计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2024/18(02)-0506-10
doi: 10.3778/j.issn.1673-9418.2210078

# 多尺寸注意力的命名实体识别方法

唐瑞雪[1,2,3+],秦永彬[2,3],陈艳平[2,3]
1. 贵州财经大学 信息学院,贵阳 550025
2. 贵州大学 计算机科学与技术学院,贵阳 550025
3. 公共大数据国家重点实验室,贵阳 550025
+ 通信作者 E-mail: trx_0401@163.com

**摘　要:**命名实体识别(NER)任务的准确性将促进自然语言领域中诸多下游任务的研究。由于文本中存在大量嵌套语义,导致命名实体识别困难,成为自然语言处理中的难点。以往研究提取特征尺度单一,边界信息利用不够充分,忽略了不同尺度下的许多细节信息,从而造成实体识别错误或遗漏的情况。针对上述问题,提出一种多尺度注意力的命名实体识别方法(MSA-NER)。首先,利用BERT模型得到包含上下文信息的表示向量,并通过BiLSTM网络加强文本的上下文表示。其次,将表示向量进行枚举拼接形成跨度信息矩阵,并融合方向信息获得更丰富的交互信息。然后,利用多头注意力构建多个子空间,通过二维卷积在每个子空间下可选地聚合不同尺度的文本信息,在每个注意力层同时进行多尺度的特征融合。最后,将融合的矩阵进行跨度分类以识别命名实体。实验表明,该方法在GENIA和ACE2005英文数据集上 $F1$ 分别达到81.7%和86.8%,与现有主流模型相比有更好的识别效果。

**关键词:**命名实体识别(NER);嵌套语义;多尺度注意力;卷积神经网络;子空间
**文献标志码:**A　　**中图分类号:**TP18

## Named Entity Recognition Based on Multi-scale Attention

TANG Ruixue[1,2,3+], QIN Yongbin[2,3], CHEN Yanping[2,3]
1. School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China
2. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
3. State Key Laboratory of Public Big Data, Guiyang 550025, China

**Abstract:** The accuracy of named entity recognition (NER) task will promote the research of multiple downstream tasks in natural language field. Due to a large number of nested semantics in text, named entities are recognized difficultly. Recognizing nested semantics becomes a difficulty in natural language processing. Previous studies have single scale of extracting feature and under-utilization of the boundary information. They ignore many details under different scales and then lead to the situation of entity recognition error or omission. Aiming at the above problems, a multi-scale attention method for named entity recognition (MSA-NER) is proposed. Firstly, the BERT model is used to obtain representation vector containing context information, and then the BiLSTM network is used to strengthen the context representation of text. Secondly, the representation vectors are enumerated and concatenated to form span information matrix. The direction information is fused to obtain richer interactive information. Thirdly, multi-head attention is used to construct multiple subspaces. Two-dimensional convolution is used to optionally ag-

**基金项目:**贵州省省级科技计划项目(黔科合基础ZK[2022]一般027)。
This work was supported by the Science and Technology Projects of Guizhou Province (ZK[2022]027).
**收稿日期:**2022-10-19　　**修回日期:**2023-01-04

gregate text information at different scales in each subspace, so as to implement multi-scale feature fusion in each attention layer. Finally, the fused matrix is used for span classification to identify named entities. Experimental results show that the $F1$ score of the proposed method reaches 81.7% and 86.8% on GENIA and ACE2005 English datasets, respectively. The proposed method demonstrates better recognition performance compared with existing mainstream models.

**Key words:** named entity recognition (NER); nested semantics; multi-scale attention; convolutional neural network; subspace

命名实体识别(named entity recognition,NER)[1]是指从给定文本中识别特定类别的实体信息,例如人名、机构名和地点等。命名实体识别作为自然语言处理(natural language processing,NLP)的核心任务之一,其结果可以促进问答系统[2]、机器翻译[3]、情感分析[4]、知识库构建[5]等下游任务的研究。

命名实体识别任务通常被看作序列标注任务[6-7],利用条件随机场(conditional random field,CRF)[8]、长短期记忆网络(long short-term memory,LSTM)[9]等输出一条最大概率的标注序列。由于每一个单词只能分配一个标签,只能解决非嵌套结构。但是实际上,文本中存在大量的具有嵌套结构的命名实体。如在GENIA数据集中,“in PWM treated B cells”中“PWM”表示一种protein实体,嵌套于另一种cell_line实体“PWM treated B cells”中。因此嵌套语义识别仍然是自然语言处理中的难点。

为了识别嵌套语义,目前研究可分为基于分层的方法、基于超图的方法及基于跨度的方法。基于分层的方法是将嵌套结构转化为平面结构来预测实体类型。虽然解决了嵌套结构,但是往往需要设计复杂解码方案,且耗时较大。基于超图的方法是通过使用神经网络来编码节点和边,显式地从超图中识别嵌套实体。该方法没有考虑到句级特征的提取。基于跨度的方法将列举句子所有可能的子序列作为候选实体来进行预测。该方法过度依赖于实体跨度的语义表示,没有充分利用实体边界信息。由于真实实体与大量负样本具有相似语义表示,极易产生错误。因此,上述方法忽略了实体的边界信息,无法从文本中捕获更多层次粒度的语义信息,导致嵌套语义识别困难,从而降低了实体识别的性能。

针对上述问题,本文提出了一种多尺度注意力的命名实体识别方法(multi-scale attention method for NER,MSA-NER)。为了丰富文本的交互信息,将通过BERT(bidirectional encoder representation from transformers)预训练模型和BiLSTM(bidirectional long short-term memory)网络的表示构建成矩阵,将矩阵中单词表示作为条件信息,体现信息之间的方向性。为了提高模型的表达能力,通过多头注意力有效地聚合不同尺度的文本信息,在提取粗粒度特征信息的同时关注细粒度特征信息。该方法在GENIA和ACE2005英文数据集上进行了实验,分别取得了81.7%和86.8%的$F1$值,与其他主流方法相比,该方法取得了最优结果。

## 1 相关工作

随着深度学习的不断发展,基于深度学习的命名实体识别方法也展现出不错的效果。BiLSTM-CRF[10]作为目前主流的命名实体识别模型,可有效地解决非嵌套语义,并取得了很好的效果。近年来,提出了各种针对嵌套命名实体识别的方法,可分为基于分层的方法、基于超图的方法及基于跨度的方法。

基于分层的方法[11-13]是将嵌套结构转化为平面结构来预测实体类型。Ju等[11]提出了一种动态分层模型。首先利用LSTM和CRF组成平面NER层,再将平面NER层进行堆叠,直至当前平面NER层预测的序列不满足检测规则。Wang等[13]提出了一种新的分层模型,将文本递归地堆叠成多个平面NER层,形成金字塔形状,每一层预测特定长度的实体。同时,设计了逆向金字塔允许层之间的双向交互,减少错误传播。虽然上述模型解决了嵌套结构,但是往往需要设计复杂解码方案,同时存在层与层之间的错误传播。

基于超图的方法[14-17]是通过使用神经网络来编码节点和边,显式地从超图中识别嵌套实体。Lu和Roth[14]在2015年首次使用超图的思想,提出了一种基于超图表示的模型。该模型首先使用节点和有向边共同对命名实体及其组合进行表示,从而将一个句子中不同类型且无限长度的嵌套命名实体表示出

来。2017年,Muis等[15]将提及分隔符与特征结合,开发了一个基于间隙的标记模型来识别嵌套实体结构。上述模型没有考虑到句级特征的提取。Wang和Lu[16]将深度神经网络用于超图模型,提出了一种新的分段超图表示模型,捕获更多的特征信息。上述模型为了避免虚假结构和结构歧义的问题,在设计超图时需要大量人工。此外,该方法也无法对嵌套命名实体之间的依赖进行编码。

基于跨度的方法[18-23]将列举句子所有可能的子序列作为候选实体来进行预测。Xia等[20]提出了一种多粒度命名实体识别框架,其中包括检测器和分类器两个模块。检测器检测所有可能的命名实体跨度,而分类器的目的是将检测到的命名实体分类为预定义的命名实体类别。Yu等[21]利用了依赖树的思想去建模,通过双仿射去计算实体跨度的开始和结束之间的分数。由于枚举所有实体跨度,模型计算复杂性高。若句子过长,还存在样本不均衡的问题。Li等[22]提出了一种基于机器阅读理解(machine reading comprehension,MRC)的方法。对于需要识别的每一类实体构造关于该类实体的问题,预测该类实体在文本中的位置。Yan等[23]采用了序列到序列(sequence-to-sequence,Seq2Seq)框架,并利用指针网络直接生成实体。但是上述方法过度依赖实体跨度表示,边界信息没有充分利用。当真实实体与负样本之间语义表示高度重叠时,易造成识别错误。

综上可知,现有命名实体方法存在提取特征尺度单一、边界信息利用不够充分的问题。本文对文本表示向量进行交互学习,通过多头注意力产生多个子空间,每个子空间通过不同卷积核的卷积神经网络来聚合不同长度的文本信息,将多尺度特征的提取统一到同一注意力层,然后利用输出的结果进行实体类别的预测。实验结果表明,本文模型得到粗细粒度的语义信息,提高实体识别的准确性。

## 2 多尺度注意力的命名实体识别模型

模型的整体结构如图1所示,主要分为编码、交互表示、多尺度注意力以及类别判断四个部分。第一部分为编码层,通过BERT获得句子的表示向量,经过BiLSTM让两个隐藏信息表示融合得到最终的上下文表示向量。第二部分为交互表示层,对上述表示向量进行枚举拼接构建成表格,并融合方向信息加强单词之间的交互。第三部分是多尺度注意力,利用多头注意力构建多个子空间,通过不同大小的二维卷积在不同子空间下进行多尺度特征的提取,并在同一注意力层进行融合。第四部分为类别判断层,得到候选实体的实体类别或非实体类别。下面分别对模型各个模块进行详细分析。

### 2.1 编码

预训练模型BERT[24]采用双向Transformer结构来抽取特征,不仅可以充分利用文本的上下文信息,而且生成的文本向量更能表达独一性,包含更多的语义信息。给定句子 $T=\{t_1,t_2,\cdots,t_n\}$ ,$n$ 为句子长度。本文采用BERT模型来获取句子的表示向量 $\boldsymbol{X}$ ,如式(1)和式(2)所示:

$$\boldsymbol{x}_i=f_e(t_i) \tag{1}$$

$$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\} \tag{2}$$

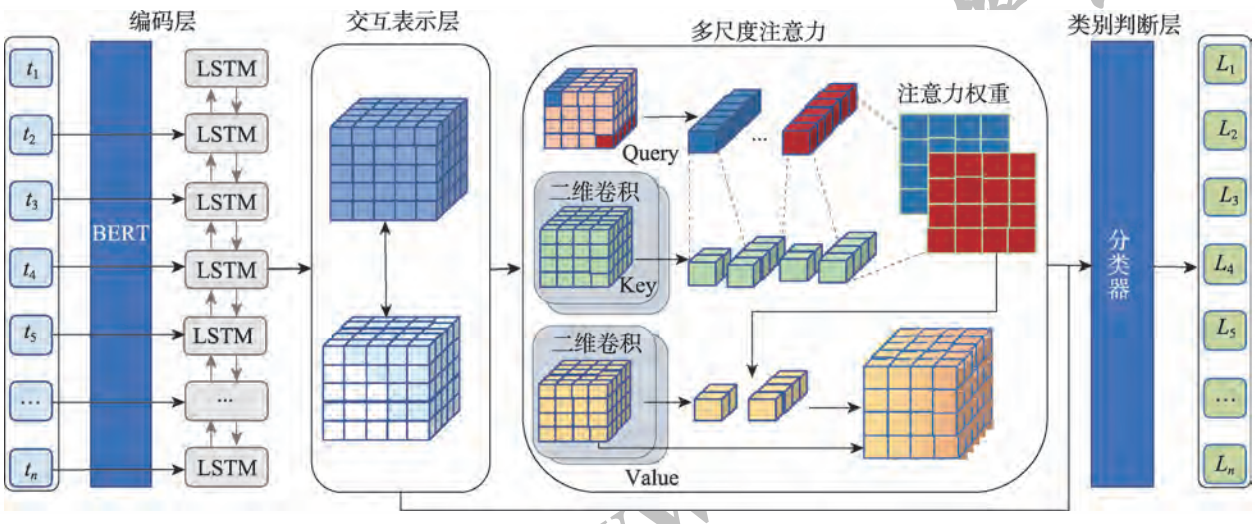


图1 MSA-NER模型架构

Fig.1 Model structure of MSA-NER

其中，$f_e$ 为 BERT 语言模型的运算；$\boldsymbol{X}\in\mathbb{R}^{n\times d}$ 表示句子的表示向量，$\boldsymbol{x}_i\in\mathbb{R}^{1\times d}$ 表示第 $i$ 个单词的表示，$d$ 为单词的维度。

相较于传统的循环神经网络，LSTM 能够有效解决因句子过长带来梯度消失和梯度爆炸问题。而双向长短期记忆网络（BiLSTM）可以同时捕获文本前向和后向的语义信息。为了更好地捕获双向的语义依赖，将得到的表示向量 $\boldsymbol{X}$ 输入 BiLSTM 网络中，通过前向 LSTM $f_L$ 和后向 LSTM $f_R$ 得到句子隐藏状态，融合得到最终的上下文表示向量，如式（3）~式（6）所示：

$$\boldsymbol{h}_{L,i}=f_L(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n) \tag{3}$$

$$\boldsymbol{h}_{R,i}=f_R(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n) \tag{4}$$

$$\boldsymbol{h}_i=[\boldsymbol{h}_{L,i};\boldsymbol{h}_{R,i}] \tag{5}$$

$$\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\} \tag{6}$$

其中，$\boldsymbol{h}_i\in\mathbb{R}^{1\times 2d}$ 为第 $i$ 个单词的双向表示；[ ; ]表示拼接操作；$\boldsymbol{H}\in\mathbb{R}^{n\times 2d}$ 为 LSTM 前向和后向拼接而得，表示编码层的输出向量。

## 2.2 交互表示层

将得到的编码层输出 $\boldsymbol{H}$ 进行枚举拼接。首先将第 $i$ 个单词和第 $j$ 个单词经过线性层，然后将长度为 $n$ 的句子拼接为 $n\times n$ 的表格 $\boldsymbol{M}\in\mathbb{R}^{n\times n\times d_i}$，其中表格中第 $i$ 行第 $j$ 列的位置 $\boldsymbol{m}_{ij}$ 对应句子中单词对 $(\boldsymbol{h}_i,\boldsymbol{h}_j)$ 进行拼接。具体如式（7）所示：

$$\boldsymbol{m}_{ij}=[\boldsymbol{h}_i\boldsymbol{W}_1;\boldsymbol{h}_j\boldsymbol{W}_2] \tag{7}$$

其中，$\boldsymbol{m}_{ij}\in\mathbb{R}^{1\times 1\times d_i}$ 为第 $i$ 个单词与第 $j$ 个单词间的表示；$\boldsymbol{W}_1\in\mathbb{R}^{2d\times d_i}$ 和 $\boldsymbol{W}_2\in\mathbb{R}^{2d\times d_i}$ 分别表示可训练权重。

由于表格 $\boldsymbol{M}$ 无法体现信息之间的方向性，使用一个能体现信息方向性的结构具有重要作用。受图像处理中流行的多信息融合结构启发，本文采用了条件层归一化（conditional layer normalization，CLN）[25]计算。CLN 将归一化结构中对应的偏置和权重变成关于待融合条件的函数。CLN 的具体计算如式（8）所示：

$$z_{ij}=\mathrm{CLN}(\boldsymbol{h}_i,\boldsymbol{h}_j)=\frac{\boldsymbol{h}_j-E(\boldsymbol{h}_j)}{V(\boldsymbol{h}_j)}*\boldsymbol{W}_3\boldsymbol{h}_i+\boldsymbol{W}_4\boldsymbol{h}_i \tag{8}$$

其中，$E(\boldsymbol{h}_j)$ 为 $\boldsymbol{h}_j$ 的均值，$V(\boldsymbol{h}_j)$ 为 $\boldsymbol{h}_j$ 的方差。由于单词对 $<t_i,t_j>$ 是具有方向性的，如“the → Senate”和“middle → east”，本文以单词 $t_i$ 对应表示作为待融合的条件信息，层归一化可通过两个不同的可训练权重 $\boldsymbol{W}_3$ 和 $\boldsymbol{W}_4$ 将条件信息 $\boldsymbol{h}_i$ 和映射到不同的空间来体现条件的方向信息，从而得到融合后的表示 $\boldsymbol{z}\in\mathbb{R}^{n\times n\times 2d}$。

将体现方向信息的表示 $\boldsymbol{z}$ 与表格 $\boldsymbol{M}$ 进行拼接得到更好的交互表示 $\boldsymbol{Z}\in\mathbb{R}^{n\times n\times(d_i+d_i)}$。具体如式（9）所示：

$$\boldsymbol{Z}=[\boldsymbol{M};f(\boldsymbol{z}\boldsymbol{W}_5+\boldsymbol{b}_5)] \tag{9}$$

式中，$f(\cdot)$ 表示 GELU 函数；$\boldsymbol{W}_5\in\mathbb{R}^{2d\times d_i}$ 为可训练权重，$\boldsymbol{b}_5$ 表示可学习参数。

## 2.3 多尺度注意力

注意力机制[26]可根据目标的重要程度进行权重分配，突出某些重要特征，从而有效捕获上下文信息。而多头注意力机制通过多个子空间表示来提升模型关注不同特征的能力。通过多次重复对矩阵 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行不同的线性映射，将每个注意力头的矩阵拼接起来，通过与随机矩阵相乘得到最终多头注意力的输出。计算如式（10）和式（11）所示：

$$\boldsymbol{head}_h=\mathrm{softmax}\left(\frac{(\boldsymbol{H}\boldsymbol{W}_h^Q)(\boldsymbol{H}\boldsymbol{W}_h^K)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{H}\boldsymbol{W}_h^V \tag{10}$$

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=[\boldsymbol{head}_1;\boldsymbol{head}_2;\cdots;\boldsymbol{head}_h]\boldsymbol{W} \tag{11}$$

式中，$\boldsymbol{head}_h$ 表示第 $h$ 个注意力头的计算结果；$\boldsymbol{W}_h^Q$、$\boldsymbol{W}_h^K$、$\boldsymbol{W}_h^V$ 分别表示可训练的权重矩阵，$d_k$ 是超参数；$\boldsymbol{W}$ 是 $h$ 个注意力头计算结果拼接后的权重矩阵。

由于多头注意力在同一注意力层的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的长度一致，仅在单一尺寸下提取特征，限制每个注意力头捕获多尺度特征的能力，同时也忽略了不同尺度特征之间的相互依赖关系，易出现实体漏检或错检的情况。因此，本文提出的多头注意力可以在不同注意力头上聚合不同长度的文本信息，以用于提取粗细粒度的特征，如图 2 所示。

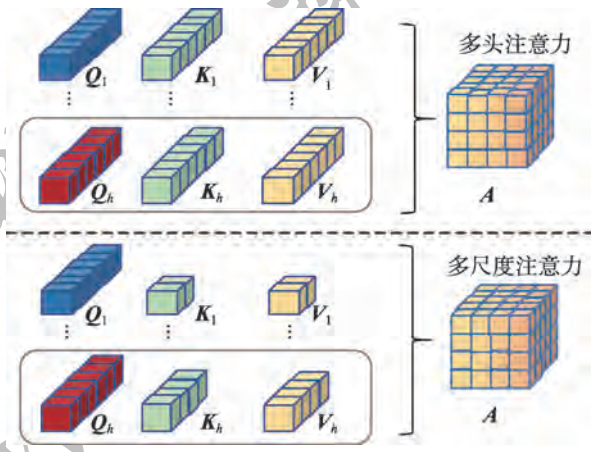


**图 2 多头注意力与多尺度注意力的对比**

Fig.2 Comparison between proposed multi-scale attention with multi-head attention

从图2可以看出,本文在同一注意力层的不同子空间中的 $\boldsymbol{K}$ 、$\boldsymbol{V}$ 的长度是不同的。具体来说,$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 将进行不同大小的下采样,如式(12)和式(13)所示:

$$\boldsymbol{Q}_h=\boldsymbol{Z}\boldsymbol{W}_h^Q;\boldsymbol{K}_h=g(\boldsymbol{Z},s_L)\boldsymbol{W}_h^K \tag{12}$$

$$\boldsymbol{V}_h=g(\boldsymbol{Z},s_L)\boldsymbol{W}_h^V;\boldsymbol{V}_h=\boldsymbol{V}_h+g(\boldsymbol{V}_h,s) \tag{13}$$

式中,$g(\boldsymbol{Z},s_L)$ 表示第 $h$ 个注意力头的多尺度融合;$s_{(\cdot)}$ 表示下采样率,这里可根据下采样率 $s_{(\cdot)}$ 的不同来捕获不同尺度的信息;$\boldsymbol{W}_h^Q$ ,$\boldsymbol{W}_h^K$ ,$\boldsymbol{W}_h^V\in\mathbb{R}^{(d_i+d_i)\times d_k}$ 分别表示可训练的权重矩阵。此外,由于在采样过程中难免出现信息丢失,特别地对采样后的 $\boldsymbol{V}_h$ 进行了局部增强,保证输出特征表达能力。

卷积神经网络(convolutional neural network,CNN)[27]是一种前馈神经网络,可以通过多个卷积核对文本特性进行自动提取。每一个卷积核都会对当前输出进行特征提取并产生对应特征信息。这里,通过 $L$ 个二维卷积进行不同尺度的下采样。通过不同大小的卷积来关注不同长度的特性信息。具体如式(14)和式(15)所示:

$$g(\boldsymbol{Z},s_l)=\varphi(\sigma(\boldsymbol{z}_{i+r_l\times s_l;j+r_l\times s_l}\boldsymbol{W}_c)) \tag{14}$$

$$g(\boldsymbol{Z},s_L)=Concat(g(\boldsymbol{Z},s_1),g(\boldsymbol{Z},s_2),\cdots,g(\boldsymbol{Z},s_L)) \tag{15}$$

其中,$\varphi$ 表示层规范(LayerNorm)运算;$\sigma$ 表示sigmoid激活函数;$s_{(\cdot)}$ 为卷积核大小;$r_l$ 为步长。不同的卷积核可以有不同的感受野。当卷积核越大时,将保留更多的长距离的信息来捕获较大目标。反之,当卷积核越小时,可以关注细节信息来捕获小目标。通过卷积大小的调整,选择性保留不同粒度的细节特征。在卷积操作后使用前馈神经网络将卷积后的结果进行归一化。将 $l$ 个归一化后表示进行拼接,输出中包含多尺度特征的矩阵。此外,在采样过程中,通过1×1二维卷积 $s$ 对 $V$ 进行局部加强表示。在特征融合时保留句子上下文特征的同时,对边界细节更加关注,以提高模型的自适应能力。

考虑到不同尺度卷积核提取的特征对模型作用不同,本文将多尺度特征映射输入注意力模块中,计算不同尺度卷积核提取特征映射的权重,为模型捕获关键特征。多尺度注意力计算过程如式(16)和式(17)所示:

$$\boldsymbol{head}_h=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_h\boldsymbol{K}_h^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_h \tag{16}$$

$$\boldsymbol{A}=[\boldsymbol{head}_1;\boldsymbol{head}_2;\cdots;\boldsymbol{head}_h]\boldsymbol{W} \tag{17}$$

式中,$\boldsymbol{W}\in\mathbb{R}^{(d_i+d_i)\times(d_i+d_i)}$ 表示可训练的权重矩阵。

与传统的多头注意力操作相比,该模块可以通过同一注意层同时进行不同尺度特征融合,进而捕获粗细粒度的特征信息。多尺度注意力的计算结果送入到下一层进行最后分类预测。

### 2.4 类别判别层

将多尺度注意力计算得到的表示向量 $\boldsymbol{A}$ 经过多层感知机,计算第 $i$ 单词和第 $j$ 个单词成为实体的开始和结束概率。具体如式(18)所示:

$$\boldsymbol{a}_{ij}'=f_m(\boldsymbol{a}_{ij}) \tag{18}$$

其中,$\boldsymbol{a}_{ij}'\in\mathbb{R}^K$ ,$K$ 为实体类型数量;$f_m(\cdot)$ 是多层感知机运算。

此外,通过编码层得到的 $\boldsymbol{H}$ ,借助双仿射计算 $\boldsymbol{h}_i$ 和 $\boldsymbol{h}_j$ 之间的得分。

$$\boldsymbol{h}_{ij}'=\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{U}\boldsymbol{h}_j+\boldsymbol{W}_6[\boldsymbol{h}_i;\boldsymbol{h}_j]+\boldsymbol{b}_6 \tag{19}$$

其中,$\boldsymbol{h}_{ij}'\in\mathbb{R}^K$ ;$\boldsymbol{U}\in\mathbb{R}^{2d_i\times K\times 2d_i}$ 是在句子中第 $i$ 个单词和第 $j$ 个单词实体类别后验概率建模,$\boldsymbol{W}_6\in\mathbb{R}^{2d_i\times K}$ 为句子中第 $i$ 个词或尾实体中第 $j$ 个词实体类别后验概率建模,$\boldsymbol{b}_6$ 表示可学习参数。

最后,通过合并式(18)和式(19)的得分来对候选实体进行判定,具体如式(20)和式(21)所示。对于某一实体类型,若实体得分超过设定的阈值,则保留为最终实体;否则将其过滤。

$$y_{ij}=a_{ij}'+h_{ij}' \tag{20}$$

$$p_{ij}=\frac{\exp(y_{ij})}{\sum_{k=1}^{K}y_{ij}} \tag{21}$$

### 2.5 训练

本文利用交叉熵训练模型,如式(22)所示。在训练过程中最小化损失函数。

$$Loss=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{K}y_{ij}\ln p_{ij} \tag{22}$$

## 3 实验与结果分析

为了验证本文提出的多尺度注意力的命名实体识别模型的效果,本文在GENIA和ACE2005英文数据集进行了实验。并选取了其他基于深度学习的命名实体识别模型进行了对比,使用准确率Precision($P$)、召回率Recall($R$)和 $F1$ 值作为主要评估指标,验证本模型的各项性能。

### 3.1 实验数据集

GENIA数据集是基于语料库GENIA3.0.2构建

而成，数据来源于生物医学。该数据集包含了DNA、RNA、protein、cell line和cell type 5种实体类型。该数据集采用文献[23]的数据集划分方法，按8.1∶0.9∶1.0进行训练集、验证集及测试集的划分。

ACE2005英文数据集是2006年由语言数据联盟发布的，数据来源于微博、广播新闻、新闻组、广播对话等，用于各项NLP任务。该数据集共599篇英文文档，包含PER、ORG、GPE、LOC、FAC、VEH和WEA 7种实体类型。该数据集采用文献[20]的数据集划分方法，按8∶1∶1进行训练集、验证集及测试集的划分。

两个数据集基本信息如表1所示。

**表1 数据集基本信息**

Table 1 Basic information of datasets

| 数据集 | 训练集 | 验证集 | 测试集 | 实体类型 |
|---|---|---|---|---|
| GENIA | 11 126 | 3 709 | 3 709 | 5 |
| ACE2005 | 7 683 | 960 | 960 | 7 |

## 3.2 评价指标

为了测试本文模型的有效性，本文采用准确率 $P$ 、召回率 $R$ 和 $F1$ 值来评价实验结果。只有当实体位置和类型两者均识别正确时结果才认为是正确的。具体计算公式如下：

$$P=\frac{M}{N}\times 100\% \tag{23}$$

$$R=\frac{M}{K}\times 100\% \tag{24}$$

$$F1=\frac{2M}{K+N}\times 100\% \tag{25}$$

其中，$M$ 表示正确识别出的实体个数，$N$ 表示识别出的实体个数，$K$ 表示标准结果中的实体个数。

## 3.3 实验环境与参数设置

本文提出的模型在Python3.7和Pytorch1.8的环境下进行实验。训练过程中使用BioBERT和BERT-Large分别作为GENIA和ACE2005数据集的预训练模型，实验采用Adam优化器。参数设置如表2所示。

**表2 参数设置**

Table 2 Setting of parameters

| 参数 | 设定值 |
|---|---|
| Batch size | 6 |
| Epoch | 30 |
| Dropout | 0.5 |
| $d$ | 768/1 024 |
| $d_l$ | 512/1 024 |
| $d_t$ | 20 |
| $s_{(\cdot)}$ | [1, 2, 4] |
| 学习率 | 1E−3 |

## 3.4 结果分析

为了验证本文模型的有效性，将本文模型与其他基线模型进行了对比分析。

(1)Ju等[11]通过动态叠加平面NER层来识别嵌套实体。

(2)Wang等[13]提出了一种新的分层模型，将文本递归地堆叠成金字塔形状。同时，设计了逆向金字塔允许层之间的双向交互。

(3)Wang和Lu[16]提出了一种新的分段超图表示模型，并利用BiLSTM学习不同级别的表示。

(4)Xia等[20]提出了包含检测器和分类器两个模块的实体识别方法。检测器检测所有可能的命名实体跨度，而分类器的目的是将检测到的命名实体分类为预定义的命名实体类别。

(5)Yu等[21]利用了依赖树的思想，通过双仿射去计算实体跨度的分数。

(6)Yan等[23]采用了Seq2Seq框架，并采用指针网络直接生成实体。

上述模型中，(1)和(2)为基于分层的方法；(3)为基于超图的方法；(4)~(6)为基于跨度的方法。不同模型在GENIA和ACE2005数据集上的性能如表3和表4所示，其中加粗表示最优结果。

**表3 GENIA数据集上的对比结果**

Table 3 Comparison results on GENIA dataset

单位：%

| 模型 | $P$ | $R$ | $F1$ |
|---|---|---|---|
| Ju等[11] | 78.5 | 71.3 | 74.7 |
| Wang等[13] | 79.5 | 78.9 | 79.2 |
| Wang和Lu[16] | 77.0 | 73.3 | 75.1 |
| Yu等[21] | **81.8** | 79.3 | 80.5 |
| Yan等[23] | 78.9 | 79.6 | 79.2 |
| 本文模型 | 81.3 | **82.1** | **81.7** |

**表4 ACE2005英文数据集上的对比结果**

Table 4 Comparison results on ACE2005 English dataset

单位：%

| 模型 | $P$ | $R$ | $F1$ |
|---|---|---|---|
| Ju等[11] | 74.2 | 70.3 | 72.2 |
| Wang等[13] | 84.0 | 85.4 | 84.7 |
| Wang和Lu[16] | 76.8 | 72.3 | 74.5 |
| Xia等[20] | 79.0 | 77.3 | 78.2 |
| Yu等[21] | 85.2 | 85.6 | 85.4 |
| Yan等[23] | 83.2 | 86.4 | 84.7 |
| 本文模型 | **86.0** | **87.6** | **86.8** |

从表3可以看出,本文模型在GENIA数据集上与其他基线模型相比召回率和*F*1值都达到了最优。从表4可以看出,本文模型在ACE2005英文数据集上各项指标均要优于其他基线模型。实验结果表明了本文方法有效地在同一注意层中聚合不同尺度的文本信息,进而可以同时关注到粗细粒度特征信息,最终将学习到的信息综合考虑来提升模型的识别能力。

对比最好的分层方法,Wang等[13]利用金字塔结构,通过逆向金字塔加强了层之间的双向交互,减少了错误传播,验证了丰富的特征表示对实体识别任务的重要性。本文模型融合方向信息获得更加丰富的交互信息,同时通过多尺度注意力同时捕获不同尺度的特征,更好地面对嵌套语义情况。

对比超图方法,Wang和Lu[16]利用了BiLSTM去提取句子中的长距离依赖,但是在两个数据集上并未取得较好的效果。这是因为两个数据集中包含大量嵌套语义,需要获得更多细节信息进行实体识别。本文模型通过不同大小的卷积核进行下采样,在捕获长距离信息的同时关注到更多的边界细节信息,避免实体漏检或错检的情况。

对比最好的跨度方法,Yu等[21]进行了字、词级信息的特征融合,并利用双仿射操作来约束预测的命名实体,在两个数据集上识别效果都低于本文模型,而优于其他基线模型。这是因为字向量和词向量的特征融合可以缓解未登陆词对实体的影响,而双仿射操作可以增加实体边界之间的信息交互。而本文模型不仅使用了双仿射操作,而且增加多尺度注意力得到粗细粒度的语义信息,提高模型识别效果。

### 3.5 不同尺度注意力对模型性能影响分析

为了研究不同尺度注意力对模型性能的影响,本节设计了三个不同尺度注意力的策略进行对比:尺度1将模型中的下采样率设置为[1,2];尺度2将下采样率设置为[2,4];尺度3将下采样率设置为[4,8],这里统一将注意力头设置为2。

从表5可以看出,相比尺度1策略,采用尺度2策略的*F*1值在两个数据集分别提升了0.3个百分点和0.6个百分点。这是因为尺度1将模型中的下采样率设置较小,可以捕获更多的细粒度的信息,但是同时也容易忽略长距离信息提取。相比尺度3策略,采用尺度2策略的*F*1值在两个数据集分别提升了1.1个百分点和1.6个百分点。这是因为尺度3将模型中的下采样率设置较大,可以更好地聚合更长距离的信息,但是同时丢失了过多文本的邻近特征,造成模型识别能力降低。实验结果表明,选择合适的下采样率将有效地提取不同尺度的特征,更有助于嵌套语义的识别,提高模型识别效果。

**表5 不同尺度注意力的实验结果**

Table 5 Experimental results of attention of different scales 单位:%

| 策略 | GENIA | | | ACE2005 | | |
|---|---|---|---|---|---|---|
| | *P* | *R* | *F*1 | *P* | *R* | *F*1 |
| 尺度1 | 81.6 | 81.2 | 81.4 | 85.8 | 86.5 | 86.2 |
| 尺度2 | 81.3 | 82.1 | 81.7 | 86.0 | 87.6 | 86.8 |
| 尺度3 | 80.7 | 80.6 | 80.6 | 84.9 | 85.6 | 85.2 |

### 3.6 消融实验分析

为了验证模型各组成部分的有效性,本节设计了以下消融实验,w/o表示去掉某模块。“w/o CLN”表示模型在交互表示层不融入方向信息;“w/o MA”表示模型中去掉多尺度注意力;“w/o BF”表示模型在类别判别层去掉双仿射操作。消融实验在两个数据集上的实验结果如表6所示。为了更直观表现模型性能,各指标对比的柱状图如图3所示。

**表6 两个数据集上的消融实验结果**

Table 6 Results of ablation experiment on two datasets 单位:%

| 模型 | GENIA | | | ACE2005 | | |
|---|---|---|---|---|---|---|
| | *P* | *R* | *F*1 | *P* | *R* | *F*1 |
| w/o CLN | **83.1** | 78.6 | 80.8 | 85.2 | 86.7 | 85.9 |
| w/o MA | 80.2 | 80.0 | 80.1 | 83.7 | 86.4 | 85.1 |
| w/o BF | 82.2 | 80.6 | 81.4 | 85.2 | 86.5 | 85.8 |
| 本文模型 | 81.3 | **82.1** | **81.7** | **86.0** | **87.6** | **86.8** |

从表6和图3可以发现,本文模型相比“w/o CLN”模型在两个数据集上均有提升,证明融合的方向信息可以帮助本文模型得到更加丰富的语义信息;本文模型相较于“w/o MA”模型在两个数据集上各评价指标均有显著提升,证明了多尺度注意力的有效性。通过在不同子空间表示下同时提取多尺度特征,关注到更多粗细粒度信息,更好地对嵌套语义进行识别。从本文模型与“w/o BF”模型的对比结果可以看出,双仿射操作可以有助于提高实体识别性能。

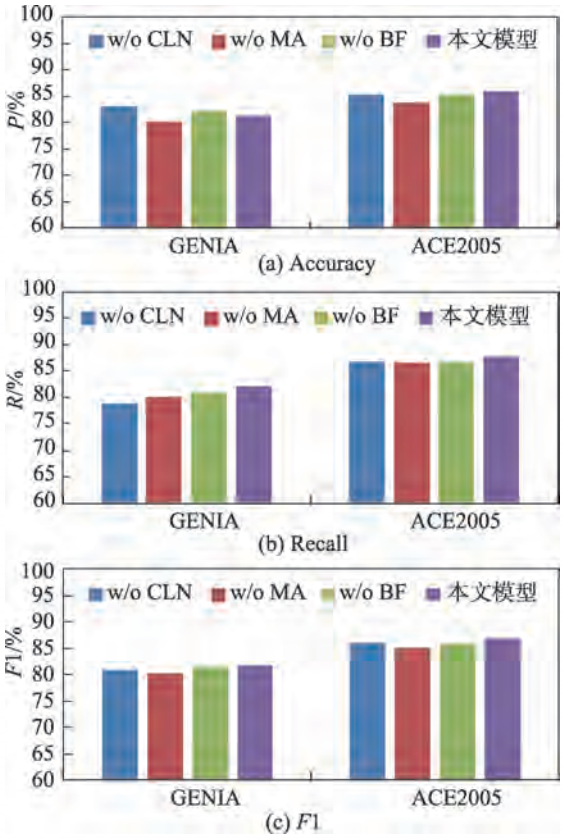


**图3 消融实验可视化**

Fig.3 Ablation experiment visualization

### 3.7 案例分析

为了更直观地展示本文模型解决嵌套语义的效果，本部分设计了具体案例分析的实验。具体地，在ACE2005和GENIA数据集上各选取一个具有代表性的例子分析本文模型的识别效果。表7、表8给出了本文模型和另两个对比模型在这两个例子上的识别结果。

从表7、表8中可以发现，本文模型可以准确识别出实例中的实体边界并进行分类。然而文献[10]的方法无法完全识别出所有实体，尤其面对嵌套语义的情况，如实例1中“a small plane that disappeared sunday in massachusetts”未被检测出来。主要是由于该方法采用了序列标注方法，每一个单词只能分配一个标签，只能解决非嵌套结构。此外，文献[20]的方法可以识别出嵌套实体，但是误检了多个实体，如实例2中“Octamer transcription factors”和“transcription factors”都被识别为protein类型。主要是由于该方法检测了所有可能的实体跨度作为候选实体来进行预测，当负样本与真实实体高度重叠时，由于它们语义表示相似，极易产生错误。以上实例表明，本文提出的模型能准确定位不同粒度的实体，并准确地对实体进行分类。

**表7 实例1分析**

Table 7 1st case study

| | |
|---|---|
| 文献 | the search for a small plane that disappeared sunday in massachusetts has a bittersweet ending. |
| Gold label | **GPE**: Massachusetts; **VEH**: that; **VEH**: a small plane that disappeared sunday in massachusetts |
| 文献[10] | **GPE**: Massachusetts; **VEH**: that |
| 文献[20] | **GPE**: Massachusetts; **VEH**:that; **VEH**: a small plane; **VEH**: a small plane that disappeared Sunday; **VEH**:a small plane that disappeared sunday in massachusetts |
| Ours | **GPE**: Massachusetts; **VEH**: that<br>**VEH**: a small plane that disappeared sunday in massachusetts |

**表8 实例2分析**

Table 8 2nd case study

| | |
|---|---|
| 文献 | Octamer transcription factors and the cell type-specificity of immunoglobulin gene expression. |
| Gold label | **protein**: Octamer transcription factors<br>**DNA**: cell type-specificity of immunoglobulin gene |
| 文献[10] | **protein**: Octamer transcription factors<br>**DNA**: immunoglobulin gene |
| 文献[20] | **protein**: Octamer transcription factors<br>**protein**: transcription factors<br>**DNA**: cell type-specificity of immunoglobulin gene<br>**DNA**: cell immunoglobulin gene |
| Ours | **protein**: Octamer transcription factors<br>**DNA**: cell type-specificity of immunoglobulin gene |

## 4 结束语

针对当前命名实体识别研究存在提取特征尺度单一、边界信息利用不足的问题，本文设计了一种多尺度注意力的命名实体识别方法，通过注意力机制可以同时提取不同尺度的特征，在同一注意力层同时提取浅层细节信息和深层语义信息，有效提高了命名识别的识别能力。本文方法在两个公共数据集上识别效果高于现有基线模型。下一步将研究如何利用外部知识来提升模型，或者将多尺度注意力机制应用到其他任务研究中。

## 参考文献：

[1] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[2] YIH S W, CHANG M W, HE X, et al. Semantic parsing via staged query graph generation: question answering with

knowledge base[C]//Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, Jul 26-31, 2015. Stroudsburg: ACL, 2015: 1321-1331.

[3] WANG X, XU Y, HE X, et al. Reinforced negative sampling over knowledge graph for recommendation[C]//Proceedings of the 2020 Web Conference, Taipei, China, Apr 20-24, 2020. New York: ACM, 2020: 99-109.

[4] GUPTA M, BENDERSKY M. Information retrieval with verbose queries[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Aug 9-13, 2015. New York: ACM, 2015: 1121-1124.

[5] 何儒汉, 唐娇, 史爱武, 等. 基于实体消岐和多粒度注意力的知识库问答[J]. 计算机工程与设计, 2022, 43(2): 560-566.
HE R H, TANG J, SHI A W, et al. Knowledge base question answering based on entity disambiguation and multiple granularity attention[J]. Computer Engineering and Design, 2022, 43(2): 560-566.

[6] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, Jun 12-17, 2016: 260-270.

[7] LI X, YAN H, QIU X, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 6836-6842.

[8] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the 7th Conference on Natural Language Learning, Edmonton, May 31-Jun 1, 2003: 188-191.

[9] LYU C, CHEN B, REN Y, et al. Long short-term memory RNN for biomedical named entity recognition[J]. BMC Bioinformatics, 2017, 18(1): 1-11.

[10] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991, 2015.

[11] JU M, 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, New Orleans, Jun 1-6, 2018: 1446-1459.

[12] 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.

[13] WANG J, SHOU L, CHEN K, et al. Pyramid: a layered model for nested named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 5918-5928.

[14] LU W, ROTH D. Joint mention extraction and classification with mention hypergraphs[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Sep 17-21, 2015: 857-867.

[15] MUIS A O, LU W. Labeling gaps between words: recognizing overlapping mentions with mention separators[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Sep 9-11, 2017: 2608-2618.

[16] WANG B, LU W. Neural segmental hypergraphs for overlapping mention recognition[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Oct 31-Nov 4, 2018: 204-214.

[17] KATIYAR A, CARDIE C. Nested named entity recognition revisited[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Jun 1-6, 2018. Stroudsburg: ACL, 2018: 861-871.

[18] XU M, JIANG H, 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, Vancouver, Jul 30-Aug 4, 2017. Stroudsburg: ACL, 2017: 1237-1247.

[19] LUAN Y, WADDEN D, HE L, 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: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 3036-3046.

[20] XIA C, ZHANG C, YANG T, et al. Multi-grained named entity recognition[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 1430-1440.

[21] YU J, BOHNET B, POESIO M. Named entity recognition as dependency parsing[C]//Proceedings of the 58th Annual

Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 6470-6476.

[22] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 5849-5859.

[23] YAN H, GUI T, DAI J, et al. A unified generative framework for various NER subtasks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Aug 1-6, 2021. Stroudsburg: ACL, 2021: 5808-5822.

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 4171-4186.

[25] DE VRIES H, STRUB F, MARY J, et al. Modulating early visual processing by language[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017: 6594-6604.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017: 5998-6008.

[27] CHEN H, LIN Z, DING G, et al. GRN: gated relation network to enhance convolutional neural network for named entity recognition[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Jan 27-Feb 1, 2019. Menlo Park: AAAI, 2019: 6236-6243.

**唐瑞雪**(1987—),女,贵州人,博士研究生,主要研究方向为自然语言处理、信息抽取、机器学习。

**TANG Ruixue**, born in 1987, Ph.D. candidate. Her research interests include natural language processing, information extraction and machine learning.

**秦永彬**(1980—),男,山东人,博士,教授,博士生导师,CCF高级会员,主要研究方向为智能计算、机器学习、算法设计。

**QIN Yongbin**, born in 1980, Ph.D., professor, Ph.D. supervisor, CCF senior member. His research interests include intelligent computing, machine learning and algorithm design.

**陈艳平**(1980—),男,贵州人,博士,副教授,CCF会员,主要研究方向为人工智能、自然语言处理等。

**CHEN Yanping**, born in 1980, Ph.D., associate professor, CCF member. His research interests include artificial intelligence, natural language processing, etc.