i^l are resized as Δs_i and Δl_i , an NE can be identified from d_i as follows: $[s_i + \Delta s_i, l_i + \Delta l_i, \tilde{c}_i]$. An example of the output is shown in Figure 3.

The output \mathbf{D} contains a large number of boxes, but many of them are overlapped. Non-maximum suppression (NMS) is implemented in the prediction process to produce the final decision, which selects true boxes from overlapped neighborhoods. The NMS algorithm is shown in Table II.

TABLE II: The NMS algorithm of the BR Model

Input: $\mathbf{D} = \{d_1, \dots, d_{ \mathbf{D} }\}$, a threshold λ .
Output: \mathbf{D}_e recognized NEs.
1: While(if \mathbf{D} is not empty){
2: Select d_i from \mathbf{D} : for $\forall d_j \in \mathbf{D}$, if $d_j \neq d_i$, $\tilde{c}_j < \tilde{c}_i$.
3: Delete d_i from \mathbf{D} .
4: For $\forall d_j \in \mathbf{D}$, if $IoU(d_j, d_i) > \lambda$, then delete d_j from \mathbf{D} .
5: Add d_i to \mathbf{D}_e . }

It is a one-dimensional NMS algorithm that selects multiscale nested NEs from overlapped positive samples. The NMS algorithm searches local maximized elements from overlapping neighborhoods in which a smaller number of high-confidence boxes are collected. The threshold is adopted to control the overlapping ratio between neighborhoods. In our experiments, the value of λ is set as 0.65. The algorithm improvement of NMS is left as our future work.

In our experiments, the ACE 2005 corpus⁶ is adopted to evaluate the BR model. This corpus is collected from broadcasts, newswires and weblogs. It is the most popular source of evaluation data for NE recognition and places more emphasis on language understanding. Compared to other corpora, the task to recognize NEs under the ACE annotation is more challenging. For example, “the six” is labeled as an NE in “President Megawati Sukarnoputri has refused clemency for the six and asked for a speedy execution”. As another example, the sentence “the 31-year-old mother of three was convicted” is annotated with a person name of “three”.

The ACE corpus contains three datasets: Chinese, English and Arabic. In this paper, the BR model is mainly evaluated on the ACE Chinese corpus. To show the extensibility of the BR model regarding other languages, it is also evaluated on the ACE English corpus. The result is presented in Section IV-D in detail.

The Chinese ACE corpus contains 633 documents with 7 entity types. It contains 499 *Vehicle* (VEH), 1,277 *Location* (LOC), 324 *Weapon* (WEA), 8,071 *Geo-Political* (GPE), 11,351 *Person* (PER), 4,837 *Organization* (ORG) and 1,194 *Facility* (FAC) instances, in which 27,553 entity mentions are collected. Approximately 30% of entity mentions⁷ are nested with each other.

In the BR model, we use the same settings as Chen et al. [2] to configure the basic neural network. The length of sentence L is set to 50. The threshold $\gamma = 0.5$ is set to collect positive samples. In the total loss function, $\alpha = 1$ is used. The parameter for boundary bounding is set to $K = 6$. In all experiments, the performance is reported on 7 positive entity types (the micro-average or “Total”) unless explicitly mentioned.

In our setting, recognizing an NE requires that start and end boundaries of an NE are precisely identified. Because the BR model uses a regression operation to locate an NE of a sentence, all entity locations are mapped into interval $[0, 1]$ for a smooth learning gradient. Therefore, the output of BR is rounded to the nearest character location.

A. Influence of Word Representation

For a better understanding of the BR performance, we first make a “close” evaluation, where BR adopts a randomly initialized lookup table to implement word embedding. This model is referred to as BR-Random. It is compared with another BR model (referred to as BR-BERT), which adopts a lookup table pretrained from BERT [4]. In the basic network, the dimension of word embedding is 300 dimensions in BR-Random and 768 dimensions in BR-BERT. Word embedding is fed into a Bi-LSTM layer, which outputs a 128×2 dimensional *recu* feature map. The result is shown in Table III.

The NE recognition task is implemented at sentence level, where a sentence is usually composed with a limited number

⁶The data are available at: <https://catalog.ldc.upenn.edu/LDC2006T06>.

⁷An entity mention is a phrase referred to as an NE in a sentence.

TABLE III: Influence of Word Embedding

TYPE	BR-Random			BR-BERT		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
VEH	56.92	25.52	35.24	84.71	49.66	62.61
LOC	51.69	21.71	30.58	79.87	43.77	56.55
WEA	64.00	22.86	33.68	71.43	42.86	53.57
GPE	65.98	62.24	64.05	86.59	82.72	84.61
PER	64.43	57.34	60.68	88.57	83.34	85.88
ORG	54.44	39.29	45.64	82.01	70.17	75.63
FAC	55.56	17.86	27.03	70.39	50.00	58.47
Total	63.59	51.60	56.97	85.79	76.24	80.73

of words. The task suffers from a serious feature sparsity problem, which worsens the performance of BR-Random. Because neural network is powerful to capture semantic information by using a pre-trained lookup table initialized from external resources, performance of BR-BERT shows a significant improvement. The result indicates a basic network, truncated from a high-quality sentence labeling task, is helpful to support the BR model.

B. Influence of the Boundary Regression

As discussed in Section III-D, \mathbf{D} is all the bounding boxes generated from an input sentence S . As shown in Table I, several bounding box sets with specific properties are generated from \mathbf{D} , e.g., \mathbf{D}_p , \mathbf{D}_n and \mathbf{D}_{g_j} . They play different roles in the BR model.

Equation 4 showed that only \mathbf{D}_p is used to compute the location loss. \mathbf{D}_n is mainly used to train the classifier for recognizing negative NEs. Using \mathbf{D}_G and Equation 3, \mathbf{D}_p can be partitioned into a set $\mathcal{D}_p = \{\mathbf{D}_{g_1}, \mathbf{D}_{g_2}, \dots\}$. \mathcal{D}_p is a partition of \mathbf{D}_p , where $\forall d_i \in \mathbf{D}_p (\exists g_j \in \mathbf{D}_G \wedge d_i \in \mathbf{D}_{g_j})$. If $i \neq j$, $\mathbf{D}_{g_i} \cap \mathbf{D}_{g_j} = \emptyset$. Therefore, the location offset of $d_i \in \mathbf{D}_p$ is only related to a ground truth box. For every ground truth box g_j , its neighborhoods \mathbf{D}_{g_j} are used to generate the location loss. This setting is natural because neighborhoods contain sufficient semantic features about an NE for boundary regression.

In the task of NE recognition, a correct identification requires that both the start and end boundaries of an output and a true NE are precisely matched. In the region proposal process, we set all bounding boxes in \mathbf{D}_{g_j} with the same class tag as the relative ground truth box g_j . This setting means that a bounding box that does not precisely match a true NE was given a positive class tag. It will lead to chaos in a traditional model. However, for the BR model, during the training process, the offset is valuable to train the linear layer. Then, every bounding box can approach the true NE by the regression operation. To show the influence of boundary regression, Table IV presents two experiments.

With the exception of the linear layer that is “closed”, “MODEL A” and “MODEL B” have the same architecture and settings as the BR model. They only output class confidence scores for all input bounding boxes. The input for “MODEL A” is the same as the BR model, where all bounding boxes in \mathbf{D}_{g_j} are labeled with the same class tag as the relative ground truth box g_j . The ratio between \mathbf{D}_p and \mathbf{D}_n is also 1:3. The

TABLE IV: Performance with Boundary Regression

TYPE	MODEL A			MODEL B		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
VEH	24.29	59.31	34.47	84.93	42.76	56.88
LOC	18.94	53.38	27.96	84.67	45.20	58.93
WEA	18.45	54.29	27.54	79.41	38.57	51.92
GPE	21.31	85.80	34.13	87.98	77.40	82.35
PER	23.89	87.63	37.55	88.56	80.15	84.14
ORG	18.23	76.86	29.47	80.49	66.25	72.68
FAC	17.33	50.40	25.79	72.06	38.89	50.52
Total	21.56	80.92	34.04	86.30	72.10	78.56

result shows that poor performance is obtained (34.04% in the F1 score). As discussed above, the reason is that many bounding boxes in \mathbf{D}_p were given false class tags. They disturb the learning process and cannot yield a coherent predication.

In “MODEL B”, only the ground truth boxes are collected as positive instances. If the start and end boundaries of a bounding box are not precisely matched to a true NE, it is labeled as a negative instance. This setting is the same as an exhaustive enumeration method, which verifies every possible NE candidate up to a certain length (e.g., Sohrab et al. [28]). Compared with the BR model, it has lower performance. There may be two reasons that lead to this problem. First, it generate a large number of negative instances, which leads to the data imbalance problem. Second, in the exhaustive enumeration method, many negative bounding boxes are highly overlapped with a ground truth box. These boxes are very ambiguous and may influence the final performance.

C. Comparison with Chinese Models

To compare the performance of BR with other models, we adopt a popular sequence model (Bi-LSTM-CRF) [12]. It consists of an embedding layer, a Bi-LSTM layer, an MLP layer and a CRF layer. The embedding layer and Bi-LSTM layer have the same settings as the basic network of BR. The MLP layer generates a distribution among NE types. The CRF layer outputs a maximized label sequence. If nested NEs occur, only innermost (or outermost) NEs are selected. Because sequence models are not effective in recognizing nested NEs, as discussed in Section I, the cascading strategy and layering strategy are adopted to solve the nesting problem [1].

In the Chinese ACE corpus, BA is a pipeline framework for nested NE recognition that has achieved state-of-the-art performance [2]. The original BA is a “Shallow” model, which uses a CRF model to identify NE boundaries and a maximum entropy model to classify NE candidates. NNBA is a neural network version, where the LSTM-CRF model is adopted to identify NE boundaries, and a multi-LSTM model is adopted to filter NE candidates.

In this experiment, the “Adam” optimizer is adopted. The learning rate, weight decay rate and batch size are set as 0.00005, 0.01 and 30, respectively. Shallow models refer to CRF-based models. All deep models use the BERT word embedding. We conduct these models with the same settings

and the same data. All entities with lengths larger than 6 are ignored. The results are shown in Table V.

TABLE V: Comparing with Other Methods

Model		P(%)	R(%)	F(%)
Shallow Models	Innermost	73.60	45.50	56.32
	Outermost	72.60	45.54	55.97
	Cascading	76.52	51.80	61.80
	Layering	71.93	56.57	63.33
	BA [3]	73.98	62.16	67.56
Deep Models	Innermost	82.00	70.70	75.93
	Outermost	80.45	69.08	74.33
	Cascading	76.96	71.39	74.07
	Layering	78.85	81.34	80.07
	NNBA [2]	80.49	79.46	79.97
BR		85.79	76.24	80.73

In Table V, all deep models outperform shallow models because neural networks can effectively use external resources by using a pretrained lookup and have the advantage of learning abstract features from raw input. In deep models, the performance of innermost and outermost models is heavily influenced by a lower recall rate, which is caused by ignoring nested NEs. The deep cascading model also suffers from poor performance because predicting every entity type by an independent classifier is not an effective approach when using annotated data. The deep layering model is impressive. This model is conducted by implementing two independent classifiers that separately recognize the innermost and outermost NEs. It has higher performance, even outperforming the NNBA model. The reason for the improvement is that, in our experiments, entities with a length larger than 6 are ignored, which decreases the nesting ratio. Most of the nested NEs have two layers, which can be handled appropriately by the layering model.

In Table V, the BR model has the best performance. Compared with traditional models, the BR model has three advantages. First, bounding boxes of the BR model are labeled with position parameters. In the training process, they can be used to detect negative examples. This technique is helpful in reducing the influence of data imbalance and in decreasing the computational complexity. Second, learning the training objective for the BR model is a multiobjective optimization problem that simultaneously locates NEs and predicts the class probability. Multiobjective learning can share parameters between tasks. The regression operation to locate NEs has the advantage of using supervised information (NE locations) in an annotation corpus. Third, The BR model is an end-to-end model, which has the ability to learn a global optimized solution and avoid the cascading failure.

The Chinese language is hieroglyphic. It has very little morphological information (e.g., the capitalization) to indicate the word usage. Because there is a lack of delimitation between words, it is difficult to distinguish monosyllabic words from monosyllabic morphemes. However, Chinese has two distinctive characteristics. First, Chinese characters have the shape of a square⁸. Their locations are symmetrical. Second,

because the meaning of a Chinese word is usually derived from the characters it contains, every character is informative. Therefore, character representations can well capture syntactic and semantic information of a sentence. The BR model works well on the Chinese corpus. To show the extensibility of the BR model, in the following experiment, the BR model is also transformed to the English corpus for further assessment.

D. Comparison with English Models

The English ACE corpus contains 506 documents, which also are annotated with 7 types. It contains 810 *Vehicle* (VEH), 1,067 *Location* (LOC), 866 *Weapon* (WEA), 7,183 *Geo-Political* (GPE), 24,940 *Person* (PER), 5,083 *Organization* (ORG) and 1,357 *Facility* (FAC) instances. As used in most previous studies, we adopt the same settings as Lu et al. [20] to evaluate the BR model. The performance is listed in Table VI.

In Table VI, Lu et al. [20] and Katiyar et al.[16] represent nested NEs as mention hypergraphs. Ju et al. [14] feed the output of a BiLSTM-CRF model to another BiLSTM-CRF model. This strategy generates layered labeling sequences. The stack-LSTM [31] uses a forest structure to model nested NEs. Then, a stack-LSTM is implemented to output a set of nested NEs. Sequence-to-nuggets [18] first identifies whether a word is an anchor word of an NE with specific types. Then, a region recognizer is implemented to recognize the range of the NE relative to the anchor word. MGEPN [32] and Merge&Label are pipeline frameworks. They first generate NE candidates. Then, all candidates are further assessed by a classifier. FLAIR [25] extracts entities iteratively from outermost ones to inner ones. Strakova et al. [29] encode an input sentence into a vector representation. Then, a label sequence is directly generated from the sentence representation. Jue et al.[15] use a CNN to condense a sentence into a stacked hidden representation with a pyramid shape, where a layer represents NE candidate representations with different granularity. Each layer is stacked with an output layer to support multiobjective learning for nested NEs.

These models are all nested NE recognition architectures. Lu et al. [20] implements a CRF model on the hypergraphs. It suffers from lower performance caused by the feature sparsity problem in shallow models. All neural network-based models have higher performance. Especially in the BERT-based models, the performance is improved considerably. Compared with related work, the BR model shows an impressive result.

E. Visualization of the Regression Operation

In traditional models, the position of NEs in a sentence is denoted as discrete values. In the BR model, the *recu* feature map can be seen as a continuous semantic space. The bounding boxes have continuous parameters denoting NE locations in a continuous semantic space. The regression operation is adopted to predict the location offset of a bounding box relative to a true bounding box. At present, no research has been conducted to identify linguistic elements by the regression operation. Therefore, in this experiment, the

⁸They are known as square-shaped characters.

TABLE VI: Comparisons with Other Methods

	Models	P(%)	R(%)	F(%)
Lu et al. [20]	Mention Hypergraphs	66.3	59.2	62.5
Katiyar et al. [16]	Neural Hypergraph	70.6	70.4	70.5
Ju et al. [14]	Layered-BiLSTM-CRF	74.2	70.3	72.2
Wang et al. [31]	Stack-LSTM	76.8	72.3	74.5
Lin et al. [18]	Sequence-to-nuggets	76.2	73.6	74.9
Xia et al. [32]	MGEPN	79.0	77.3	78.2
Fisher et al. [9]	BERT+Merge&Label	82.7	82.1	82.4
Shibuya et al. [25]	BERT+FLAIR	85.94	85.69	85.82
Strakova et al. [29]	BERT+Seq2Seq	-	-	84.33
Jue et al. [15]	BERT-Pyramid	85.30	87.40	86.34
Ours	BR	86.62	86.76	86.69

regression process is visualized to show the feasibility of the BR model in locating NEs in a sentence.

A sentence “克林顿总统也将前往家乡助战” (translated as: “President Clinton will also go to his hometown to help”) is selected from the testing data. It contains two nested person names: “克林顿” (Clinton) and “克林顿总统”(President Clinton), and a location name “家乡” (hometown). A bounding box is denoted by 3 parameters s_i , l_i and c_i , which represent the start position of the box, the length of the box and the class probability of the box, respectively. In Figure 4, a bounding box is visualized as a rectangle. The horizontal ordinate represents the boundary positions of the bounding boxes in a sentence, which are normalized to $[0, 1]$. The vertical coordinate represents the class probability that a bounding box is an NE. The colors of the bounding box represent NE types.

In each subfigure of Figure 4, the selected sentence is put into the BR model with parameters learned from the training data. All outputted bounding boxes are collected and drawn with the sentence. In Figure 4(a), bounding boxes were predicted by the BR model with parameters initialized randomly without training (0 iteration). In our experiment, the BR model achieves the best performance after the iterations reached 300 rounds. From Figure 4(c) to Figure 4(f), the BR model is trained with training data with different rounds, which is denoted by the titles of the subfigures. Because the regression operation may output negative values for parameters s_i and l_i , we filter bounding boxes with $l_i \leq 0$ or $s_i + l_i > 1$ (beyond the sentence range).

In Figure 4(a), there is no tendency between bounding boxes. They are distributed evenly across the whole sentence and all NE types. In Figure 4(c), the BR model is implemented on the training data in only one round. One interesting phenomenon is that red bounding boxes and blue bounding boxes are grouped around the person name and location name. Furthermore, other positive entity types are depressed appropriately. From Figure 4(c) to Figure 4(f), when the number of iterations is increased, there are two tendencies in bounding boxes. First, the BR model is becoming more confident in the entity type prediction, which increases the classification confidence of the bounding boxes. Second, the locations of bounding boxes are approaching the true NEs. It means that the regression operation to locate NEs is feasible.

Overlapped bounding boxes is the key to solve the nested NE problem. In Figure 4(f), the bottom lines of bounding boxes matching the entity “克林顿” (Clinton) are lowered for better understanding, which shows that nested NEs are distinguished appropriately. We have tracked several bounding boxes and found that bounding boxes are not approaching true NEs smoothly and directly. There are some oscillations between them. In the training process, a bounding box may match the true NE perfectly, while moving away in the next iteration. However, in pace with the increase in the number of training steps, the oscillation tends toward stability.

V. RELATED WORK

Because the BR model is motivated by object detection from computer vision, in the following, we divide the related work into two parts: object detection and NE recognition.

Object detection is implemented in a multistage pipeline in the early stage. A typical object detection model is often composed of three stages: segmentation, feature extraction and classification. Segmentation is implemented to generate possible object locations for prediction. Generic algorithms (e.g., selective search) are often adopted to avoid exhaustively searching. Feature extraction is implemented to extract higher-order abstract features from raw input images. The output of this process is often denoted as a feature map. The feature extraction process can be encapsulated as a basic network truncated from a standard architecture for high-quality image classification, including the VGG-16 network [26], GoogLeNet [30], etc. Finally, an output layer (e.g., a linear SVM or a softmax layer) is used to predict confidence scores for each proposed region.

End-to-end object detection models can be optimized globally and share computing between inputs. These models are often similar in the feature extraction layer and output layer, where a basic network is adopted to extract a *conv* feature map, and two fully connected layers synchronously output class probabilities and object locations. The main difference is the strategy to generate the region proposal. For example, the Faster R-NN adopts anchor boxes to generate region proposals per feature map [24]. Erhan et al. [6] use a single deep neural network to generate a small number of bounding boxes. Redomon et al. [23] divide an image into grids associated with a number of bounding boxes. Liu et al. [19] use a basic

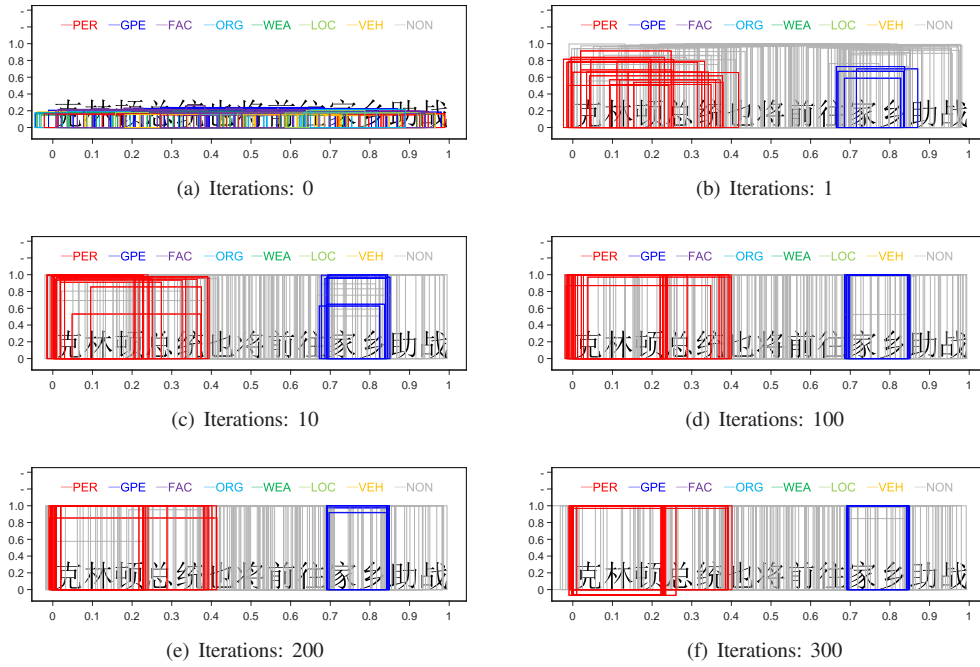


Fig. 4: Visualization of the Bounding Box Regression

network that maps an image into multiple feature maps for generating default boxes with different aspect ratios and scales.

In the field of NE recognition, neural networks have also received great attention. Early models usually adopt a sequence model to output flattened NEs (e.g., LSTM, Bi-LSTM or Bi-LSTM-CNN). To handle the nesting problem, the sequence model is redesigned. It has three variants: the layering, cascading and joint models [1]. Parsing trees are also widely used to represent nested NEs into a tree structure [8]. For example, Finkel et al. [7] use internal and structural information of parsing trees to flatten nested NEs. Zhang et al. [35] adopted a transition-based parser. Jie et al. [13] tried to capture the global dependency of a parsing tree.

Recently, many models are designed to recognize nested NEs directly. Lu et al. [20] resolve nested NEs into a hypergraph representation. Xu et al. [33] and Sohrab et al. [28] verify every possible fragment up to a certain length. Wang et al. [31] map a sentence with nested mentions to a designated forest. Ju et al. [14] proposed an iterative method that implements a sequence model in the output of a previous model. Lin et al. [18] propose a head-driven structure. Li et al. [17] combined outputs of a Bi-LSTM-CRF network with another Bi-LSTM network. Strakova et al. [29] proposed a sequence-to-sequence model. Zheng et al. [36] proposed an end-to-end boundary-aware neural model.

In Chen et al. [3], a boundary assembling (BA) model is designed to recognize nested NEs. The BA model identifies NE boundaries, assembles them into NE candidates, and picks up the most probable ones. Lately, the BA model was developed into a neural network-based approach, referred to as NNBA [2], which is effective in capturing semantic

information of a sentence by pretrained word embedding. Because NE boundaries have smaller granularity, the task to recognize them is less ambiguous and depends more on local features. Therefore, the two models are useful in recognizing nested NEs. The main drawback of them is that both utilize the cascading framework. They cannot guarantee a global optimized solution.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a boundary regression model for nested NE recognition. This model is an end-to-end multiobjective learning framework for nested NE recognition. In the future, the BR model can be extended from three aspects: 1) a more sophisticated basic network can be imported to learn high-order abstract features from input sentences. 2) The context of bounding boxes can be explored to improve performance. 3) A new non-maximum suppression algorithm can be developed to select multiscale nested NEs from overlapped positive samples.

VII. ACKNOWLEDGMENT

This work is supported by the Major Research Program of the National Natural Science Foundation of China under Grant No. 91746116, the Joint Funds of the National Natural Science Foundation of China under Grant No. U1836205, the Major Special Science and Technology Projects of Guizhou Province under Grant No. [2017]3002 and the Key Projects of Science and Technology of Guizhou Province under Grant No. [2020] 1Z055.

REFERENCES

- [1] Beatrice Alex, Barry Haddow, and Claire Grover. Recognising nested named entities in biomedical text. In *Proceedings of the BioNLP '07*, pages 65–72. ACL, 2007.
- [2] Yanping Chen, Yuefei Wu, Yongbin Qin, Ying Hu, Zeyu Wang, Ruizhang Huang, Xinyu Cheng, and Ping Chen. Recognizing nested named entity based on the neural network boundary assembling model. *IEEE IS*, 2019.
- [3] Yanping Chen, Qinghua Zheng, and Ping Chen. A boundary assembling method for chinese entity-mention recognition. *IEEE IS*, 30(6):50–58, 2015.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [5] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, pages 837–840, 2004.
- [6] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the CVPR '14*, pages 2147–2154, 2014.
- [7] Jenny Rose Finkel and Christopher D Manning. Joint parsing and named entity recognition. In *Proceedings of the HLT-NAACL '09*, pages 326–334. ACL, 2009.
- [8] Jenny Rose Finkel and Christopher D Manning. Nested named entity recognition. In *Proceedings of the EMNLP '09*, pages 141–150. ACL, 2009.
- [9] Joseph Fisher and Andreas Vlachos. Merge and label: A novel neural network architecture for nested ner. *arXiv preprint arXiv:1907.00464*, 2019.
- [10] Ross Girshick. Fast r-cnn. In *Proceedings of the ICCV '15*, pages 1440–1448, 2015.
- [11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [12] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
- [13] Zhanming Jie, Aldrian Obaja Muis, and Wei Lu. Efficient dependency-guided named entity recognition. In *Proceedings of the AAAI '17*, pages 3457–3465, 2017.
- [14] Meizh Ju, Makoto Miwa, and Sophia Ananiadou. A neural layered model for nested named entity recognition. In *Proceedings of the NAACL-HLT '19*, pages 1446–1459, 2018.
- [15] WANG Jue, Lidan Shou, Ke Chen, and Gang Chen. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the ACL '20*, pages 5918–5928, 2020.
- [16] Arzoo Katiyar and Claire Cardie. Nested named entity recognition revisited. In *Proceedings of the NAACL-HLT '18*, pages 861–871, 2018.
- [17] Fei Li, Meishan Zhang, Bo Tian, Bo Chen, Guohong Fu, and Donghong Ji. Recognizing irregular entities in biomedical text via deep neural networks. *Pattern Recognition Letters*, 2017.
- [18] Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In *Proceedings of the ACL '19*, page 5182–5192, 2019.
- [19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the ECCV '16*, pages 21–37. Springer, 2016.
- [20] Wei Lu and Dan Roth. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the EMNLP '15*, pages 857–867, 2015.
- [21] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the HLT-NAAC '03*, pages 188–191. ACL, 2003.
- [22] Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the HLT '02*, pages 82–86. Morgan Kaufmann Publishers Inc., 2002.
- [23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the CVPR '16*, pages 779–788, 2016.
- [24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *proceedings of the NIPS '15*, pages 91–99, 2015.
- [25] Takashi Shibuya and Eduard Hovy. Nested named entity recognition via second-best sequence learning and decoding. *arXiv preprint arXiv:1909.02250*, 2019.
- [26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [27] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.
- [28] Mohammad Golam Sohrab and Makoto Miwa. Deep exhaustive model for nested named entity recognition. In *Proceedings of the EMNLP '18*, pages 2843–2849, 2018.
- [29] Jana Straková, Milan Straka, and Jan Hajič. Neural architectures for nested ner through linearization. *arXiv preprint arXiv:1908.06926*, 2019.
- [30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the CVPR '15*, pages 1–9, 2015.
- [31] Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. A neural transition-based model for nested mention recognition. *arXiv preprint arXiv:1810.01808*, 2018.
- [32] Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. Multi-grained named entity recognition. *arXiv preprint arXiv:1906.08449*, 2019.
- [33] Mingbin Xu and Hui Jiang. A fofe-based local detection approach for named entity recognition and mention detection. *arXiv preprint arXiv:1611.00801*, 2016.
- [34] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *arXiv preprint arXiv:1911.03977*, 2019.
- [35] Xiantao Zhang, Dongchen Li, and Xihong Wu. Parsing named entity as syntactic structure. In *Proceedings of the ISCA '14*, 2014.
- [36] Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the EMNLP-IJCNLP '19*, pages 357–366, 2019.