International Journal of Computational Intelligence Systems (2024) 17:190
https://doi.org/10.1007/s44196-024-00595-5

RESEARCH ARTICLE

CLESR: Context-Based Label Knowledge Enhanced Span Recognition for Named Entity Recognition

Xi Chen[1] · **Wei Zhang**[1] · **Shuai Pan**[1]

Received: 16 April 2024 / Accepted: 5 July 2024

Abstract

Named entity recognition (NER) stands as a pivotal task in natural language processing, bearing profound implications for diverse downstream applications. Recent scholarship underscores the substantial performance gains attainable by integrating label knowledge into token representations, effectively treating named entity recognition as a machine reading comprehension task. Nevertheless, extant methodologies often inadequately leverage the potential of label knowledge. In response to this limitation, we introduce the Context-based Label Knowledge Enhanced Span Recognition (CLESR) architecture, designed to augment label knowledge through the assimilation of contextual information. We formulate an annotation paradigm tailored to nested scenarios, concurrently training external context-based label knowledge using conventional word association learning algorithms. The ensuing context-based label knowledge is seamlessly integrated into the model via a dedicated label attention module, thereby fortifying label learning capabilities during training. To adeptly manage both flat and nested entities, we implement a global pointer as our decoding strategy, enabling direct predictions of the positions and corresponding categories of named entities. Rigorous experimentation across six widely recognized benchmarks substantiates the efficacy of CLESR in both flat and nested tasks. Impressively, our model achieves performance enhancements surpassing those of the baseline BERT-MRC model, with gains of +0.29%, +0.77%, +0.39%, +1.9%, and +0.47% on English CoNLL2003, English OntoNotes 4.0, Chinese MSRA, English ACE2004, and English ACE2005, respectively.

Keywords Context-based label knowledge · Global pointer · Label attention · Named entity recognition

1 Introduction

The task of Named Entity Recognition (NER) has recently been approached as a span recognition problem, encompassing two subtasks: span extraction and span classification [1, 2]. The aim of the span extraction task is to determine the start and end indexes of all entities. On the other hand, the span classification task assigns one or more categories to the extracted entity. NER is a crucial component for several downstream Natural Language Processing (NLP) applications, including relation extraction [3], entity linking [4], co-reference resolution [5], and event extraction [6]. The task of NER can be further divided into two categories: flat NER and nested NER. Flat NER is typically modeled as a sequence labeling problem where each token in a sentence is assigned a single tagging class, usually using the Beginning-Inside-Outside (BIO) tagging scheme. Early approaches to this problem mainly relied on Long Short-Term Memory (LSTM) encoders with conditional random field decoders and utilized character-level and lexical features to improve performance [7–9]. With the advent of powerful contextual modeling techniques, such as BERT and RoBERTa, flat NER can be reformulated as a token-level multi-classification problem.

However, current formulations fall short in addressing the complexity of nested NER, where a shorter entity is encompassed within a longer entity. This results in a single token being assigned multiple categories. A demonstration of this scenario can be found in the ACE2004 corpora (as shown

Xi Chen and Wei Zhang contributed equally to this work.

✉ Shuai Pan
pans@aiit.org.cn

Xi Chen
xchen@aiit.org.cn

Wei Zhang
ai-lab@pku.edu.cn

[1] Advanced Institution of Information Technology Peking University, No. 233, Yonghui Rd, Hangzhou 311215, Zhejiang, China

Published online: 22 July 2024

in Fig. 1). Here, the entity "border" with the label "LOC" is nested within the entity "a border town" labeled as "GPE," which is, in turn, nested within the entity "the square in the center of Dundalk, a border town" labeled as "FAC."

Many efforts have been dedicated to the problem of nested NER. These efforts can be broadly classified into three categories: label engineering, hypergraph-based approaches, and span recognition approaches. Label engineering in nested NER mainly involves concatenating multiple labels into a single label, resulting in a sparsity problem, or grouping the nested labels in an inner-out or outside-in manner and training multiple models for each group, which is time-consuming and resource-intensive [10]. Hypergraph-based approaches require handcrafted labels, making it challenging to generalize to other domains [11–13]. In recent developments, the Machine Reading Comprehension (MRC) method has shown effectiveness in nested Named Entity Recognition (NER) tasks. This strategy involves using a question to guide the model's focus on a specific domain and predict named entities related to that question. During training and inference, both the question and text are combined as input, resulting in the model producing probabilities for named entity positions. This MRC approach helps the model concentrate on the query domain and introduces external label knowledge, significantly enhancing NER performance.

However, despite its promising performance, the MRC strategy has three limitations. Firstly, it guides the model with prompting statements but does not provide adequate contextual semantics for the labels. Secondly, it struggles to fully utilize label information in the question, often getting distracted by the text and primarily focusing on the beginning of the question. Lastly, the questions used are optimized with label annotation guidelines from previous work, making adaptation to new domains challenging.

To address these limitations and further enhance NER performance, a proposed solution is the context-based label knowledge enhancement strategy within the MRC framework. This involves simplifying the question into a basic label name and integrating context-based label knowledge pre-trained on domain-specific datasets using the word2vec algorithm. Unlike previous methods, which initialized label knowledge randomly, experiments demonstrate that this new strategy consistently outperforms random label knowledge. Additionally, simplifying the question into a label name seems to make it more adaptable to different domain datasets. Surprisingly, experiments suggest that the question might not be necessary in this framework and could even hinder performance in certain cases, potentially diverting attention away from the main query or introducing unexpected information.

Regarding the decoding strategy, the original MRC framework utilizes a pipeline span recognition method, which predicts start and end positions separately and then combines them before using a classification layer to identify the most probable spans as named entities. However, this method is prone to error propagation. Recent studies have introduced an end-to-end span recognition framework for both nested and flat NER tasks, aiming to predict named entities and categories simultaneously, thereby addressing the error propagation issue. A particularly effective framework is the global pointer [14], which has been adopted in this work to further improve NER performance.

To sum up, our contributions are as follows:

1. Integration of pre-trained label embeddings from domain-specific datasets significantly improved performance compared to random label knowledge, enhancing adaptability to different domains.
2. Simplifying questions into basic label names demonstrated potential benefits, challenging the necessity of detailed questions and their potential negative impact on performance.
3. Embracing the global pointer decoding strategy addressed error propagation issues in the original MRC framework, potentially enhancing NER performance compared to conventional methods.

2 Related Work

In the field of natural language processing, Named Entity Recognition (NER) is a commonly studied task that involves identifying and classifying named entities in text. While most existing methods approach NER as a sequence labeling problem, early approaches relied on handcrafted features and probabilistic models such as hidden Markov models and conditional random fields to predict labels for each token [10]. In recent years, deep neural networks such as convolutional neural networks and long short-term memory models have gained popularity for their ability to extract both local and global features [15]. Additionally, pretrained word and character-level embeddings have been found to complement these models [7, 8]. One such model, the deep bidirectional transformer model (e.g., BERT, Roberta), has shown a strong ability to understand context and word meaning through pretraining on large corpuses [16, 17]. By simply treating NER as a tagging classification problem with BERT, similar performance can be obtained compared to probabilistic decoding methods. However, a major drawback of these methods is their inability to handle nested entities, as only one tagging class can be assigned to a token.

Several approaches have been proposed to address the issue of nested entities in NER. One such approach is label engineering, which involves combining nested labels into a single new label or breaking them down into several unnested label groups and training multiple models [10]. Hypergraph approaches construct handcrafted tags and hyperarcs to rep-

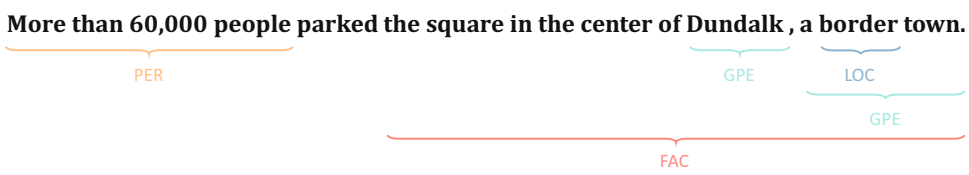

Fig. 1 Examples for nested entities from ACE2004 corpora

resent all mentions in a hypergraph and decode them similarly to linear chain conditional random fields [11–13]. Exhaustive methods enumerate all mentions and classify them using deep neural features, while boundary selection methods reduce the number of classified mentions by using another module like a pointer network [18–20]. However, these approaches are prone to high time complexity, error propagation, and label isolation problems. In comparison, the global pointer model [14] approaches nested NER by concurrently classifying both the positions of named entities and their respective categories in an end-to-end fashion.

Furthermore, in addition to different decoding methods, recent studies have introduced two innovative approaches to improve NER performance: the MRC framework and label knowledge enhancement techniques [1, 7]. The MRC framework treats NER as a machine reading comprehension task, combining plain text with domain-specific queries. On the other hand, label knowledge enhanced methods integrate label information at the feature level using a label attention module. However, the previous MRC framework lacks consideration for contextual semantics related to labels. Additionally, the domain-specific nature of task-specific queries poses challenges when applying these methods to other domains, especially with larger entity types. To overcome these limitations, we propose a new approach that enriches token representation through context-based label knowledge and simplifies task-specific queries by using complete label names, thereby facilitating application in various domains. Our experiments in Sect. 4.4 validate the effectiveness of our proposed model.

3 Proposed Method

In this section, we provide a detailed introduction of CLESR, which is depicted in Fig. 2. The framework of CLESR comprises three critical modules, which are explained as follows: (1) The semantics encoding module that encodes token representations using contextualized information; (2) The label attention module that enhances token features by adding pretrained, context-based label knowledge; and (3) The span recognizing module that predicts whether each token span is the specified category or not. We will describe all these components in detail in the following sections.

3.1 Task Formulation

In this work, we propose a method for named entity recognition that utilizes both the input text and query to generate the final model input. Specifically, given an input text $X = (x_1, x_2, \ldots, x_n)$ consisting of n tokens and an input query $Q = (q_1, q_2, \ldots, q_m)$ consisting of m tokens, we define the combination of query and text as $QX = (q_1, \ldots, q_m, x_1, \ldots, x_n)$. We use (QX, Y) to denote a training example, where Y is a set of triples $Y = \langle Y_c^l, Y_c^r, Y_c \rangle$, with $Y_c^l \in [0, N-1]$, $Y_c^r \in [0, N-1]$, and $Y_c \in C$ denoting the left boundary, right boundary indices, and c-th entity type, respectively. Here, C is a predefined set of total categories. To incorporate label knowledge, we use a label attention module to integrate M, the label knowledge represented as $|C|$ vectors of size 300, into the feature level. Instead of random label initialization, we propose incorporating robust context-based label instances to enhance the label knowledge learning ability. Therefore, the task is defined as follows: given an input sentence X and input queries Q, the aim is to extract the entities Y based on the queries Q and the label instances M.

3.2 Semantic Encoding Module

In recent years, deep bidirectional attention-based models pretrained on large corpora have demonstrated their strong ability to learn contextual information and the meaning of words. Several architectures, such as BERT, Roberta, XLNet, and Electra, have been proposed. To make a fair comparison with previous methods, we adopt Roberta as our Semantic Encoding Module (SEM). Since Roberta uses the same architecture as BERT, we use the term "BERT" in the later sections for brevity. As previously described, BERT adopts the concatenation of an input text and query, denoted as QX, and the final token representation $H^t \in \mathcal{R}^{n \times d}$ is then fed into the next layer which can be shown in Eq. 1:

$$H^t = BERT(QX). \tag{1}$$

3.3 Label Attention Module

The goal of the Label Attention Module (LAM) is to incorporate label information with contextual semantics to each

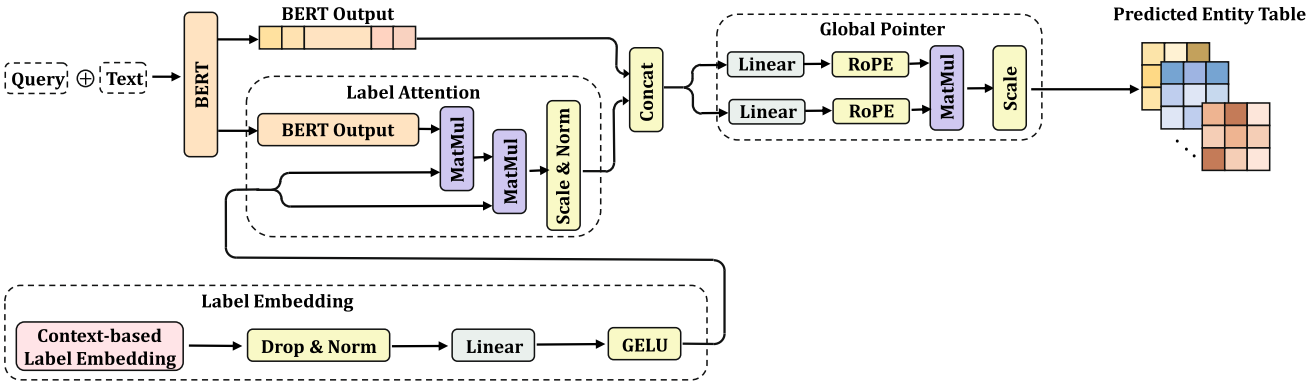

Fig. 2 The overall model architecture of our CLESR

token, supplementing the MRC framework and enhancing the learning ability of label knowledge. To explain the label attention module, we first introduce the methods used to obtain the context-based label instance. Traditional label instances are initialized randomly and updated during the training process, but this approach can fall into local optima due to the reliance on random seeds. To address this limitation, we construct the context-based label instance by concatenating the original text with the annotated text. Previous work [8] annotated the sentence by masking all entities with their corresponding entity types. However, this approach has two drawbacks. Firstly, it can not properly deal with nested entities, when annotate longer named entity, the short one is ignored which leading to information loss. Secondly, it is challenging in addressing the situation where multiple named entities are adjacent to each other. If all adjacent named entities are annotated, each label is hard to get proper context knowledge. In order to overcome the aforementioned limitations, we propose a new annotation approach which annotates sentences separately, i.e., one entity per sentence. This enables the entity type to fully exploit its context and learn complete contextual knowledge. More details can be found in Appendix C. To train the context-based label instance, we utilize word2vec.

The pre-trained context-based label instance is first fed into the drop and norm layer to alleviate overfitting. Next, a linear layer is used to modify the dimension of the label instance to the same dimension as the BERT output. After that, it is activated by the Gaussian Error Linear Unit (GELU) operation for better non-linear approximation. The final label embedding input to the label attention is formulated in Eqs. 2 and 3.

$$M = LayerNorm(Dropout(\mathcal{M})), \tag{2}$$

$$\mathcal{M}' = GELU(MW + b), \tag{3}$$

where $\mathcal{M}$ and $M \in \mathcal{R}^{n\times 300}$ denote the pretrained context-based label instance and the regularized label instance, respectively. $W \in \mathcal{R}^{300\times d}$ and $b \in \mathcal{R}^{d}$ denote the learnable weights and bias. $\mathcal{M}' \in \mathcal{R}^{n\times d}$ denotes the final label embedding.

We incorporate label knowledge into token representation using the attention mechanism (i.e., scaled dot product). Here, the token representation is treated as a query and the label representation is treated as a key and value. The attention mechanism produces an attention matrix α, which consists of a label distribution for each token. The label-enhanced token representation H'^{l} is then calculated in Eqs. 4, 5 and 6.

$$\alpha = Softmax\left(\frac{H^{t}\mathcal{M}'^{T}}{\sqrt{d}}\right), \tag{4}$$

$$H^{l} = \alpha\mathcal{M}', \tag{5}$$

$$H'^{l} = LayerNorm(Dropout(H^{l})). \tag{6}$$

The final input $H \in \mathcal{R}^{n\times 2d}$ to the span recognition module is the concatenation of the token representation and label enhanced token representation which is shown in Eq. 7.

$$H = [H^{w}; H'^{l}]. \tag{7}$$

3.4 Span Recognition Module

The goal of the Span Recognition Module (SRM) is to simultaneously extract and classify the token spans. Accurate position information is necessary to extract token spans precisely. Previous methods [21, 22] typically employ multi-layer LSTMs to enhance token representation with position information, resulting in increased time complexity. Instead, we propose the use of a global pointer [14] as the decoding layer, which adds position information simply by using Rotary Position Embedding (RoPE [23]). RoPE provides consistent position information and has been shown to be effective in many large pretraining networks [24, 25].

Global pointer employs a biaffine model with RoPE to create a scoring tensor table $s \in \mathcal{R}^{|C|\times n\times n}$. It first uses two linear layers to get the token start representation $H^s \in \mathcal{R}^{|C|\times n\times d_g}$ and token end representation $H^e \in \mathcal{R}^{|C|\times n\times d_g}$ as shown in Eqs. 8 and 9.

$$H^s = Reshape(HW_s + b_s), \tag{8}$$
$$H^e = Reshape(HW_e + b_e), \tag{9}$$

where W_s and $W_e \in \mathcal{R}^{2d\times|C|\times d_g}$ are learnable weights, and b_s and $b_e \in \mathcal{R}^{|C|\times d_g}$ are learnable biases. It is worth noting that the number of entity types can be treated as head numbers, and d_g is the head dimension. Equations 8 and 9 can be used to build multi-head inputs. RoPE is then added to the token representations above calculated in Eqs. 10 and 11.

$$H'^s = RoPE(H^s), \tag{10}$$
$$H'^e = RoPE(H^e). \tag{11}$$

The final scoring tensor is calculated by dot product shown in Eq. 12:

$$s = H'^s H'^{eT} + m. \tag{12}$$

The upper triangle loss is calculated using a mask tensor, $m \in \mathcal{R}^{n\times n}$, where the upper triangles are filled with 0 and the lower triangles are filled with $-inf$. As the score matrix is symmetric, only the upper triangle loss is computed to avoid duplication of the same entities during inference.

3.5 Training Objective

In the span recognition task, each span in the upper triangle of the prediction table is assigned one label from a set of categories. However, this approach leads to a class imbalance problem as there are $n(n+1)/2$ classification tasks for a sentence and more than 90% of the entries are negative with the label NAN. To address this issue, we used the class-balanced loss proposed in [14], which is a variant of cross-entropy. The class-balanced loss helps in mitigating the class imbalance problem. The formula for the loss function is given by Eq. 13.

$$\mathcal{L} = log\left(1 + \sum_{i\in\Omega_{neg}} exp(s_i)\right) + log\left(1 + \sum_{j\in\Omega_{pos}} exp(-s_j)\right). \tag{13}$$

The class balanced loss aims to ensure that all scores of negative samples are less than zero and all scores of positive samples are greater than zero. Consequently, during inference, we simply select the token spans with scores greater than 0 as our predicted entities, which is more effective compared to first extracting the start and end positions and then classifying the spans using the permutations of start and end indexes. The complete training procedure for our proposed approach is outlined in Algorithm 1.

Algorithm 1 The training algorithm

Input: The training Query Q and training Text X
Output: The configurations of the model with parameters θ.
Initialize: The model with parameters θ.
for e ∈ {1,...,epoch} **do**
  for $q_i, x_i \in (Q, X)$ **do**
    Concatenate query and text → $[q_i, x_i]$
    BERT outputs query and text embed → h_i^t
    Preprocess pretrained label embed → $m_i^{'}$
    LAM outputs label knowledge enhanced token representation → h_i
    SRM outputs token predictions → s
    UPDATE θ with $\mathcal{L}$
  end for
end for

4 Experiments

In this section, we present CLESR results on 6 widely-used benchmarks.

4.1 Dataset

In this study, we evaluate our proposed model for NER on both flat and nested tasks. For flat NER, we conduct experiments on three datasets, namely Chinese MSRA, Chinese OntoNote 4.0, and English CoNLL2003. On the other hand, for nested NER, we evaluate our model's performance on three different datasets: English GENIA, English ACE2004, and English ACE2005.

The six datasets used in our experiments have varying numbers of entity categories. For the GENIA and MSRA datasets, which contain five and three types of categories, respectively, we use the same data split as that used in [1] to ensure a fair comparison. For Chinese OntoNotes 4.0, which has four entity types, we use the data split provided by [26] in our experiments. In the case of the CoNLL2003 dataset, which contains four named entities, we follow the data processing protocols described in [15]. Both the ACE2004 and ACE2005 datasets have seven entity categories, and we adopt the data processing strategy presented in [11, 27] for these datasets.

4.2 Implementation Details

We employ BERT architecture as our contextualized representation encoder for the input texts and queries. More specifically, we implement our model based on the RoBerta-large model for all datasets including both flat and nested NER tasks. For Chinese datasets, we use the Chinese-RoBerta-wwm-ext-large [28] and for English datasets, we use the English-RoBerta-large [17]. The context-based label embedding is pretrained by word2vec with CBOW and negative sampling strategies. We train the word2vec model to learn 300-dimensional vectors with a minimum word frequency cutoff of 5, a context window size of 5, and 10 epochs. The head dimension d_g in global pointer is set to 64.

For all datasets, we adopted the adam optimizer [29] with the peak learning rate of 2e−5 and a linear decay schedule with a warmup proportion of 0.1. The training is under the setting of 8 batch size and 16 gradient accumulation steps for 10 epcohs. Our test results is from the checkpoints of best scores in validation set. All of the models are implemented using Pytorch and are trained on GPU GeForce GTX 2080 Ti.

4.3 Main Results

In contrast to prior approaches that consider flat Named Entity Recognition (NER) tasks as a sequence labeling challenge and necessitate novel methods for handling nested tasks, our proposed Context-based Label Knowledge Enhanced Span Recognition (CLESR) offers a unified framework for both flat and nested NER tasks. We have showcased the performance of CLESR in Tables 1 and 2 across six widely recognized benchmark datasets. Notably, it outperforms the baseline model BERT-MRC and achieves comparable results to state-of-the-art work such as PIQN.

In the realm of flat Named Entity Recognition (NER) datasets, our CLESR demonstrates superior performance across three prominent benchmarks: English CoNLL2003, English OntoNotes 4.0, and Chinese MSRA. Notably, we achieve performance enhancements of +0.29%, +0.77%, and +0.39%, respectively. In the context of nested NER datasets, our CLESR outperforms BERT-MRC on English ACE2004 and English ACE2005 by margins of +1.9% and +0.47%, respectively. Particularly noteworthy is the substantial improvement observed in the English ACE2005 dataset, a phenomenon shared with the state-of-the-art model PIQN. We posit that this improvement stems from a shared characteristic between PIQN and CLESR, wherein both models eschew the utilization of protracted queries. Instead, they derive label knowledge from instance features during training. This departure from lengthy queries is theorized to mitigate noise introduction and mitigate adverse effects on specific datasets. In this vein, our context-based label knowledge, augmented by label attention, emerges as a mechanism that effectively fortifies label learning without the encumbrance of extraneous information.

Table 1 Experimental results for flat tasks

Model	P	R	F1
English CoNLL2003			
BiLSTM-CRF [15]	–	–	91.03
ELMo [30]	–	–	92.22
CVT [31]	–	–	92.6
BERT-Tagger [16]	–	–	92.8
PIQN [32]	93.29	92.46	92.87
BERT-MRC [1]	92.33	94.61	93.04
CLESR	0.9297	0.937	**93.33 (+ 0.29)**
Chinese OntoNotes 4.0			
Lattice-LSTM [33]	76.35	71.56	73.88
BERT-Tagger [16]	78.01	80.35	79.16
Glyce-BERT [27]	81.87	81.40	81.63
BERT-MRC [1]	82.98	81.25	82.11
CLESR	0.8213	0.8363	**82.88 (+ 0.77)**
Chinese MSRA			
Lattice-LSTM [33]	93.57	92.79	93.18
PIQN [1]	93.61	93.35	93.48
BERT-Tagger [16]	94.97	94.62	94.80
Glyce-BERT [27]	95.57	95.51	95.54
BERT-MRC [1]	96.18	95.12	95.75
CLESR	96.45	95.83	**96.14 (+ 0.39)**

Our findings are highlighted by the improvements we achieved over the baseline model BERT-MRC which is emphasized in italic
Bold values represent the best-performing metrics among all methods

The observed outcomes underscore the efficacy of CLESR, a phenomenon attributed to several key factors. Notably, the adoption of end-to-end span recognition strategies for simultaneous prediction of named entity positions and categories presents a departure from conventional two-step extraction processes. This departure mitigates error propagation, consequently elevating the overall performance of the model. The incorporation of a label attention module further distinguishes CLESR by integrating context-based label knowledge. This integration steers the model toward an optimal state in contrast to random label embeddings. Recognizing the tendency of queries to be perturbed by textual distractions and underutilize available label knowledge, we introduce concise full names for entity types, encapsulating pivotal information. This context-based label knowledge synergistically complements the question-answering formulation, thereby augmenting the likelihood of effective label knowledge acquisition and, consequently, enhancing model performance. Furthermore, the strategic integration of position information, facilitated by RoPE position embeddings,

Table 2 Experimental results for nested tasks

Model	P	R	F1
English GENIA			
Hyper-Graph [34]	77.7	71.8	74.6
ARN [20]	75.8	73.9	74.8
Path-BERT [35]	78.07	76.45	77.25
DYGIE [36]	–	–	76.2
Seq2Seq-BERT [37]	–	–	78.31
PIQN [32]	83.24	80.35	81.77
BERT-MRC [1]	85.18	81.12	**83.75**
CLESR	0.8172	0.8239	82.05
English ACE2004			
Seq-Graph [12]	78.0	72.4	75.1
Seq2Seq-BERT [37]	–	–	84.40
DYGIE [36]	–	–	84.7
Biaffine-NER [22]	87.3	86.0	86.7
PIQN [32]	88.48	87.81	88.14
BERT-MRC [1]	85.05	86.23	85.98
CLESR	86.96	88.82	**87.88 (+ 1.9)**
English ACE2005			
Seq-Graph [12]	76.8	72.3	74.5
DYGIE [36]	–	–	82.9
Seq2Seq-BERT [37]	–	–	84.33
Biaffine-NER [22]	85.2	85.6	85.4
PIQN [32]	86.27	88.60	87.42
BERT-MRC [1]	87.16	86.59	86.88
CLESR	87.62	87.09	**87.35 (+ 0.47)**

Our findings are highlighted by the improvements we achieved over the baseline model BERT-MRC which is emphasized in italic
Bold values represent the best-performing metrics among all methods

contributes to the overall efficiency of CLESR in contrast to the utilization of multi-layer LSTMs.

4.4 Ablation Study

In this section, we conduct a comprehensive evaluation of various iterations of the proposed Context-based Label Knowledge Enhanced Span Recognition (CLESR) model. The evaluation, presented in Table 3, encompasses six benchmarks, revealing that the integration of context-based label knowledge consistently enhances performance in comparison to models lacking a label attention module. Specifically, notable performance gains of +0.68%, +0.47%, +0.35%, +0.84%, +0.97%, and +0.96% are discerned across CoNLL2003, OntoNotes4.0, MSRA, GENIA, ACE2004, and ACE2005, respectively. Moreover, it is noteworthy that the model employing random initialization yields inferior results in Chinese OntoNotes4.0 and ACE2005 when compared to the model without a label attention module. This observation suggests that random label knowledge may not consistently furnish beneficial information and could potentially impede the model's efficacy in certain scenarios. These findings underscore the consistent and reliable provision of valuable information by context-based label knowledge, substantiating its role in effecting a persistent enhancement in model performance.

Furthermore, Table. 4 showcases the impact of three other components of CLESR on GENIA and ACE2004 datasets. The outcomes reveal that the global pointer and the proposed annotation paradigm outperform pipelined span recognition methods and traditional annotation paradigms. Additionally, regarding the necessity of detailed questions, the results demonstrate that a simple label name suffices, making it easily applicable to various domain-specific datasets.

4.5 Effect on LSTM Encoding

To investigate the effectiveness of our proposed label knowledge integration methods, we conducted experiments on the Chinese OntoNotes 4.0 dataset without utilizing any

Table 3 Impact of context-based label knowledge on six benchmarks

	CoNLL2003			OntoNotes 4.0			MSRA		
	P	R	F1	P	R	F1	P	R	F1
Default	92.97	93.7	**93.33**	82.13	83.63	**82.88**	96.45	95.83	**96.14**
w/o CLK	92.80	93.38	93.09	80.99	83.38	82.17	95.85	95.68	95.77
w/o LA	92.14	93.17	92.65	84.04	80.84	82.41	95.97	95.61	95.79
	GENIA			ACE2004			ACE2005		
	P	R	F1	P	R	F1	P	R	F1
Default	81.72	82.39	**82.05**	86.96	88.82	**87.88**	87.62	87.09	**87.35**
w/o CLK	80.81	82.17	81.48	86.43	88.46	87.43	86.2	86.06	86.13
w/o LA	80.19	82.27	81.21	85.84	88.02	86.91	85.89	86.90	86.39

w/o LA remove the label attention module, *w/o CLK* replace the context-based label knowledge with random initialization
The bold values indicate the best-performing metrics for CLESR

Table 4 Assessing the performance on GENIA and ACE2004 datasets using different modifications of the proposed methods

	GENIA			ACE2004		
	P	R	F1	P	R	F1
Default	81.72	82.39	**82.05**	86.96	88.82	87.88
w/o GP	81.52	82.14	81.82	86.52	88.66	87.57
w/o Anno	81.57	81.96	81.76	86.79	88.54	87.65
w/o LN	81.66	82.23	81.94	86.97	88.92	**87.93**

w/o GP replace global pointer with pipelined span recognition, *w/o Anno* replace the proposed annotation paradigm with traditional annotation, *w/o LN* replace full label name with question
The bold values indicate the best-performing metrics for CLESR

Table 5 The effectiveness of label attention when using LSTM encoder

Chinese OntoNotes 4.0 (Flat)			
Model	P	R	F1
Default	70.01	57.24	**62.99**
w/o CLK	73.63	52.02	60.97
w/o LA	67.12	55.18	60.57

The bold value indicates the best-performing metrics for CLESR

pre-trained deep neural networks. We replaced the RoBerta encoder with a single bidirectional LSTM layer, and used the BERT tokenizer and a randomly initialized lookup table to minimize modification. We set the character size, LSTM hidden size, total epochs, and dropout rate to 21,128, 384, 10, and 0.5, respectively. Our results, presented in Table 5, demonstrate that our proposed label knowledge method significantly outperforms the random knowledge integration method, increasing the F1 score by up to 2.42%. These findings imply that leveraging context-based label knowledge offers valuable insights and points towards a promising path for creating effective text encoders in industrial sectors with potentially limited computational resources.

4.6 Analysis of the Context-Based Label Knowledge

In Fig. 3, a visual representation in two dimensions illustrates the context-based label representation and their closest words on the GENIA dataset, employing the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. Notably, the proximity of words surrounding the label name forms clusters that predominantly consist of annotated named entities. This suggests that the context-based label instances encapsulate a wealth of knowledge specific to a distinct class of entities, demonstrating a grasp of contextual semantics.

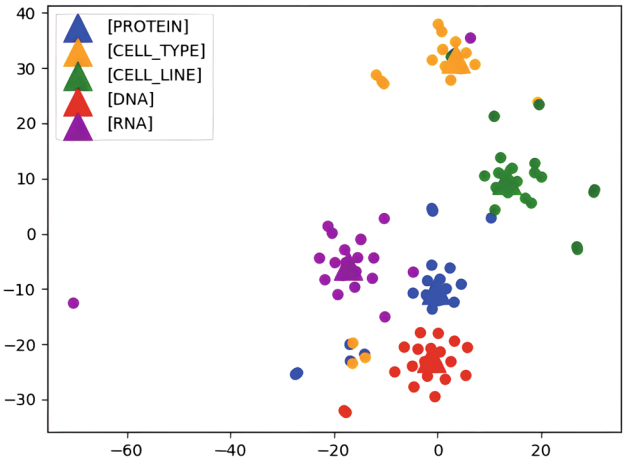

Fig. 3 Representation of entity categories and their nearest words on two-dimensional space using T-SNE algorithm on GENIA dataset. Nearest words examples: (1) Protein: ["Ca(2+)-regulated", "kinase/c-Jun", "4G10",...]; (2) Cell-Type: ["lymphocytes", "monocytes", "peripheral",...]; (3) Cell-Line: ["myelomonoblasts", "ML-3", "HPB.ALL",...]; (4) DNA: ["promoter", "gene", "(CT)19/(CA)16",...]; (5) RNA: ["mRNA", "polyadenylated", "RNA",...]

5 Conclusion

This study introduces the Context-based Label Knowledge Enhanced Span Recognition (CLESR) model, designed to address both flat and nested NER. The integration of a label attention module plays a pivotal role in infusing context-based label knowledge into the model, thereby augmenting label learning capabilities and leading to a noteworthy enhancement in overall performance. Our devised annotation strategy serves to further optimize the model's efficacy, particularly in the context of nested NER tasks. Moreover, the incorporation of a global pointer is introduced to effectively handle both flat and nested NER assignments, resulting in expedited inference times. Rigorous experimentation conducted across a diverse set of nested and flat NER datasets serves to empirically validate the efficacy of our proposed model.

Appendix 1: Dataset Statistics

We report the number of sentences, the average sentence length, the total number of entities on all datasets, and report the number of sentences containing nested entities, the number of nested entities on nested datasets. The statistics are shown in Table 6.

Appendix 2: Baseline Models

We use the following models as baselines:

Table 6 Statistics of the datasets used in experiments

	CoNLL2003			OntoNotes 4.0			MSRA		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
$\#S$	14,041	3250	3453	15,724	4301	4347	41,728	4636	4365
$\#NS$	–	–	–	–	–	–	–	–	–
$\#AL$	14.50	15.80	13.45	31.28	46.61	47.87	46.87	46.17	39.54
$\#TE$	23,499	5942	5648	13,372	6950	7684	70,446	4257	6181
$\#NE$	–	–	–	–	–	–	–	–	–
	GENIA			ACE2004			ACE2005		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
$\#S$	16,692	–	1854	6200	745	812	7194	969	1047
$\#NS$	3522	–	446	2712	294	388	2691	338	320
$\#AL$	25.35	–	25.99	22.50	23.02	23.05	19.21	18.93	17.2
$\#TE$	50,509	–	5506	22,204	2514	3035	24,441	3200	2993
$\#NE$	9064	–	1199	2712	294	388	2691	338	320

$\#S$ number of sentences, $\#NS$ number of sentences containing nested entities, $\#AL$ average sentence length, $\#TE$ total number of entities, $\#NE$ number of nested entities

- **BiLSTM-CRF** [15] is the model utilized for flat NER, employing BiLSTM to encode tokens and CRF to predict token labels.
- **Lattice-LSTM** [33] constructs a word-character lattice to enhance the LSTM performance.
- **ELMo** [30] is the model utilized for flat NER with deep contextualized word representation.
- **CVT** [31] is the model utilizes Cross-View Training to enhance the representations of a Bi-LSTM encoder.
- **BERT-Tagger** [16] treats the flat NER as a tagging problem with pretrained BERT model.
- **Glyce-BERT** [27] combines glyph information for Chinese NER
- **Hyper-Graph** [34] proposes a hypergraph-based approach built on LSTMs for nested NER tasks.
- **ARN** [20] introduces Anchor Region Networks (ARN) which initially identifies anchor words and subsequently identifies the boundaries of mentions.
- **Path-BERT** [35] processes the tag sequence as the second-best path within the span of their parent entity relying on BERT.
- **DYGIE** [36] is the model that predicts token spans with dynamically constructed span graphs.
- **seq2seq-BERT** [37] treats the nested NER as a sequence to sequence problem with pretrained BERT model.
- **Biaffine-NER** [22] predicts token spans using a dependency parsing approach.
- **BERT-MRC** [1] approaches NER as a machine reading comprehension task.
- **PIQN** [32] is a model that simultaneously learns the label instance and extracts entities in parallel.

Appendix 3: Demonstration of Annotation Paradigm

In this section, we demonstrate comparisons between our proposed annotation approaches and traditional annotation approaches. As show Table 7, in plain text 1, "The same part of the world" is the entity labeled with "[LOC]" and the inner entity "the world" is also labeled with "[LOC]," so the shorter embedded entities will be ignored if we choose to annotate longer entities. On the other hand, if we choose to annotate shorter entity, the longer one will be lost. Therefore, the traditional annotation methods will loss information in nested NER. There is another problem, for instance, in plain text 2, "the c-fos and c-jun genes" can be labeled as "[DNA] and [DNA]," but the first "[DNA]" loses the latter context "c-jun genes", and the second "[DNA]" loses the previous context "c-fos".

Appendix 4: Case Study

In this section, we showcase the outcomes derived from the CONLL2003 dataset and the GENIA dataset, as illustrated in Fig. 4. During the inference phase, each text undergoes concatenation with a specific query at the beginning, enabling the model to focus on forecasting entities related to the query. For CONLL2003 dataset, the CLESR accurately identifies "JAPAN" as a location in the provided examples. However, an erroneous prediction occurs with "CHINA,"

Table 7 Comparison of context label annotation approaches

Plain text 1	(1) The same part of the world , high winds, high seas, three fishermen rescued near Vancouver by a royal Caribbean cruise ship
Previous annotation	(2) [LOC], high winds, high seas, [PER] rescued near [GPE] by [VEH]
Our annotation	(3) [LOC], high winds, high seas, three fishermen rescued near vancouver by a royal caribbean cruise ship
	(4) The same part of [LOC], high winds, high seas, three fishermen rescued near vancouver by a royal caribbean cruise ship
	(5) The same part of the world, high winds, high seas, three fishermen rescued near vancouver by [VEH]
	(6) The same part of the world, high winds, high seas, three fishermen rescued near vancouver by a [ORG] cruise ship
Plain text 2	(7) The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes
Previous annotation	(8) The inhibition of [DNA] and c-jun expression by [protein] in [cell_type] was shown to be due to a lower transcription rate of the [DNA] and [DNA]
Our annotation	9) The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the [DNA] and c-jun genes
	(10) The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and [DNA]
	(11) The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the [DNA]

This table demonstrates the difference between our approach to context label annotation and previous approaches, highlighting the challenges in properly handling nested entities and continuous entities in the ACE2005 and GENIA corpora

Input: *Location* SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . **Output**: {"*Location*": ["JAPAN", "CHINA"]}
Input: *Organization* SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . **Output**:{"*Organization*": []}
Input: *Name of miscellaneous* SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . **Output**:{"*Name of miscellaneous*": []}
Input: *Persen* SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT . **Output**:{"*Persen*": []}
Input: *DNA* The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes . **Output**: {"DNA": ["c-fos", "c-fos", "c-jun genes", "c-fos and c-jun genes"]}
Input: *RNA* The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes . **Output**: {"RNA": []}
Input: *protein* The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes . **Output**: {"protein": ["IL-4"]}
Input: *cell type* The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes . **Output**: {"cell type": ["LPS-treated cells"]}
Input: *cell* line The inhibition of c-fos and c-jun expression by IL-4 in LPS-treated cells was shown to be due to a lower transcription rate of the c-fos and c-jun genes . **Output**: {"cell line": []}

Fig. 4 Illustrative examples of input and output from the CONLL2003 dataset and the GENIA dataset. The query is highlighted in italics, and the predicted results that are correct are shown in blue, while the predicted results that are incorrect are displayed in red

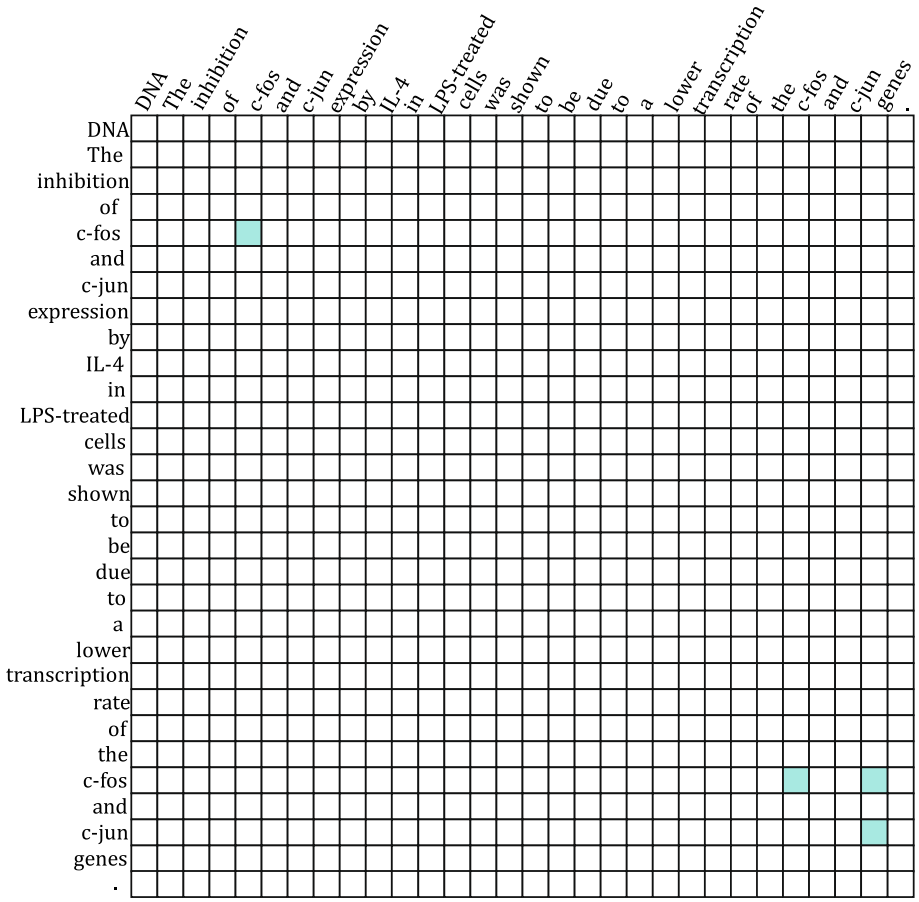

Fig. 5 The prediction table for a nested entity sample from the GENIA dataset. The extracted location is highlighted in aqua green

where the CLESR wrongly labels it as a location despite it being annotated as a person. This clearly demonstrates the capability of CLESR to mitigate the influence of incorrectly annotated labels by emphasizing the correct label, which in this case is "location," and disregarding the erroneous label of "person." The CLESR model demonstrates accurate identification of all entities, including nested entities like "c-fos," "c-jun genes," and "c-fos and c-jun genes," within the GENIA dataset. To provide a clear understanding of how nested entities are extracted, we present the prediction table in Fig. 5. Each input sentence contains a query that instructs the model to focus on extracting entities relevant to that query. In this example, the location of nested entities such as (25, 25), (25, 28), and (27, 28) are predicted and highlighted in aqua green. The first value in each tuple represents the start index, while the second value indicates the end index of the extracted entities. As a result, CLESR effectively handles nested entities by associating them with the corresponding query.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xi Chen, Wei Zhang and Shuai Pan. The first draft of the manuscript was written by Xi chen and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The research leading to these results received funding from the National Key Research and Development Program of China under Grant Agreement No 2022YFF0903302.

Data and code availability and access Our data and code are available at https://github.com/hanggun/CLESR.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical and informed consent for data used The data employed in this study has been obtained with proper ethical considerations and informed consent.

References

1. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5849–5859. Association for Computational Linguistics, Online (2020). https://aclanthology.org/2020.acl-main.519
2. Yang, P., Cong, X., Sun, Z., Liu, X.: Enhanced language representation with label knowledge for span extraction. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4623–4635. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://aclanthology.org/2021.emnlp-main.379
3. Wei, Z., Su, J., Wang, Y., Tian, Y., Chang, Y.: A novel cascade binary tagging framework for relational triple extraction. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, pp. 1476–1488. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.136
4. Le, P., Titov, I.: Improving entity linking by modeling latent relations between mentions. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1595–1604. Association for Computational Linguistics, Melbourne, Australia (2018). https://aclanthology.org/P18-1148
5. Kirstain, Y., Ram, O., Levy, O.: Coreference resolution without span representations. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 14–19. Association for Computational Linguistics, Online (2021). https://aclanthology.org/2021.acl-short.3
6. Yang, H., Sui, D., Chen, Y., Liu, K., Zhao, J., Wang, T.: Document-level event extraction via parallel prediction networks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 6298–6308. Association for Computational Linguistics, Online (2021). https://aclanthology.org/2021.acl-long.492.pdf
7. Cui, L., Zhang, Y.: Hierarchically-refined label attention network for sequence labeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4115–4128. Association for Computational Linguistics, Hong Kong, China (2019). https://aclanthology.org/D19-1422
8. Ghaddar, A., Langlais, P.: Robust lexical features for improved neural network named-entity recognition. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1896–1907. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://aclanthology.org/C18-1161
9. Ma, R., Peng, M., Zhang, Q., Wei, Z., Huang, X.: Simplify the usage of lexicon in Chinese NER. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5951–5960. Association for Computational Linguistics, Online (2020). https://aclanthology.org/2020.acl-main.528
10. Alex, B., Haddow, B., Grover, C.: Recognising nested named entities in biomedical text. In: Biological, Translational, and Clinical Language Processing, pp. 65–72. Association for Computational Linguistics, Prague, Czech Republic (2007). https://aclanthology.org/W07-1009
11. Lu, W., Roth, D.: Joint mention extraction and classification with mention hypergraphs. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, pp. 857–867. The Association for Computational Linguistics, Online (2015). https://doi.org/10.18653/v1/d15-1102
12. Wang, B., Lu, W.: Neural segmental hypergraphs for overlapping mention recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 204–214. Association for Computational Linguistics, Brussels, Belgium (2018). https://aclanthology.org/D18-1019
13. Muis, A.O., Lu, W.: Labeling gaps between words: Recognizing overlapping mentions with mention separators. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2608–2618. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://aclanthology.org/D17-1276
14. Su, J., Murtadha, A., Pan, S., Hou, J., Sun, J., Huang, W., Wen, B., Liu, Y.: Global pointer: novel efficient span-based approach for named entity recognition. arXiv (2022). arXiv:2208.03054
15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (2016). https://aclanthology.org/P16-1101
16. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Online (2019). https://doi.org/10.18653/v1/n19-1423
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: a robustly optimized BERT pretraining approach (2019). arXiv:1907.11692
18. Xu, M., Jiang, H., Watcharawittayakul, S.: A local detection approach for named entity recognition and mention detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1237–1247. Association for Computational Linguistics, Vancouver, Canada (2017). https://aclanthology.org/P17-1114
19. Sohrab, M.G., Miwa, M.: Deep exhaustive model for nested named entity recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2843–2849. Association for Computational Linguistics, Brussels, Belgium (2018). https://aclanthology.org/D18-1309
20. Lin, H., Lu, Y., Han, X., Sun, L.: Sequence-to-nuggets: nested entity mention detection via anchor-region networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5182–5192. Association for Computational Linguistics, Florence, Italy (2019). https://aclanthology.org/P19-1511
21. Wan, J., Ru, D., Zhang, W., Yu, Y.: Nested named entity recognition with span-level graphs. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 892–903. Association for Computational Linguistics, Dublin, Ireland (2022). https://aclanthology.org/2022.acl-long.63
22. Yu, J., Bohnet, B., Poesio, M.: Named entity recognition as dependency parsing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6470–6476. Association for Computational Linguistics, Online (2020). https://aclanthology.org/2020.acl-main.577
23. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: RoFormer: enhanced transformer with rotary position embedding. arXiv (2021). arXiv:2104.09864

24. al., C.: PaLM: Scaling language modeling with pathways. arXiv (2022). arXiv:2204.02311
25. Hua, W., Dai, Z., Liu, H., Le, Q.: Transformer quality in linear time. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 9099–9117. PMLR, Online (2022). https://proceedings.mlr.press/v162/hua22a.html
26. Sun, Z., Li, X., Sun, X., Meng, Y., Ao, X., He, Q., Wu, F., Li, J.: ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. 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), pp. 2065–2075. Association for Computational Linguistics, Online (2021). https://aclanthology.org/2021.acl-long.161
27. Meng, Y., Wu, W., Wang, F., Li, X., Nie, P., Yin, F., Li, M., Han, Q., Sun, X., Li, J.: Glyce: Glyph-vectors for chinese character representations. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 2742–2753 (2019). https://proceedings.neurips.cc/paper/2019/hash/452bf208bf901322968557227b8f6efe-Abstract.html
28. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. Audio Speech Lang. Proc. **29**, 3504–3514 (2021)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015). arXiv:1412.6980
30. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://aclanthology.org/N18-1202
31. Clark, K., Luong, M.-T., Manning, C.D., Le, Q.V.: Semi-supervised sequence modeling with cross-view training. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1914–1925. Association for Computational Linguistics, Online (2018). https://aclanthology.org/D18-1217.pdf
32. Shen, Y., Wang, X., Tan, Z., Xu, G., Xie, P., Huang, F., Lu, W., Zhuang, Y.: Parallel instance query network for named entity recognition. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 947–961. Association for Computational Linguistics, Dublin, Ireland (2022). https://aclanthology.org/2022.acl-long.67
33. Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1554–1564. Association for Computational Linguistics, Melbourne, Australia (2018). https://aclanthology.org/P18-1144
34. Katiyar, A., Cardie, C.: 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), pp. 861–871. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://aclanthology.org/N18-1079
35. Shibuya, T., Hovy, E.: Nested named entity recognition via second-best sequence learning and decoding. Trans. Assoc. Comput. Linguist. **8**, 605–620 (2020)
36. Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., Hajishirzi, H.: A general framework for information extraction using dynamic span graphs. 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), pp. 3036–3046. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://aclanthology.org/N19-1308
37. Straková, J., Straka, M., Hajic, J.: Neural architectures for nested NER through linearization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5326–5331. Association for Computational Linguistics, Florence, Italy (2019). https://aclanthology.org/P19-1527