# 跨度语义增强的命名实体识别方法

耿汝山[1,2]，陈艳平[1,2]，唐瑞雪[1,2]，黄瑞章[1,2]，秦永彬[1,2]，董博[3]
(1.公共大数据国家重点实验室，550025，贵阳；2.贵州大学计算机科学与技术学院，550025，贵阳；
3.西安交通大学计算机科学与技术学院，710049，西安)

**摘要**：针对命名实体识别方法存在字与字之间语义信息丢失、模型召回率不佳等问题，提出了一种跨度语义信息增强的命名实体识别方法。首先，使用ALBERT预训练语言模型提取文本中包含上下文信息的字符向量，并使用GloVe模型生成字符向量；其次，将两种向量进行拼接作为模型输入向量，对输入向量进行枚举拼接形成跨度信息矩阵；然后，使用多维循环神经网络和注意力网络对跨度信息矩阵进行运算，增强跨度之间的语义联系；最后，将跨度信息增强后的矩阵进行跨度分类以识别命名实体。实验表明：与传统的跨度方法相比该方法能够有效增强跨度之间的语义依赖特征，从而提升命名实体识别的召回率；该方法在ACE2005英文数据集上比传统的方法召回率提高了0.42%，并且取得了最高的$F_1$值。

**关键词**：命名实体识别；跨度语义增强；多维循环神经网络；ALBERT预训练语言模型
**中图分类号**：TP391.1 **文献标志码**：A
**DOI**：10.7652/xjtuxb202207013 **文章编号**：0253-987X(2022)07-0118-09

OSID码

# Named Entity Recognition Based on Span Semantic Enhancement

GENG Rushan[1,2], CHEN Yanping[1,2], TANG Ruixue[1,2], HUANG Ruizhang[1,2],
QIN Yongbin[1,2], DONG Bo[3]
(1. State Key Laboratory of Public Big Data, Guizhou University, Guizhou 550025, China;
2. College of Computer Science and Technology, Guizhou University, Guizhou 550025, China;
3. School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China)

**Abstract**: For problems of the named entity recognition such as word-to-word semantic information loss and poor model recall, the named entity recognition based on span semantic enhancement was proposed in this study. First, character vectors containing contextual information in the text were extracted using the ALBERT pre-trained language model, and character vectors were generated using the GloVe model. Second, the two types of vectors were spliced as the model input vectors, and the input vectors were enumeratively spliced to form the span information matrix, and then the multi-dimensional recurrent neural network and attention network were used to operate on the span information matrix and enhance semantic connections between spans. Finally, the enhanced span information matrix was used for span classification to recognize a named entity. The results show that the proposed approach can effectively enhance the semantic dependency between spans compared with the traditional approach, thus improving the recall rate (increased by 0.42% for ACE2005 English dataset), and the highest $F_1$ value was obtained.

收稿日期：2022-01-07。 作者简介：耿汝山(1997—)，男，硕士生；陈艳平(通信作者)，男，副教授，硕士生导师。 基金项目：国家自然科学基金资助项目(62166007)。
网络出版时间：2022-03-25 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220323.1722.004.html

**Keywords**: named entity recognition; span semantic enhancement; multi-dimensional recurrent neural network; ALBERT pre-trained language model

命名实体识别(named entity recognition, NER)是自然语言处理(nature language processing, NLP)中一项重要的基本任务[1]，文本中的实体往往携带重要的信息，因此识别实体对于下游任务如关系提取[2]、问题生成[3]和实体链接[4]等有重要的影响和积极的意义。由于语言在表达上具有迭代、递归的特点，文本中存在着大量的嵌套语义。例如，“秘鲁总统藤森撤换三军司令”，“秘鲁总统藤森”中嵌套着“藤森”这个实体，且它们归类的实体类型都为“人”。嵌套语义识别依然是自然语言处理中的研究难点。

目前，基于深度学习的命名实体识别的方法可以分为基于序列模型、基于跨度模型和超图的方法。基于序列信息的方法是对语句中的每个字符打标签，其主要思想是通过神经网络的方法对每个字符进行向量表示然后通过一定的编码运算分类。依据每个字符分类的类型再合并为实体[5-8]。基于跨度模型的方法是对字符进行组合以形成不同的跨度，跨度信息中包含有每个字符的信息，因此特定的跨度信息代表着特定的实体信息，所以通过对跨度信息的分类来识别不同的实体[9-12]。基于超图的方法是允许一条边连接到多个节点，以代表不同的实体，使用神经网络对这些边和节点进行编码最后从超图标签中恢复实体[13-14]。虽然深度学习的方法可以加入额外的语义特征信息，但是以上这些方法没有考虑字与字之间的交互问题，会出现语义不足的情况。语义不足会降低模型对于实体的识别精确度，以及无法识别出存在嵌套情况的实体。

循环神经网络(recurrent neural networks, RNN)模型在传统的神经网络模型的基础上发展起来，因而具有很强的非线性数据处理能力，在机器翻译、自动问答、句法解析等自然语言处理任务中被广泛的使用。RNN 是一类用于处理序列数据的神经网络模型，序列数据具有前后数据相关联的特点[15]。但是，RNN 在参数过多的时候容易发生梯度爆炸问题，为了改善其局限性，研究人员提出了许多改进的循环神经网络结构，其中长短期记忆网络(long short term memory, LSTM)和门控循环单元(gate recurrent unit, GRU)就是最常用于处理文本循环神经网络变种模型。文献[16]模型使用分层的 RNN 对人体的姿势进行识别，文献[17]模型使用门控 RNN 模型增强了模型对长距离信息的学习能力，文献[18]将残差网络引入 RNN 中处理更长的文本序列。通常，RNN 用来处理一维信息，无法处理超越一维的信息。对于此问题，多维循环神经网络(multi-dimensional recurrent neural networks, MD-RNN)[19]被提出并可以在高维度的信息上进行语义增强。目前，多维循环神经网络已经在图像分类[20]和语音识别[21]等领域都取得了优异的效果。MD-RNN 是 RNN 的一种泛化，通过将单个循环神经网络的连接替换为与数据维度相同的连接来处理多维数据。为了增强 MD-RNN 利用上下文信息的能力，同时避免环节 RNN 在处理长序列数据时梯度爆炸或梯度消失的问题，通常将长短期记忆网络(LSTM)和门控循环神经网络(gated recurrent unit, GRU)用作隐藏单元。

基于上述研究，MD-RNN 为融合跨度语义信息提供了一种有效的手段，在一定程度上解决了语义不足的问题。本文提出了一种增强跨度信息的方法，在利用跨度信息进行实体识别的基础上，增强跨度之间的语义特征，从而有效地增加了跨度的语义信息。首先，使用预训练语言模型提取字符向量信息，对字符向量进行枚举拼接形成矩阵。其次，对矩阵使用 MD-RNN 进行跨度级的语义增强。最后，对矩阵中的跨度进行分类。模型在 ACE2005 英文数据集与中文数据集上进行了实验，分别取得了 86.92 和 81.16 的 $F_1$ 值，与其他主流的方法相比，该方法取得了最优的结果。

## 1　跨度语义增强模型

模型的整体结构如图 1 所示，主要分为编码、表填充、跨度语义增强、输出 4 个模块，其中表填充和跨度意义增强可以反复叠加多层，下面分别介绍各个模块。

### 1.1　编码

假设输入模型的一句话 $\boldsymbol{S}=\langle w_1, w_2, \cdots, w_n\rangle$，$w_k$ 表示句子中的第 $k$ 个字。对 $\boldsymbol{S}$ 通过 ALBERT 的预训练语言模型提取出含有上下文信息的连续稠密向量 $\boldsymbol{a}\in\mathbf{R}^{N\times d_a}$，$N$ 表示句子长度，同时使用 GloVe 预训练字向量将其编码为连续稠密向量 $\boldsymbol{x}\in\mathbf{R}^{N\times d_x}$，公式如下

$$\boldsymbol{a}, \boldsymbol{c} = f(\boldsymbol{S}) \tag{1}$$

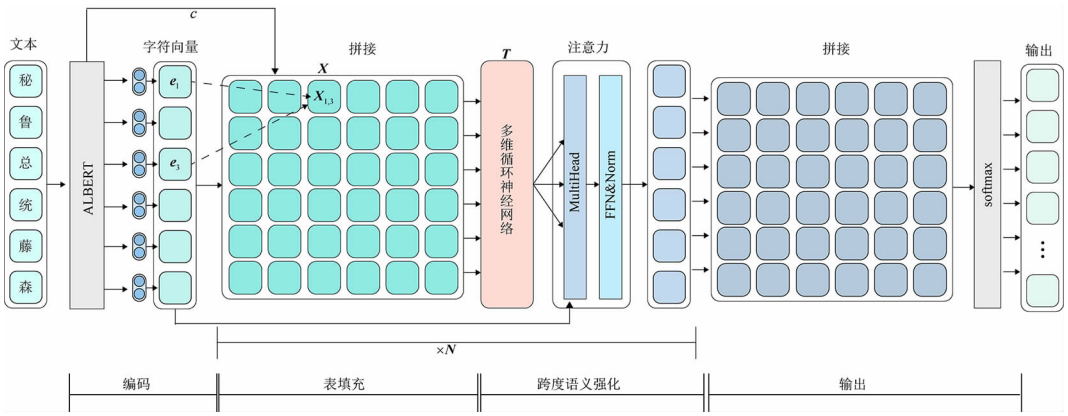



图 1 跨度语义增强模型

Fig. 1 Span semantic enhancement model

$$\boldsymbol{x}=g(\boldsymbol{S}) \tag{2}$$

$$\boldsymbol{E}=([\boldsymbol{a};\boldsymbol{x}])\boldsymbol{W}_1+\boldsymbol{b}_1 \tag{3}$$

式中：$f$ 表示 ALBERT 语言模型的运算；$g$ 表示 GloVe 语言模型的运算；$\boldsymbol{c}\in\mathbf{R}^{N\times N\times d_c}$ 表示 ALBERT 的注意力向量；$\boldsymbol{E}\in\mathbf{R}^{N\times d_e}$ 表示编码层的输出向量；$\boldsymbol{W}_1\in\mathbf{R}^{N\times d_1}$ 和 $\boldsymbol{b}_1$ 表示可学习参数。通过将两种向量进行拼接，并经过线性层来代表编码层输出向量。

### 1.2 表填充

首先将得到的编码层输出 $\boldsymbol{E}=[e_1,e_2,e_3,\cdots,e_n]$ 进行枚举拼接。将一个长度为 $N$ 的句子拼接为 $N\times N$ 的表格 $\boldsymbol{X}\in\mathbf{R}^{N\times N\times d_x}$，其中表格 $i$ 行 $j$ 列的位置对应句子中第 $i$ 个词与第 $j$ 个词进行拼接，此时 $\boldsymbol{X}$ 并不包含上下文的信息，公式如下

$$\boldsymbol{X}_{l,i,j}=\max(0,[\boldsymbol{E}_{l-1,i}\boldsymbol{W}_2;\boldsymbol{E}_{l-1,j}\boldsymbol{W}_3]) \tag{4}$$

式中 $l$ 表示当前运算的层数。表格 $\boldsymbol{X}$ 中每一个格子都代表着一个子序列跨度，这样就将句子中所有可能的子序列都表示出。同时，将 ALBERT 预训练模型中得到的注意力权重 $\boldsymbol{c}$ 与表格拼接，作为补充信息，更新公式如下

$$\boldsymbol{X}_{l,i,j}=\max(0,[\boldsymbol{E}_{l-1,i}\boldsymbol{W}_2;\boldsymbol{E}_{l-1,j}\boldsymbol{W}_3;\boldsymbol{c}_{i,j}]) \tag{5}$$

通过这种方式，可以将预训练模型的权重 $\boldsymbol{c}$ 添加到跨度语义增强中，补充语义信息。

### 1.3 跨度语义增强

通过拼接得到的每个表格只有边界的信息，缺少句子的上下文信息，因此使用 MD-RNN 增强跨度语义信息以弥补缺失的信息。由于传统循环神经网络是在一维的信息上进行计算，因此只接受上一个位置的隐藏层输出。然而，表格具有二维结构的特点，所以二维循环神经网络可以充分利用表格信息。在表格中每个位置的隐藏层不仅接受表格横向的上一个位置的隐藏层输出，也接受表格纵向的上一个位置的隐藏层输出。

将表格 $\boldsymbol{X}$ 输入到使用 GRU 作为隐层的 MD-RNN 中计算，得到表格 $\boldsymbol{T}$。整个计算过程分为门控计算和隐层计算两部分。计算门控的过程如下

$$\boldsymbol{T}'_{l,i,j}=[\boldsymbol{T}_{l-1,i,j};\boldsymbol{T}_{l,i-1,j};\boldsymbol{T}_{l,i,j-1}] \tag{6}$$

$$\boldsymbol{r}_{l,i,j}=\sigma([\boldsymbol{X}_{l,i,j}\boldsymbol{W}_r;\boldsymbol{T}'_{l,i,j}\boldsymbol{U}_r+\boldsymbol{b}_r]) \tag{7}$$

$$\boldsymbol{z}_{l,i,j}=\sigma([\boldsymbol{X}_{l,i,j}\boldsymbol{W}_z;\boldsymbol{T}'_{l,i,j}\boldsymbol{U}_z+\boldsymbol{b}_z]) \tag{8}$$

$$\tilde{\boldsymbol{\lambda}}_{l,i,j,m}=[\boldsymbol{X}_{l,i,j}\boldsymbol{W}_{\lambda_m};\boldsymbol{T}'_{l,i,j}\boldsymbol{U}_{\lambda_m}+\boldsymbol{b}_{\lambda_m}] \tag{9}$$

$$\boldsymbol{\lambda}_{l,i,j,0},\boldsymbol{\lambda}_{l,i,j,1},\boldsymbol{\lambda}_{l,i,j,2}=\mathrm{softmax}(\tilde{\boldsymbol{\lambda}}_{l,i,j,0},\tilde{\boldsymbol{\lambda}}_{l,i,j,1},\tilde{\boldsymbol{\lambda}}_{l,i,j,2}) \tag{10}$$

式中：$\sigma$ 表示 sigmoid 函数；$\boldsymbol{T}'_{l,i,j}\in\mathbf{R}^{3\times d_t}$ 表示用于计算门控的中间向量；$\boldsymbol{r}_{l,i,j}\in\mathbf{R}^{d_r}$ 表示恢复门(reset gate)，决定是否遗忘之前的隐层状态；$\boldsymbol{z}_{l,i,j}\in\mathbf{R}^{d_z}$ 表示更新门(update gate)，选择隐层状态是否更新到新的状态；$\boldsymbol{\lambda}_{l,i,j,m}\in\mathbf{R}^{d_\lambda}$ 表示更新门的权重；$\boldsymbol{W}^r$、$\boldsymbol{W}^z$、$\boldsymbol{W}_{\lambda_m}$、$\boldsymbol{U}^r$、$\boldsymbol{U}^z$、$\boldsymbol{U}_{\lambda_m}$、$\boldsymbol{b}_r$、$\boldsymbol{b}_z$ 和 $\boldsymbol{b}_{\lambda_m}$ 表示可学习参数，用于对矩阵信息进行调节；softmax 函数将实数映射到[0,1]之间。接着，计算隐藏状态

$$\widetilde{\boldsymbol{T}}_{l,i,j}=\tanh(\boldsymbol{X}_{l,i,j}\boldsymbol{W}_x+\boldsymbol{r}_{l,i,j}(\boldsymbol{T}^{\mathrm{prev}}_{l,i,j}\boldsymbol{W}_p)+\boldsymbol{b}_p) \tag{11}$$

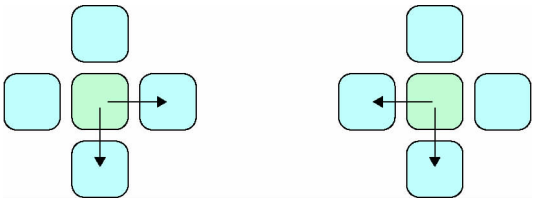
$$\widetilde{\boldsymbol{T}}'_{l,i,j}=\boldsymbol{T}_{l-1,i,j}\boldsymbol{\lambda}_{l,i,j,0}+\boldsymbol{T}_{l,i-1,j}\boldsymbol{\lambda}_{l,i,j,1}+\boldsymbol{T}_{l,i,j-1}\boldsymbol{\lambda}_{l,i,j,2} \tag{12}$$

$$\boldsymbol{T}_{l,i,j}=\widetilde{\boldsymbol{T}}_{l,i,j}\boldsymbol{z}_{l,i,j}+\widetilde{\boldsymbol{T}}'_{l,i,j}(\boldsymbol{1}-\boldsymbol{z}_{l,i,j}) \tag{13}$$

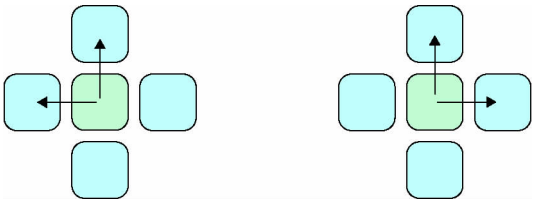
式中：$\widetilde{\boldsymbol{T}}_{l,i,j}\in\boldsymbol{R}^{d_t}$ 和 $\widetilde{\boldsymbol{T}}'_{l,i,j}\in\mathbf{R}^{d_t}$ 表示计算的中间状态；$\boldsymbol{T}\in\mathbf{R}^{d_t}$ 表示 MD-RNN 的输出；$\boldsymbol{W}^x$、$\boldsymbol{W}^p$ 和 $\boldsymbol{b}^h$ 表示可学习的参数。

将序列信息升为表格信息带来了计算量上的指数增加，因此需要考虑对表格和计算进行一定的优化。在表格存储的信息中反对角线是相同的，因此为了加速计算以减少运算量可以只关注矩阵的上三

角信息。一个长度为 $N$ 的序列更新为表格信息后只需要计算 $(N+1)\times N/2$ 个格子，MD-RNN 可以在行与列方向上任意挑选方向组合。如图 2 所示，MD-RNN 在 $N$ 维上有 $2^N$ 种计算方向，在二维的数据上共有 4 种方式。



(a)方向左到右上到下　　(b)方向右到左上到下



(c)方向右到左下到上　　(d)方向左到右下到上

图 2　MD-RNN 计算方向

Fig. 2　MD-RNN calculation direction

根据文本信息的特点，使用所有方向上的信息会导致一些计算的重复利用，并不能很好增强跨度语义信息。为了降低计算量同时尽可能使当前跨度学习到临近跨度的语义信息，可只选择两个方向 $(a)$ $(c)$ 进行计算，并将计算的结果拼接作为跨度语义增强后的输出，计算公式如下

$$\boldsymbol{T}_{l,i,j}^{(a)}=\mathrm{GRU}^{(a)}(\boldsymbol{X}_{l,i,j},\boldsymbol{T}_{l-1,i,j}^{(a)},\boldsymbol{T}_{l,i-1,j}^{(a)},\boldsymbol{T}_{l,i,j-1}^{(a)}) \tag{14}$$

$$\boldsymbol{T}_{l,i,j}^{(c)}=\mathrm{GRU}^{(c)}(\boldsymbol{X}_{l,i,j},\boldsymbol{T}_{l-1,i,j}^{(c)},\boldsymbol{T}_{l,i+1,j}^{(c)},\boldsymbol{T}_{l,i,j+1}^{(c)}) \tag{15}$$

$$\boldsymbol{T}_{l,i,j}=[\boldsymbol{T}_{l,i,j}^{(a)};\boldsymbol{T}_{l,i,j}^{(c)}] \tag{16}$$

如图 3 所示，绿色方块代表当前的跨度，蓝色的方块代表了临近的跨度，$(a)$ 方向上的当前跨度会增强后一个跨度的信息，$(c)$ 方向上的会增强到之前跨度的信息，最后将两个方向上的结果进行拼接，得到包含跨度上下文信息的表格 $\boldsymbol{T}_l$。

注意力的使用是为了梳理跨度信息，跨度信息在经过语义增强后会出现语义信息混乱的情况。注意力运算将根据序列信息调整跨度信息，输入是一条句子的向量，第 $i$ 个向量代表着句子中的第 $i$ 个字，以及 MD-RNN 的输出表格 $\boldsymbol{T}_l$，其整体结构是基于 Transformer，将自注意力机制(self-attention)中的缩放点乘替换成跨度引导的注意力机制[22]。Transformer 在自然语言处理领域取得了优异的成

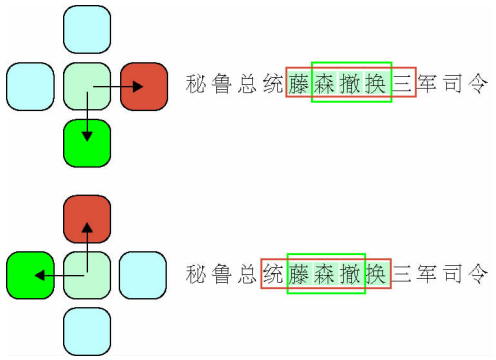



图 3　相邻跨度的语义增强

Fig. 3　Semantic reinforcement of adjacent spans

绩，其采用的自注意力机制是成功的关键因素[23]。传统的自注意力机制有查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)和值($\boldsymbol{V}$)3 个输入，计算过程如下

$$z(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}}{\sqrt{d}}\right)\boldsymbol{V} \tag{17}$$

式(17)表示 3 种输入之间的运算关系。

本模型的数据输入来自前一个序列表示层 $\boldsymbol{E}_{l-1}$。在 self-attention 中考虑 $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{E}$，可以发现函数 softmax 的输出是一个基本的注意力权重，而 MD-RNN 输出的 $\boldsymbol{T}_l$ 来自于 $\boldsymbol{E}_{l-1}$。因此，可以用 $\boldsymbol{T}_l$ 来代替计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似度，这样就可以把 $\boldsymbol{T}_l$ 带入到 self-attention 公式。这个过程称为跨度引导的注意力机制，公式如下

$$z(\boldsymbol{T},\boldsymbol{V})=\boldsymbol{V}\boldsymbol{T}_l\boldsymbol{W}_{\mathrm{att}} \tag{18}$$

式中：$\boldsymbol{W}_{\mathrm{att}}$ 表示可学习参数，用于调整 $\boldsymbol{T}_l$ 的维度；$\boldsymbol{V}$ 来自于字符序列 $\boldsymbol{E}_{l-1}$。与传统的注意力的计算方法相比，使用跨度信息来引导计算会有一些相对应的优势：①不需要计算相似度，而使用 $\boldsymbol{T}_l$ 来减少计算量；②$\boldsymbol{T}_l$ 的计算过程有关于行、列和前几层的信息，这可以更好地获得上下文信息；③使用 $\boldsymbol{T}_l$ 允许表格编码器参与序列表示学习过程。

本模型的注意力的计算流程如图 4 所示。将注意力机制扩展到多个头，并且都有独立的参数。拼接输出，使用全连接层来得到最终的注意力向量，其余部分与 Transformer 类似。

在注意力机制后使用前馈神经网络(FFNN)对关注权重进行串联和归一化，公式如下

$$\tilde{\boldsymbol{E}}_l=\varphi(\boldsymbol{E}_{l-1},z(\boldsymbol{E}_{l-1},\boldsymbol{T}_{l-1})) \tag{19}$$

$$\boldsymbol{E}_l=\varphi(\tilde{\boldsymbol{E}}_l,z(\tilde{\boldsymbol{E}}_l,\boldsymbol{T}_l)) \tag{20}$$

式中 $\varphi$ 表示层规范(LayerNorm)运算。

### 1.4　预测与损失函数

在最后一层跨度意义增强后，对输出的序列信息进行枚举拼接，并使用拼接的表格信息 $\boldsymbol{T}_l$ 去预测

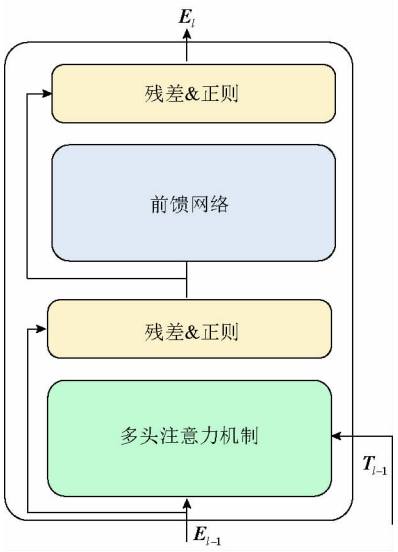



图 4 跨度引导注意力模块

Fig. 4 Span guide attention module

嵌套的实体标签，公式如下

$$P_\theta(Y)=\mathrm{softmax}(\boldsymbol{T}_l\boldsymbol{W}_2+b_2) \tag{21}$$

式中：$Y$ 表示预测标签的随机变量；$P_\theta$ 表示模型参数的估计概率函数。使用交叉熵函数被用作损失函数，给定输入 $x$ 和标签 $y$ 计算损失函数，公式如下

$$L=-\sum_{i,j\in[i,N]}\log P_\theta(Y_{i,j}=y_{i,j}) \tag{22}$$

在训练的过程中，只考虑矩阵中上三角的内容，选择最可能的标签作为答案进行输出，公式如下

$$\arg\max\sum_{i\in[0,N],j\in[i,N]}P_\theta(Y_{i,j}^{\mathrm{nest}}=x)+P_\theta(Y_{j,i}=x) \tag{23}$$

## 2 实验与结果分析

为了验证本文提出的基于跨度语义增强的嵌套命名实体识别模型的效果，所以选择与其它基于深度学习的命名实体识别模型进行对比。本文在 ACE2005 中文数据集和 ACE2005 英文数据集上进行实验。使用精确率 $P$、召回率 $R$ 以及 $F_1$ 值作为实验的主要评估指标，通过在相同数据集上进行实验，验证本模型的各项性能。

### 2.1 实验设置

#### 2.1.1 数据集介绍

ACE2005 中文数据集和 ACE2005 英文数据集是语言数据联盟（Linguistic Data Consortium，LDC）于 2006 年发布用于各项自然语言处理任务的公共数据集。ACE2005 英文数据集包含 7 种实体类别，共 599 篇英语文档，分别为 PER（人物，person）、ORG（组织机构，organization）、GPE（地理/社会/政治实体，geo-political）、LOC（处所，locations）、FAC（设施，facilities）、VEH（交通工具，vehicle）、WEA（武器，weapon）。ACE2005 中文数据集包括同样的 7 种实体类别，共 633 篇中文文档。

两个数据集都按照 8∶1∶1划分位训练集、验证集和测试集，其中 ACE2005 英文数据集采用文献[5]的数据集划分方法。

#### 2.1.2 评价指标

实体识别需要同时识别出实体准确的位置和类别，只有当实体位置和类别都识别正确时结果才算正确。本文实验使用综合指标 $F_1$ 值来衡量模型性能，公式如下

$$P=\frac{M}{N}\times 100\% \tag{24}$$

$$R=\frac{M}{K}\times 100\% \tag{25}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{26}$$

式中：$M$ 表示正确识别出的命名实体个数；$N$ 表示识别出的命名实体个数；$K$ 表示标准结果中的命名实体个数。

#### 2.1.3 实验环境与参数设置

本文提出的跨度语义增强模型在 Python3.7 和 Pytorch1.8 的环境下进行实验。设置模型在训练、验证与测试时的 Batch_size 分别为 16、8、8。为降低神经网络在训练过程中过拟合造成的影响，设置 Dropout 为 0.5。初始化的字嵌入向量使用了 GloVe 模型进行训练并将其保存，其维度为 100。在模型训练中不更新单词的 GloVe 向量，将隐藏层的大小设置为 200，并且将每个 MD-RNN 的隐藏层大小设置为 100。学习率设置为 0.001，Adam 作为模型的优化工具，上下文的信息使用 ALBERT 进行训练并保存最后一层的输出作为文字的上下文信息。ALBERT 是一个共享 Transformer 参数的轻量型 BERT 模型[24]，在模型中使用的版本是 albert-xxlargev2，其隐藏层的维度为 4 096。

### 2.2 结果分析

本文选取了几个在命名实体识别上取得了较高性能的模型进行对比，表 1 展示了与这些方法在 ACE2005 英文数据集上的性能区别，通过使用 ALBERT 作为预训练语言模型来提供上下文的信息。本文模型的 $F_1$ 值达到了最优的 86.92，$P$、$R$ 值也达到了最好的性能，表明了本文方法的优越性。

从整体上看，文献[8]的模型使用次佳最优路径的方法识别嵌套命名实体在精确率 $P$ 和召回率 $R$ 的

表 1　ACE2005 英文数据集上的实验结果

Table 1　The results of the contrast experiments on the ACE2005 English dataset

| 模型来源 | $P$/% | $R$/% | $F_1$ 值 |
|---|---|---|---|
| 文献[8] | 83.30 | 84.69 | 83.99 |
| 文献[25] | 85.20 | 85.60 | 85.40 |
| 文献[7] | 85.30 | 87.40 | 86.34 |
| 文献[26] | 86.09 | 87.27 | 86.67 |
| 本文 | 86.17 | 87.69 | 86.92 |

结果上明显低于其他方法。这与其使用的选择次佳路径的方式有关，选择次佳路径的算法会对结果有着明显的影响和限制。文献[26]的回归模型性能与其他传统方法比取得了最好的 $F_1$ 值，其模型在设计的过程中考虑到了跨度信息利用不充分的问题，以及长实体识别较困难的现象，用回归来生成跨度建议来定位实体位置，然后利用相应的类别标记边界来对跨度建议再次调整。与其相比，本文的模型在精确率、召回率和 $F_1$ 值上都高于回归模型，这说明本文模型识别出的实体绝大多数是真实有效的。这是由于结合了跨度级上下文信息为模型提供了更加准确的实体表示，使得分类器有能力确定矩阵中的每一个单位是否为有效的实体跨度。

表 2 展示模型在 ACE2005 中文数据集上的性能。“Shallow BA”模型采用两个独立的 CRF 模型来识别起点与终点边界，候选实体建立起来后，再对候选实体进行分类。“Cascading-out”模型每种实体被单独的识别，如果出现嵌套的相同类型的实体，则会识别最外层的实体。“Layering-out”模型是由两个独立的 BERT-BiLSTM-CRF 模型去识别外层的实体与内层的实体，“NNBA”模型通过两个 BiLSTM-CRF 模型分别识别实体的开始与结束位置，然后通过组合的方式生成候选实体再分类。本文模型的精确率较高，大幅度超过了之前的模型，这可能与堆叠多层的跨度语义增强模块有关。

表 2　ACE2005 中文数据集对比实验结果

Table 2　The results of the contrast experiments on the ACE2005 Chinese dataset

| 模型 | $P$/% | $R$/% | $F_1$ 值 |
|---|---|---|---|
| Shallow BA[27] | 73.98 | 62.16 | 67.56 |
| Cascading-out | 76.96 | 71.39 | 74.07 |
| Layering-out | 78.85 | 81.34 | 80.07 |
| NNBA[28] | 79.31 | 80.94 | 80.12 |
| 本文 | 83.00 | 79.39 | 81.16 |

本文模型在两种语言的数据集上进行实验是为了验证不同语言上的适用性。整体上，本文的模型都取得了非常高的结果。

### 2.3　句子长度(字数)对模型性能影响分析

长句子会导致模型对实体信息不敏感[27]，提高模型对长句子中实体识别的能力对模型性能有积极的作用，因此探究去掉跨度语义增强模块对不同长度句子的性能影响。本实验在 ACE2005 英文数据集上分析不同长度句子对于跨度语义增强模块的影响，模型参数保持不变，其结果如图 5 所示。

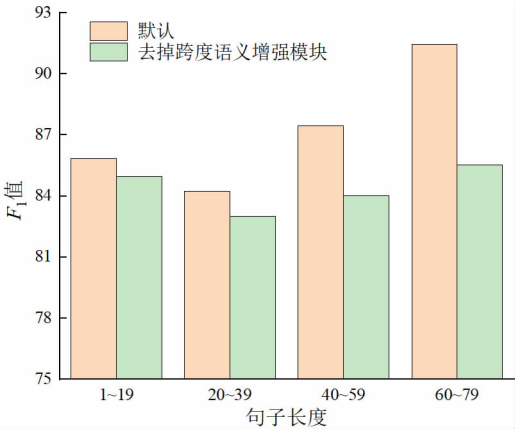



图 5　不同句子长度模型性能

Fig. 5　Different results of different sentence length

通过对两种方法进行对比，完整的模型比去掉跨度语义增强模块的模型在长句子的区间(40 到 59 与 60 到 79)上有了明显的性能提升。在较短句子的区间上，也有一定的提升。

### 2.4　注意力权重可视化

为了更好地探索注意力机制在模型中的作用，将注意力权重进行了可视化，如图 6 所示，使用 bertvit 工具进行可视化处理。为了更好地突出与 self-attention 的不同，同时也对模型进行了修改并做了 self-attention 的可视化。重点比较两种算法下的单词与单词之间的关系，模型的其他参数与之前保持一致，在此使用一个例子“We're paying attention to all those details for us, Chris Plante at the pentagon.”来进行说明。这句话中有 4 个实体，即“we”“Chris Plante at the pentagon”“the pentagon”和“us”。

在图 6 中深色的线代表相对较高的注意权重，右侧的层数代表着堆叠的跨度增强模块的层数，可以看出，不同层之间注意力关注到的内容是不同的，“Chris”对“we”“us”和“Plante”的注意权重较高，它们的实体类型是相同的，同一类型的实体彼此之间

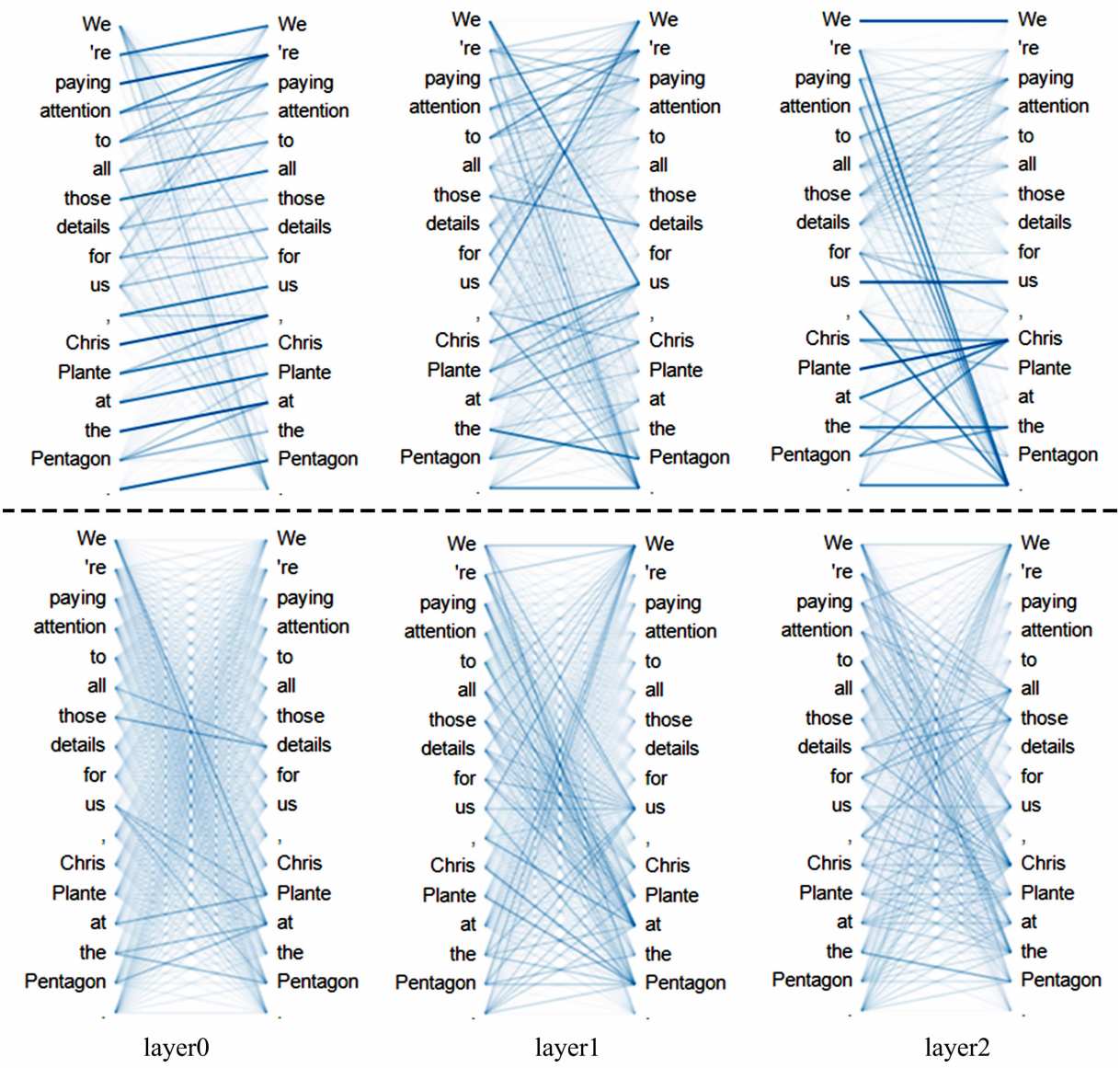



图 6 注意力可视化

Fig. 6 Attention visualization

有相对较高的相似度。因此，同一类型的实体之间的注意权重是相互关联的，与 self-attention 算法相比，self-attention 的关注重点不够突出，即图中的深色线不够明显。由此可以证明，本模型的注意力算法优于原始的算法，并且可以关注到同种类型实体之间的关系。

### 2.5 消融实验

为了验证注意力机制和 MD-RNN 的有效性，本文设计了消融实验，结果如表 3 所示，w/o 表示去掉某模块。注意力机制在模型中用于梳理跨度信息的权重，在去掉注意力机制后表格内的序列信息会不可避免的产生一定的混乱，这对于实体的准确性有较大的影响。在去掉跨度引导的注意力模块后，模型性能有了明显下降。

表 3 ACE2005 英文数据集上消融实验结果

Table 3 Results of ablation experiment on ACE2005 English dataset

| 模型设置 | $P$/% | $R$/% | $F_1$ 值 |
|---|---|---|---|
| 默认 | 86.17 | 87.69 | 86.92 |
| w/o 注意力机制 | 84.31 | 85.62 | 84.96 |
| w/o MD-RNN | 84.33 | 86.14 | 85.22 |

MD-RNN 在模型中用于进行跨度语义增强，并且融合跨度级语义特征信息，去掉 MD-RNN 后性能有所下降，这证明了跨度信息增强对于模型性能提升有帮助。跨度语义信息的增强可以补充因利用

边界信息导致丢失的部分语义信息，并且跨度信息对于嵌套实体识别有着较大的帮助。

### 2.6 层数与性能

如表 4 所示，探讨了表填充和跨度引导注意力机制层数与模型性能之间的关系。为了保持与前面实验的一致性，选择 ACE2005 英文数据集进行实验，除了编码器的层数之外其他的超参数均保持不变。一般来说，增加编码器数量可以提高模型的性能，但当编码器的数量超过一个范围时，性能将开始下降。在实验中选择编码器层数为 1、2、3、4、5。通过对实验结果进行分析发现，在层数大于 3 之后，模型的性能就已经开始下降，这应该是由过拟合而导致的性能下降。

表 4　多层编码器对模型的影响

Table 4 The influence of multi-layer encoder on model

| 层数 | $P$/% | $R$/% | $F_1$ 值 |
|---|---|---|---|
| 1 | 84.48 | 86.37 | 85.41 |
| 2 | 85.36 | 86.37 | 85.86 |
| 3 | 86.17 | 87.69 | 86.92 |
| 4 | 85.77 | 87.28 | 86.52 |
| 5 | 85.22 | 86.43 | 85.82 |

## 3 结　语

本文提出一种跨度语义增强的命名实体识别方法，解决了命名实体识别中出现的语义不足、召回率不高等问题，并将传统的一维文本信息拼接成二维的表格信息，使用 MD-RNN 和注意力模型作为跨度语义增强模块，进而有效增强了跨度之间的联系。实验结果表明，本文提出的方法相比之前基于跨度模型的嵌套命名实体识别方法有着明显优势。

在下一步的工作中，将探究如何将跨度语义增强方法应用到其他领域，以及探究一种新的跨度语义融合机制去融合跨度间的语义信息。

**参考文献：**


[1] 黄铭，刘捷，戴齐. 融合字词特征的中文嵌套命名实体识别 [J]. 现代计算机，2021，27(34)：21-28.
HUANG Ming，LIU Jie，DAI Qi. Chinese Nested named entity recognition integrating improved representation learning method [J]. Modern Computer，2021，27(34)：21-28.

[2] XIONG Chenyan，LIU Zhengzhong，CALLAN J，et al. Towards better text understanding and retrieval through kernel entity salience modeling [C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York，NY，USA：ACM，2018：575-584.

[3] ZHOU Qingyu，YANG Nan，WEI Furu，et al. Neural question generation from text：a preliminary study [C]// Natural Language Processing and Chinese Computing. Berlin：Springer，2018：662-671.

[4] GUPTA N，SINGH S，ROTH D. Entity linking via joint encoding of types，descriptions，and context [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Seattle，Washington，USA：ACL，2017：2681-2690.

[5] JU Meizhi，MIWA M，ANANIADOU S. A neural layered model for nested named entity recognition [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Seattle，Washington，USA：ACL，2018：1446-1459.

[6] FISHER J，VLACHOS A. Merge and label：a novel neural network architecture for nested NER [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Seattle，Washington，USA：ACL，2019：5840-5850.

[7] WANG Jue，SHOU Lidan，CHEN Ke，et al. Pyramid：a layered model for nested named entity recognition [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle，Washington，USA：ACL，2020：5918-5928.

[8] SHIBUYA T，HOVY E. Nested named entity recognition via second-best sequence learning and decoding [J]. Transactions of the Association for Computational Linguistics，2020，8：605-620.

[9] XIA Congying，ZHANG Chenwei，YANG Tao，et al. Multi-grained named entity recognition [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Seattle，Washington，USA：ACL，2019：1430-1440.

[10] XU Mingbin，JIANG Hui，WATCHARAWITTAYAKUL S. A local detection approach for named entity recognition and mention detection [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Seattle，Washington，USA：ACL，2017：1237-1247.

[11] SOHRAB M G，MIWA M. Deep exhaustive model for nested named entity recognition [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Seattle，Washington，USA：ACL，2018：2843-2849.

[12] LUAN Yi, WADDEN D, HE Luheng, et al. A general framework for information extraction using dynamic span graphs [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 1 Human Language Technologies. Seattle, Washington, USA: ACL, 2019: 3036-3046.

[13] LU Wei, ROTH D. Joint mention extraction and classification with mention hypergraphs [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: ACL, 2015: 857-867.

[14] MUIS A O, LU Wei. Labeling gaps between words: recognizing overlapping mentions with mention separators [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: ACL, 2017: 2608-2618.

[15] 张西宁，郭清林，刘书语. 深度学习技术及其故障诊断应用分析与展望 [J]. 西安交通大学学报，2020，54(12)：1-13.
ZHANG Xining, GUO Qinglin, LIU Shuyu. Analysis and prospect of deep learning technology and its fault diagnosis application [J]. Journal of Xi'an Jiaotong University, 2020, 54(12): 1-13.

[16] DU Yong, WANG Wei, WANG Liang. Hierarchical recurrent neural network for skeleton based action recognition [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 1110-1118.

[17] TANG Duyu, QIN Bing, LIU Ting. Document modeling with gated recurrent neural network for sentiment classification [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: ACL, 2015: 1422-1432.

[18] WANG Yiren, TIAN Fei. Recurrent residual learning for sequence classification [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: ACL, 2016: 938-943.

[19] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Multi-dimensional recurrent neural networks [C]// International Conference on Artificial Neural Networks. Heidelberg: Springer, 2007: 549-558.

[20] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks [C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2013: 6645-6649.

[21] GRAVES A, SCHMIDHUBER J. Offline handwriting recognition with multidimensional recurrent neural networks [C]// Proceedings of the 21st International Conference on Neural Information Processing Systems. New York, NY, USA: ACM, 2008: 545-552.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, NY, USA: ACM, 2017: 6000-6010.

[23] 方军，管业鹏. 基于双编码器的会话型推荐模型 [J]. 西安交通大学学报，2021，55(8)：166-174.
FANG Jun, GUAN Yepeng. A session recommendation model based on dual encoders [J]. Journal of Xi'an Jiaotong University, 2021, 55(8): 166-174.

[24] 李军怀，陈苗苗，王怀军，等. 基于 ALBERT-BGRU-CRF 的中文命名实体识别方法 [J/OL]. 计算机工程 [2021-10-01]. https: // doi. org/10. 19678/j. issn. 1000-3428. 00616 30.
LI Junhuai, CHEN Miaomiao, WANG Huaijun, et al. Chinese named entity recognition method based on ALBERT-BGRU-CRF [J/OL]. Computer Engineering [2021-10-01]. https: // doi. org/10. 19678/j. issn. 1000-3428. 00616 30.

[25] YU Juntao, BOHNET B, POESIO M. Named entity recognition as dependency parsing [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 6470-6476.

[26] SHEN Yongliang, MA Xinyin, TAN Zeqi, et al. Locate and label: a two-stage identifier for nested named entity recognition [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Volume 1. Seattle, Washington, USA: ACL, 2021: 2782-2794.

[27] CHEN Yanping, ZHENG Qinghua, CHEN Ping. A boundary assembling method for Chinese entity-mention recognition [J]. IEEE Intelligent Systems, 2015, 30(6): 50-58.

[28] CHEN Yanping, WU Yuefei, QIN Yongbin, et al. Recognizing nested named entity based on the neural network boundary assembling model [J]. IEEE Intelligent Systems, 2020, 35(1): 74-81.

（编辑 荆树蓉 刘杨）