·计算机技术与应用·

基于 LSTM-SNP 的命名实体识别

邓　琴, 陈晓亮*, 陈龙齐
(西华大学计算机与软件工程学院, 四川 成都 610039)

摘　要：脉冲神经 P 系统(SNPs)是抽象于生物神经元信息交互机制的高效并行计算系统。LSTM-SNP 首次结合非线性 SNP 和长短期记忆神经网络(LSTM), 从而形成门控机制可解释的深度学习通用模型。LSTM-SNP 作为传统序列分析模型 LSTM 的最新变体, 在处理典型自然语言处理序列分析问题的性能表现未见相关研究。文章以命名实体识别任务为基础, 通过在 LSTM-SNP 上增补不同的深度学习组件, 对 LSTM-SNP 与传统 LSTM 以及其变体 BiLSTM 的性能差异进行了全面分析, 为在自然语言处理任务中使用 LSTM-SNP 模型提供可靠的改进参考。通过以 CoNLL-2003 和 OntoNotes5.0 为标准数据集的对比实验, 发现: LSTM-SNP 模型与 LSTM 模型具有类似的实体识别性能, 但随着预处理的操作, LSTM-SNP 模型的整体性能提升更为显著; LSTM-SNP 模型对命名实体的识别是一种行之有效的方法, 且具有较大的应用潜力。

关键词：LSTM-SNP 模型; 命名实体识别; 模型性能测评

中图分类号：TP311.52；TP391　　**文献标志码：**A　　**文章编号：**1673–159X(2023)05 − 0028 − 10

doi:10.12198/j.issn.1673 − 159X.4844

Named Entity Recognition Based on the LSTM-SNP

DENG Qin, CHEN Xiaoliang*, CHEN Longqi
(*School of Computer and Software Engineering*, *Xihua University*, *Chengdu* 610039 *China*)

Abstract: Spiking neural P systems (SNPs) are efficient parallel computing systems abstracted from the mechanism of information exchange between biological neurons. For the first time, LSTM-SNPs combine nonlinear SNPs and long short-term memory (LSTM) to form a universal deep learning model that gating mechanisms can explain. LSTM-SNPs, the latest variant of the traditional sequence analysis model LSTM, has yet to be studied on the performance of typical sequence analysis in natural language processing. This paper comprehensively analyzes the performance difference in the named entity recognition tasks between LSTM-SNPs, traditional LSTMs, and its variant BiLSTM by adding different deep learning components. The study provides a reliable reference for applying the LSTM-SNP model in natural language processing tasks. The results of comparative experiments based on CoNLL-2003 and OntoNotes 5.0 data sets indicate the LSTM-SNP model has a similar entity recognition performance to the LSTM model.

收稿日期:2023 − 02 − 16

基金项目:国家自然科学基金项目(61902324); 四川省科技厅重点研发项目(2023YFS0424)。

*** 通信作者:**陈晓亮(1984—), 男, 教授, 博士, 硕士生导师, 主要研究方向为自然语言处理。
ORCID: 0000 − 0002 − 8201 − 9631　　E-mail: chenxl@mail.xhu.edu.cn

引用格式:邓琴, 陈晓亮, 陈龙齐. 基于 LSTM-SNP 的命名实体识别[J]. 西华大学学报(自然科学版), 2023, 42(5): 28 − 37.
DENG Qin, CHEN Xiaoliang, CHEN Longqi. Named Entity Recognition Based on the LSTM-SNP[J]. Journal of Xihua University(Natural Science Edition), 2023, 42(5): 28 − 37.

In further research,the overall model performance can be improved significantly with the pretreatment operation. The results show the LSTM-SNP model is an effective method for named entity recognition and has great application potential.

Keywords: LSTM-SNP model；named entity recognition；model performance evaluation

脉冲神经 P 系统(SNPs)是从生物信息学的神经元之间的脉冲通信机制中抽象出来的一类分布式并行计算模型[1]。1 个脉冲神经 P 系统,通常由 4 个基本元素构成:结构、数据、规则集和对规则的控制方法[2]。结构细分为膜结构和数据结构 2 部分。膜结构由有向图进行刻画,其中图的节点和边分别表示神经元和神经元间的突触。数据结构形式化为脉冲多重集。1 个神经元的内部机制包含脉冲和规则。系统中的数据一般是由神经元的脉冲统计个数来描述。规则是脉冲神经 P 系统完成信号传递的核心[3]。SNP 系统的规则分为 2 个类别:脉冲规则和遗忘规则[4]。前者又叫作点火规则,表示消耗脉冲且同时产生新脉冲,后者仅消耗而不会产生新脉冲。

与传统 SNP 系统的区别在于,非线性脉冲神经 P 系统(NSNPs)[5] 通过预定义的神经元状态非线性函数实现脉冲的消耗和产生。因此,NSNP 系统适用于捕获复杂系统中的非线性特征。长短记忆神经网络(LSTMs)[6] 属于循环神经网络(RNNs)的变体。1 个 LSTM 模型包含 1 个隐藏状态和 3 个门结构(遗忘门、输入门和输出门),共同实现神经元信息传递的调节。受到 NSNP 系统脉冲和遗忘规则的启发,Liu 等 [7] 基于 LSTM 模型提出了新的循环神经网络模型,即 LSTM-SNP 模型。该模型只由一个非线性脉冲神经元组成,具有非线性脉冲机制(非线性脉冲消耗和产生)和非线性门函数(重置、消耗和生成)。

循环模式可以较好地解决序列分析问题,例如,时间序列的预测。然而 LSTM-SNP 作为传统序列分析模型 LSTM 的最新变体,在处理典型自然语言处理序列分析问题,如命名实体识别(NER)的性能表现未见相关研究。本文将序列分析模型 LSTM-SNP 用于解决命名实体识别任务,通过添加不同的深度学习组件,模型的性能得到显著的提升,同时,设计了多组对比实验,比较 LSTM-SNP 模型、传统的 LSTM 和双向长短记忆网络(BiLSTM)的性能。

1　相关工作

本文旨在研究 LSTM-SNP 模型在命名实体识别任务上的适应性问题,以评估模型在自然语言处理底层任务中的性能和潜力。命名实体识别任务是指在不规则文本中识别具有代表性的特定实体。其主要研究策略、方法根据时间的先后顺序分为:基于规则、基于机器学习和基于深度学习。

1)基于规则的命名实体处理,以语法为基础。Etzioni 等[8] 和 Wang 等[9] 分别提出了基于地名词典和基于词汇句法模式引用规则的 2 种经典方法。这类方法具有设计简单、复杂性低等优点,但识别效果严重依赖领域专家对语料库的标注[10]。此外,在处理大规模数据集时基于规则的模型性能具有局限性。

2)机器学习已经成为研究命名实体识别的主流技术。该任务在机器学习领域被定义为多分类序列标注问题。主要技术包括最大熵(MaxEnt)、支持向量机(SVMs)、隐藏马尔可夫模型(HMMs)、条件随机场(CRFs)等。Makino 等[11] 基于语音和单词形式构建了系列人工特征,继而用 HMM 提取特征,将其合并,并使用 SVM 计算实体识别结果。Krishnan 等[12] 利用 2 个 CRFs 来提取实体识别中的局部特征,并输出由逻辑前向 CRF 提取的特征信息。这些模型克服了基于规则的缺陷,然而,由于无法捕获更多的上下文信息,当面对句子过长的场景会导致模型性能的降低。

3)近期一些文献强调了神经网络方法在解决 NER 问题中的作用,包括长短期记忆神经网络(LSTMs)、卷积神经网络(CNNs)及其变体。神经网络方法避免了手动特征提取。Luo 等[13] 提出了一个基于注意力的具有 CRF 层的双向长短期记忆神经网络(Att-BiLSTM-CRF),继而训练一个高准

确度的模型来识别已命名的实体。Li 等[14] 建立的 BiLSTM-CNN 模型表明, CNN 作为模型组件可以显著提高实体识别的精度。Li 等[15] 提出了一种新的替代方法 W2NER, 它将 NER 建模为词—词关系分类。此外, Bert、LSTM 和多重二维扩展卷积(DConv)的有机组合可以较优地处理 NER 问题。

2 LSTM-SNP 模型实验设计

本章对文中提到的命名实体任务进行形式化表述。设$L=\{L_1,\cdots,L_t,\cdots,L_n\}$为一个有标签的实例文本训练集, $\theta=\{\theta_1,\cdots,\theta_t,\cdots,\theta_m\}$为一组类别标签, 如位置、组织、其他等。实例$L_t$中的词语, 记为tk, 都被分配了一个标签$\varepsilon\in\theta$。如果实例的词语是已命名实体的元素, 则它的标签是该实体的类别。NER 模型的基本处理步骤为: 首先, 通过在模型的嵌入层中使用独热编码技术, 将具有n个单词的句子X表示为向量序列$X=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_n\}$; 其次, 将嵌入操作后的向量输入到 LSTM、BiLSTM 或 LSTM-SNP 中, 以生成区分实体的预测标签; 最后, 使用相关的评价指标来评估模型的有效性。

2.1 LSTM-SNP 模型结构

RNN 技术在自然语言处理 (NLP) 研究中得到了广泛的应用。但是, 随着模型层数的叠加, RNN 网络容易发生梯度消失和梯度爆炸的问题。长短期记忆神经网络(LSTMs)[6] 作为 RNN 的变体, 设计了 3 个门控机制来调整细胞状态, 如图 1 所示。LSTM 单元在时间步t的遗忘门、输入门、输出门通常分别形式化为函数f_t, i_t, o_t。其输入的向量用$\boldsymbol{x}_t$表示。C_t表示单元在时间步骤t的状态。

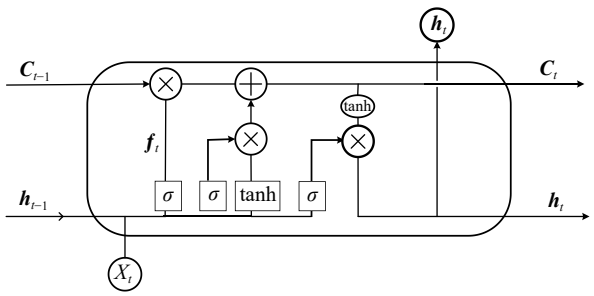

图 1 LSTM 模型细胞结构图

Fig. 1 Cell structure diagram of LSTM model

由于使用了 sigmoid 型函数, LSTM 能够有效地选择放弃和保留的信息。3 个门控单元的连接与控制的计算公式为:

$$\left.\begin{aligned}
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_f\cdot[h_{t-1},\ x_t]+b_f)\\
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i\cdot[h_{t-1},\ x_t]+b_i)\\
\tilde{\boldsymbol{C}}_t &= \tanh(\boldsymbol{W}_c\cdot[h_{t-1},\ x_t]+b_c)\\
\boldsymbol{C}_t &= \boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\tilde{\boldsymbol{C}}_t\\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_o\cdot[h_{t-1},\ x_t]+b_o)\\
\boldsymbol{h}_t &= \boldsymbol{O}_t*\tanh(\boldsymbol{C}_t)
\end{aligned}\right\}\quad(1)$$

Liu 等[7] 认为, LSTM-SNP 是在 LSTM 基础上用不同的非线性门函数、状态方程和基于膜计算规则的输入输出进行的模型重构。图 2 表示了新型门机制: 复位门r_t、消耗门c_t和生成门o_t。复位门根据当前的输入、上一时刻的状态和偏置, 决定前一状态的复位程度。

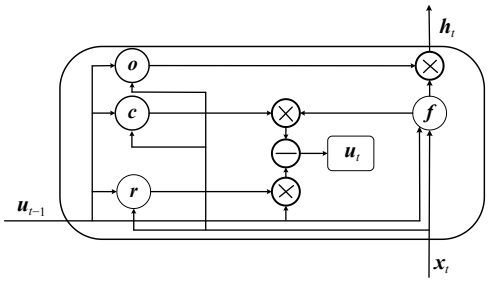

图 2 LSTM-SNP 模型细胞结构图

Fig. 2 Cell structure diagram of LSTM-SNP model

生成门根据当前输入、上一时刻的状态和偏置来指定输出多少个生成的脉冲。$\boldsymbol{x}_t$表示输入句子的向量。总体而言, 3 个门控装置之间的连接与控制取决于公式(2)。

$$\left.\begin{aligned}
\boldsymbol{r}_t &= \rho(\boldsymbol{W}_r\cdot\boldsymbol{x}_t+\boldsymbol{U}_r\cdot\boldsymbol{u}_{t-1}+b_r)\\
\boldsymbol{c}_t &= \rho(\boldsymbol{W}_c\cdot\boldsymbol{x}_t+\boldsymbol{U}_c\cdot\boldsymbol{u}_{t-1}+b_c)\\
\boldsymbol{o}_t &= \rho(\boldsymbol{W}_o\cdot\boldsymbol{x}_t+\boldsymbol{U}_o\cdot\boldsymbol{u}_{t-1}+b_o)
\end{aligned}\right\}\quad(2)$$

神经元σ产生峰值, 为

$$\boldsymbol{\alpha}_t = f(\boldsymbol{W}_\alpha\cdot\boldsymbol{x}_t+\boldsymbol{U}_\alpha\cdot\boldsymbol{u}_{t-1}+b_\alpha)\quad(3)$$

根据 3 个非线性门和产生的脉冲信号, 计算神经元σ在t时刻的状态和输出为:

$$\left.\begin{aligned}
\boldsymbol{u}_t &= \boldsymbol{r}_t\odot\boldsymbol{u}_{t-1}-\boldsymbol{c}_t\odot\boldsymbol{\alpha}(t)\\
\boldsymbol{h}_t &= \boldsymbol{o}_t\odot\boldsymbol{\alpha}_t
\end{aligned}\right\}\quad(4)$$

其中, $\boldsymbol{h}_t$表示 LSTM-SNP 层的输出, 即 NLP 任务的上下文隐层向量。图 1、图 2 以及 LSTM 和 LSTM-SNP 的相关公式均在神经元细胞的水平上进行描述。当数百个神经元连接起来执行计算时, 一个真正的仿生神经网络就形成了。

2.2 LSTM-SNP-CRF 模型

Lafferty 等[16] 在 2001 年提出条件随机场(CRFs)。CRF 是统计关系学习的重要框架, 具有较强的描述、逻辑推理, 以及对不确定性的处理能力。作为

典型的判别模型经常被构造为 NER 或其他 NLP 学习模型的增强组件。本节阐述 LSTM-SNP 和 CRF 组件的兼容性, 目的在于研究 CRF 组件能否在 NER 任务中提高 LSTM-SNP 的识别准确性。

本文选择 LSTM 和 BiLSTM 这 2 个模型作为 LSTM-SNP 模型的参照实验组, 通过实体识别精度来评估 3 个模型对 CRF 层的性能提高敏感性。模型的处理流程为首先将文本经过词嵌入转换为特征向量, 然后分别送入 LSTM-SNP、LSTM、BiLSTM 这 3 个模型获取词语之间的关系特征, 最后将其输入到 CRF 处理层, 获得标签的分值, 最大分值对应的标签即为模型认定的标签。模型整体处理流程如图 3 所示。LSTM-SNP 层的功能与 LSTM 层和 BiLSTM 层的功能相同。这 3 层都是用来提取句子的特征。LSTM-SNP 层将依次被 LSTM 或 BiLSTM 层取代, 用于 CRF 敏感性比较。这些模型使用了 BIO(begin, inside, outside)标签方案。

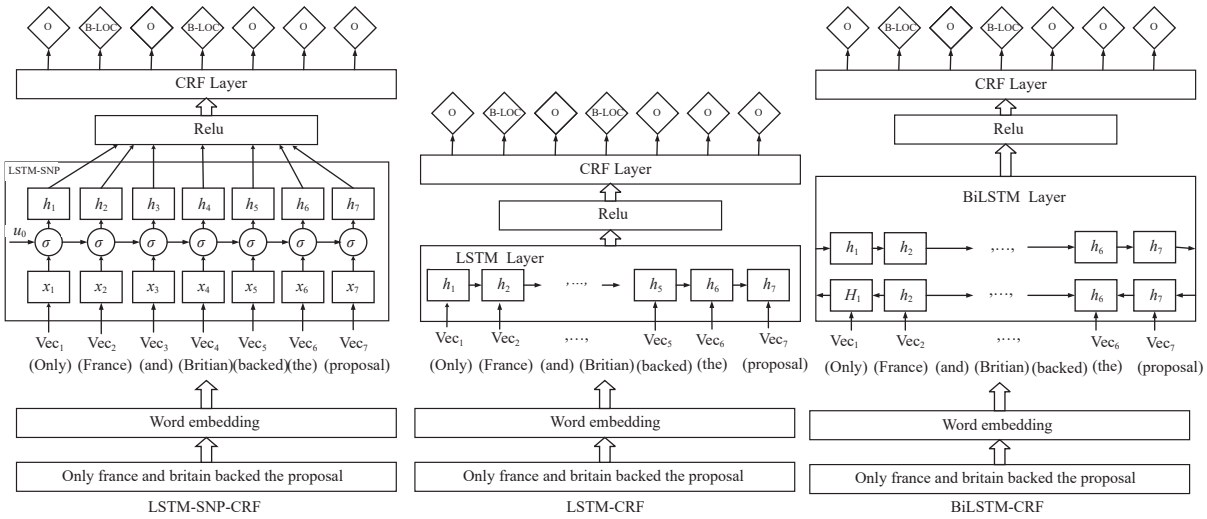

图 3　LSTM-SNP-CRF 模型、LSTM-CRF 模型、BiLSTM-CRF 模型的结构

Fig. 3　Structure of the LSTM-SNP-CRF model, the LSTM-CRF model, and the BiLSTM-CRF model

2.3　GloVe-CNN- LSTM-SNP 模型

本节提出带有 GloVe[17] 和卷积神经网络(CNN)[18] 的 LSTM-SNP 的整体体系架构, 如图 4 所示。LSTM-SNP、LSTM-SNP-CRF 这 2 种模型中的嵌入表示仅采用了独热编码连接层。这种嵌入方法会造成编码稀疏、维度大、词间相似性反应能力弱等问题。因此, LSTM-SNP 在命名实体识别任务的有效性有待进一步的实验证明。区别于传统 LSTM-SNP, 本文采用了更高效的特征提取方法, 具体分为基于词级别的特征提取和字符级别的特征提取。词级别的特征通过词嵌入方法 GloVe 以及手动定义词大小写特征的方法分别获取语言特征和词大小写信息。基于字符级别的特征提取是通过卷积神经网络 CNN 以获得词更加细粒度的特征表示。CNN 模型提取单词的字符级特征的过程如图 5 所示。模型将 GloVe 向量、CNN 向量和单词大小写信息向量通过拼接操作相结合, 并通过 LSTM-SNP 层进行处理。同时, 为了验证 LSTM-SNP 模型在实际应用中的优越性, 本文将 LSTM-SNP 层分别替换为 LSTM 层和 BiLSTM 层用于性能比对。下

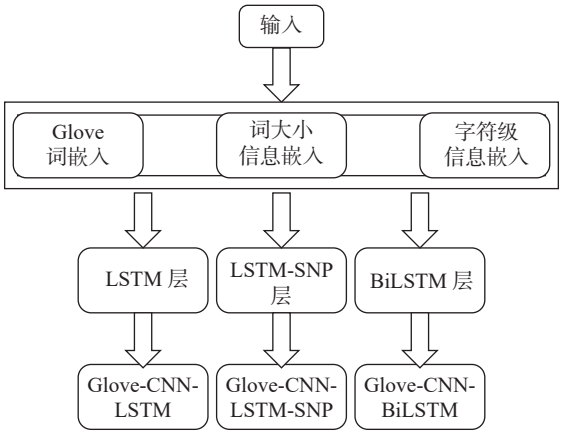

图 4　GloVe-CNN- LSTM-SNP 模型、GloVe-CNN- LSTM 模型、GloVe-CNN-BiLSTM 模型的结构

Fig. 4　Structure of the GloVe-CNN-LSTM-SNP model, the GloVe-CNN-LSTM model, and the GloVe-CNN-BiLSTM model

面将基于词级别和字符级别介绍各项特征提取技术。

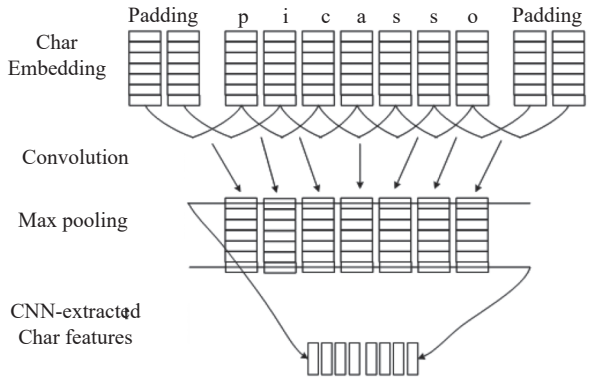

图 5 CNN 模型提取单词的字符级特征的过程

Fig. 5 The process of extracting character level features of words from CNN model

2.3.1 基于词级别的语义特征提取

近年来，一些工具，如 word2vec 和 GloVe，已被广泛应用于命名实体识别(NER)。GloVe 是一种用于获取词的向量表示的无监督学习算法。简而言之，GloVe 允许获取文本语料库，并将该语料库中的每个单词直观地转换为高维空间位置。这意味着相似的词将被放在一起，而这一技术也是词嵌入技术的重要组成部分。本文受到 Chiu 等[18]的启发，提出了一种基于预训练的字符嵌入方法，将来自维基百科和网络文本的 60 亿个单词作为训练资料，设计了一组基于 GloVe embeddings3[17]的对比实验。

2.3.2 基于词级别的大小写信息特征提取

因为在使用 GloVe 词嵌入方法时会丢失大量的字母大写信息，所以本文借鉴 Collobert[18]的方法获取词嵌入过程缺少的信息。该方法使用一个单独的查找表来添加大写选项：全为大写、初始大写、初始小写、大小写混合、其他。

本文的 GloVe-CNN-LSTM-SNP 模型应用了 Collobert 等[18]的方法以在单词嵌入期间获得词语大小写信息，同时将该查找表选项扩展。选项包括：所有字母全小写、所有字母全大写、仅首字母大写、全为数字、多部分为数字、少部分数字(包含数字)、其他、填充标记这 8 个选项。将此选项表命名为查找表 C 中，用于做基于词级别的单词大小写信息嵌入。

2.3.3 基于字符级别的特征提取

CNNs[18]是当前深度学习技术中最具有代表性的一种神经网络结构，近年来受到了众多学科的广泛关注。实验设置通过采用 CNN 技术，从英文文本资料中提取指定实体的字符级特征。

英语中的单词通常由细粒度的字母构成，CNN 技术被用于处理这些字母。这些字母包含了诸如前缀/后缀等隐藏特征。对于不同类型的字符，实验设置了不同的随机字符向量，以区分字符和字符类型(字母、数字、标点符号、特殊字符等)。例如，大写字母‘A’和小写字母‘a’对应于 2 组不同的字符向量集。图 5 展示了 CNN 从一个单词中提取字符级特征的过程。

结合词级别和字符级别的特征表示，并将 2 种级别的特征表示向量进行拼接，得到完整的单词嵌入表示。该词嵌入表示包括了词的语言相关特征、词语的字符特征、词的大小写信息。图 6 展示了在 GloVe 和 LSTM-SNP 基础上加入 CNN 模块后的整体模型，即完整的 GloVe-CNN-LSTM-SNP 模型。LSTM-SNP 层的功能与 LSTM 层和 BiLSTM 层的功能相同。这 3 层都是用来提取句子的特征。LSTM-SNP 层将依次被 LSTM 或 BiLSTM 层取代，用于比较 3 种模型对于 CNN 的敏感程度。

3 实验分析

3.1 数据集

本研究优先采用 2 个经典的命名实体识别数据集 CoNLL-2003 和 OntoNotes5.0，对基于 CRF、基于 GloVe 和基于 CNN 的 LSTM-SNP 模型性能进行评估。所有的数据集都可以在网站公开获得。CoNLL-2003 数据集可以通过文献 [19] 网站下载。OntoNotes5.0 数据集可以通过文献 [20] 网站下载。关于数据集的训练测试和验证集的句子数量划分如表 1 所示。

3.2 评估标准

根据前期工作[21]，为正确评估 LSTM-SNP 在命名实体识别任务中使用 CRF、CNN 和 GloVe 时模型的有效性，本文选择了 NLP 领域的通用评估度量系统，即精度(P)、召回率(R)和准确率(Acc)。测试样本被分为实际的实体类别和预测的实体类别。实验结果分为 4 类，如表 2 所示。预测

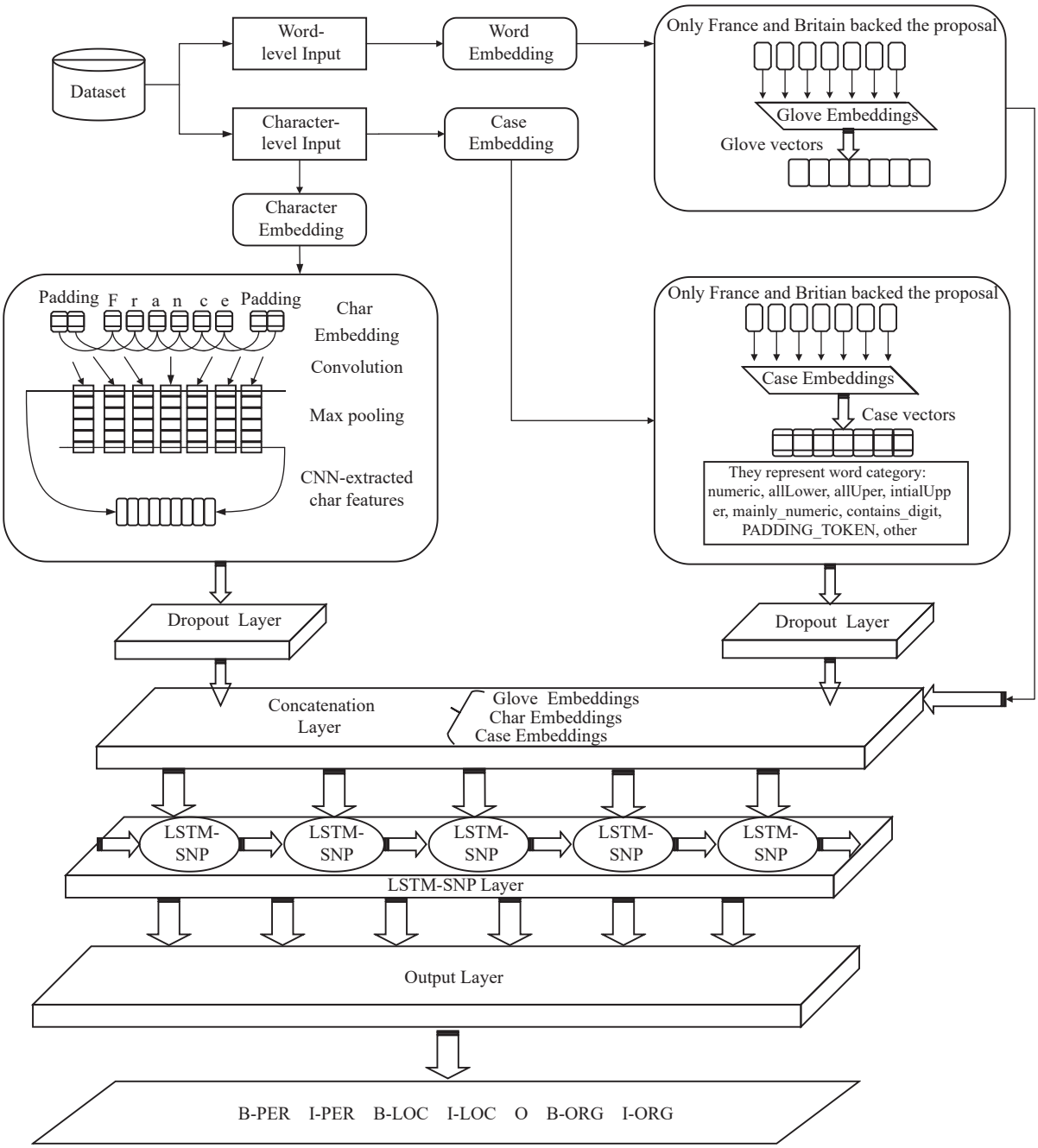

图 6　GloVe-CNN-LSTM-SNP 模型的架构

Fig. 6　Architecture of GloVe-CNN-LSTM-SNP model

表 1　语料库句子统计

Tab. 1　Corpus sentence statistics

数据集	全集	训练集	验证集	测试集
CoNLL-2003	20744	17291	—	3453
OntoNotes5.0	76714	59924	8528	8262

的实体代表由模型得出的实体标签,实际的实体代表人工标注的真实标签。

本文采用的精度(P)、召回率(R)和准确率(Acc)定义为:

表 2　混淆矩阵

Tab. 2　Confusion matrixs

Actual entities	Predicted entities	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{5}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{6}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\times 100\% \quad (7)$$

式中：TP(true positive)表示模型正确地将一个实体标记为正类，即模型正确地将一个实体标记为实体，并且这个实体与真实标签一致；FN(false negative)表示模型错误地将一个实体标记为负类，即模型没有将一个实体标记为实体，或者将实体标记为了错误类型；FP(false positive)表示模型错误地将一个非实体标记为正类，即模型将一个非实体错误地标记为了实体；TN(true negative)表示模型正确地将一个非实体标记为负类，即模型正确地将一个非实体标记为非实体。

命名实体识别任务涉及多种分类。因此，微平均F_1($F_{1\mathrm{macro}}$)值也被用作性能评估指标，定义为：

$$P_{\mathrm{macro}}=\mathrm{average}(P_1+P_2+\cdots+P_n) \quad (8)$$

$$R_{\mathrm{macro}}=\mathrm{average}(R_1+R_2+\cdots+R_n) \quad (9)$$

$$F_{1\mathrm{macro}}=\frac{2\times P_{\mathrm{macro}}\times R_{\mathrm{macro}}}{P_{\mathrm{macro}}+R_{\mathrm{macro}}} \quad (10)$$

式中：P_1, P_2,⋯,P_n分别代表第 1 种实体类别，第 2 种实体类别以及第 n 种实体类别的精度值；R_1, R_2,⋯,R_n分别代表第 1 种实体类别，第 2 种实体类别以及第 n 种实体类别的召回率。

3.3 参数配置

根据 LSTM-SNP、LSTM 和 BiLSTM 的模型结构，首先在实验中实现了这 3 种模型。当对比实验依次添加 CNN 和 GloVe 作为嵌入模型时，模型的内部参数保持不变。

LSTM-SNP 模型除了需要学习的权重矩阵和偏差向量外，还有一些通过实验确定的先验参数，包括迭代计数(iterations)、Dropout 率和神经元数量(neurons)。

图 7 展示了 LSTM-SNP 在 CoNLL-2003 数据集上经过不同迭代次数和提前停止次数训练时的性能差异，包括 15-4、15-5、15-6、15-10、20-5、20-10 和 100-50。由图可知，当迭代次数和提前次数为 15-5 时，LSTM-SNP 在 CoNLL-2003 数据集上取得了 72.5% 的$F_{1\mathrm{macro}}$值。因此不同迭代次数和提前次数采用 15-5。

本文还考察了不同的 Dropout 率对 LSTM-SNP 模型$F_{1\mathrm{macro}}$分数的影响，即当迭代次数与提前停止次数定为 15-5 时，Dropout 率分别为 0%、5%、10%、25%、50% 和 75%。图 8 所示的对比实验结果表明，当 Dropout 率为 50% 时，$F_{1\mathrm{macro}}$值为 72.8%，适合于模型训练。

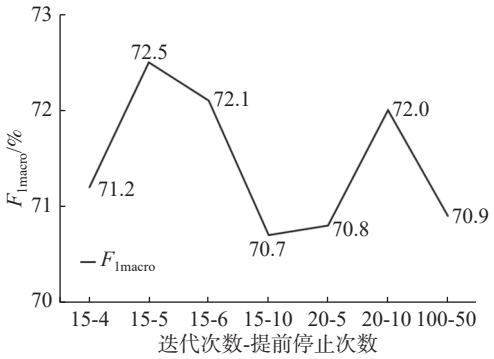

图 7 基于数据集 CoNLL-2003 上的不同的迭代次数与提前停止次数对 LSTM-SNP 模型$F_{1\mathrm{macro}}$分数的影响

Fig. 7 Influence of iterations and early-stops on $F_{1\mathrm{macro}}$ based on LSTM-SNP model and dataset CoNLL-2003

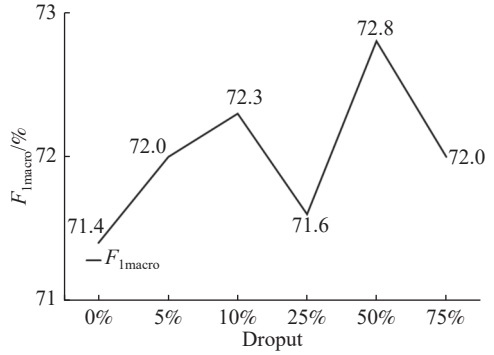

图 8 基于数据集 CoNLL-2003 上的不同 Dropout 率对 LSTM-SNP 模型$F_{1\mathrm{macro}}$分数的影响

Fig. 8 Influence of Dropout rates on $F_{1\mathrm{macro}}$ based on LSTM-SNP model and dataset CoNLL-2003

当迭代次数和提前停止次数为 15-5、Dropout 率为 50% 时，设置不同的神经元个数(32、64、128、256 和 512)进行实验。如图 9 所示的当前结果发现，神经元数量设为 256 有利于模型训练。

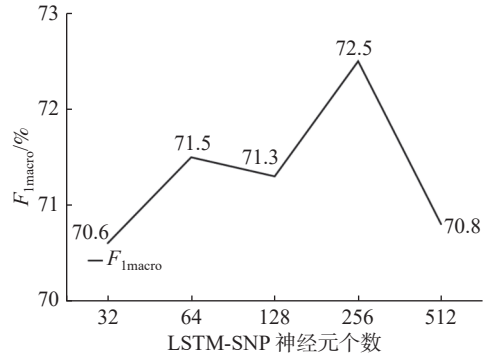

图 9 基于数据集 CoNLL-2003 上的不同的神经元个数率对 LSTM-SNP 模型$F_{1\mathrm{macro}}$分数的影响

Fig. 9 Influence of neurons on $F_{1\mathrm{macro}}$ based on LSTM-SNP model and dataset CoNLL-2003

3.4　对比实验结果分析

本研究的最初目标是确定 LSTM-SNP 模型对 CRF、GloVe 词嵌入和 CNN 等传统深度学习组件的适应性。在保持相同的超参数基础上,如表 3 所示,本文将 A、B、C、D 4 组模型在 2 个数据集上进行实验,获得以实体为单位的识别结果。模型得分情况如表 4 所示。其中:LSTM-SNP 在 2 个数据集上的表现与 LSTM 相似;所有模型在数据集 CoNLL-2003 的 F_{1macro}平均得分比数据集 OntoNotes5.0 高出 10 分左右。一个重要的原因是数据集 CoNLL-2003 和 OntoNotes5.0 中存在着不同的实体类别数量。前者分 4 类,后者分 18 类。这说明,LSTM-SNP 模型以及这类循环神经网络模型在对少量实体进行分类是具有更好的识别效果。本文从 2 个不同的视角对实验结果进行分析。

表 3　对比实验设置

Tab. 3　Contrast experiment settings

组别	本文模型	对比模型1	对比模型2
A组	LSTM-SNP	LSTM	BiLSTM
B组	LSTM-SNP-CRF	LSTM-CRF	BiLSTM-CRF
C组	GloVe-LSTM-SNP	GloVe-LSTM	GloVe-BiLSTM
D组	GloVe-CNN-LSTM-SNP	GloVe-CNN-LSTM	GloVe-CNN-BiLSTM

表 4　LSTM-SNP、LSTM、BiLSTM 在数据集 CoNLL-2003 和 OntoNotes5.0 的性能结果

Tab. 4　Performance results of LSTM-SNP, LSTM, and BiLSTM in the dataset CoNLL-2003 and OntoNotes5.0

组别	Models	CoNLL-2003			OntoNotes5.0		
		P	R	F_{1macro}	P	R	F_{1macro}
A	LSTM-SNP	***77.25***	***69.82***	***73.35***	***40.50***	***38.16***	***39.30***
	LSTM	76.45	68.66	72.35	40.05	38.77	39.40
	BiLSTM	83.19	72.01	77.20	61.29	54.24	57.55
B	LSTM-SNP-CRF	***82.20***	***70.94***	***76.16***	***65.16***	***55.51***	***59.95***
	LSTM-CRF	82.18	71.37	76.40	65.38	55.18	59.85
	BiLSTM-CRF	84.10	71.55	77.32	79.25	76.89	78.05
C	GloVe-LSTM-SNP	***83.25***	***72.22***	***77.34***	***79.25***	***69.00***	***73.77***
	GloVe-LSTM	82.32	80.55	81.41	79.25	76.89	78.05
	GloVe-BiLSTM	86.38	84.90	85.63	82.42	80.74	81.57
D	GloVe-CNN-LSTM-SNP	***76.72***	***79.65***	***78.12***	***74.42***	***75.85***	***75.12***
	GloVe-CNN-LSTM	81.55	81.55	81.55	78.55	78.74	78.64
	GloVe-CNN-BiLSTM	84.96	87.00	85.96	81.12	82.68	81.89

3.4.1　基于 A、B、C、D 组的实验分析

A 组的 LSTM-SNP 模型在数据集 CoNLL-2003 中的 F_{1macro}分数为 73.35%,在数据集 OntoNotes5.0 中为 39.30%。BiLSTM 的 F_{1macro}得分在 2 个数据集下都是最高的。一个未预料到的发现是,LSTM-SNP 模型和 LSTM 模型在 NER 任务上的 F_{1macro}分数方面没有显著差异。整体而言,在处理 4 种实体类型问题时,3 种模型的表现均优于 18 种实体类型的模型。

B 组添加 CRF 后,LSTM、BiLSTM 和 LSTM-SNP 3 种模型的性能均得到改善。在数据集 CoNLL-2003 和 OntoNotes5.0 中,LSTM-SNP 模型在 2 个数据集的 F_{1macro}得分分别为 76.16% 和 59.95%。添加了 CRF 的 BiLSTM 在 2 个数据集中性能仍然是最好的。值得一提的是,LSTM-SNP-CRF 在数据集 OntoNotes5.0 拥有更明显的性能改进。与 A 组相比,F_{1macro}约提高了 20%。

C 组的得分体现了 GloVe 词嵌入对模型的贡献。实验结果显示,在使用 GloVe 词嵌入时,3 个模型的学习效果均比 A 组和 B 组更优。在 2 个数据集上,模型 GloVe-LSTM-SNP 的 F_{1macro}分别比 A 组的 LSTM-SNP 高 4% 和 34.47%。但是,从 2 个数据集的 F_{1macro}来看,GloVe-LSTM-SNP 的得分在 2 个数据集上最低,分别为 77.34% 和 73.77%。

D 组与 C 组相比,3 个模型均无显著差异。与 C 组 3 个没有添加 CNN 组件的模型相比,D 组的每个模型性能有轻微的提高。在数据集 OntoNotes5.0 中,GloVe-CNN-GloVe 相比 A 组的 F_{1macro}提高 35.82%。显然,D 组的 3 个模型在 2 个数据集上都取得了最高得分。

3.4.2　3 种模型的消融实验分析

将 CRF 添加到 LSTM-SNP、LSTM、BiLSTM 3 个模型时,各模型在数据集 CoNLL-2003 上的

F_{1macro}分别提高了 2.81%、4.05% 和 0.12%。3 个模型在数据集 OntoNotes5.0 上的F_{1macro}分别提高 20.65%、20.45% 和 20.5%。实验结果表明,LSTM 和 LSTM-SNP 模型的性能提升程度在数据集 CoNLL-2003 上比 BiLSTM 模型更大,在数据集 OntoNotes5.0 上提升程度相似。

需要指出的是,数据集 CoNLL-2003 的数据量相对于数据集 OntoNotes5.0 来说较小。数据集 CoNLL-2003 大约是数据集 OntoNotes5.0 的 3.5 倍。因此,模型 LSTM-SNP、LSTM 与 BiLSTM 相比,LSTM-SNP 和 LSTM 对数据集的质量和数量更为敏感。

在 LSTM-SNP、LSTM、BiLSTM 这 3 个模型中添加 GloVe 词嵌入,当处理数据集 CoNLL-2003 时,F_{1macro}值相比于 A 组分别增加了 3.99%、9.06% 和 8.43%。将 GloVe 词嵌入应用到 LSTM-SNP、LSTM 和 BiLSTM 模型中,在 CoNLL-2003 数据集上与 A 组比较,F_{1macro}值分别提高了 3.99%、9.06% 和 8.43%。此外,在 OntoNotes5.0 数据集下,与原始模型 A 组相比,分别提高了 34.47%、38.65% 和 24.02%。实验结果表明:F_{1macro}改善非常显著,GloVe 词嵌入对模型的性能有很大的改善。

然而,随着数据集数量的增加,BiLSTM 模型的F_{1macro}得分提高幅度不如其他 2 个模型。这一现象在某种程度上证明了 BiLSTM 受数据量影响相对较小,而其他 2 种模型受数据量影响较大。对于 LSTM-SNP 模型,通过添加预处理方法,如词嵌入,该模型能够表现得更好。

在 LSTM-SNP、LSTM、BiLSTM 中同时加入 CNN 和 GloVe 进行处理时,在 CoNLL-2003 数据集下,3 个模型的性能均比 A 组的基线模型提高了 4.77%、9.20% 和 8.76%。在数据集 OntoNotes5.0 下,F_{1macro}分别提高 35.82%、39.24% 和 24.34%。这些数据共同证明了添加 CNN 和词嵌入的有效性。类似地,LSTM-SNP 和 LSTM 都有较大的增加,也说明了上述结论。此外,通过添加 CNN 从单词嵌入中学习字符级特征,3 个模型在 2 个数据集上的性能提升很小。结果显示,基于文字嵌入的特征方法可以有效地改善 LSTM-SNP 模型的性能,并具有较好的效果。

4 总结与展望

本文旨在评价 LSTM-SNP 模型在序列问题(命名实体识别)应用中的有效性。同时,为了探索 LSTM-SNP 模型是否具有在自然语言处理领域的研究潜力,本文在 LSTM-SNP 模型以及其对比模型 LSTM 和 BiLSTM 中有序添加了一些深度学习组件,包括 CRF、单词嵌入等,以对比组件对不同模型的性能提升幅度,从而为 LSTN-SNP 模型的未来研究提供可靠数据参考。

实验表明,传统 LSTM-SNP 模型在命名实体任务中的性能与 LSTM 模型基本相似,但与 BiLSTM 的良好性能仍存在一定的差距。此外,实验发现,LSTM-SNP 模型受数据集领域知识的影响较大。在 LSTM-SNP 模型中加入 CRF、词嵌入和 CNN,该模型的性能有了显著的提高。加入词嵌入、CNN 等特征预处理模块可以极大地改善模型的总体性能。总体而言,LSTM-SNP 模型在命名实体识别任务中具有潜力,并且有比较大的改进空间。

未来的工作将考虑使用 LSTM-SNP 模型提取实体局部特征。本文仅考虑了实体上下文特征,其粒度不够细腻。因此,将注意机制引入到 LSTM-SNP 模型中,利用注意机制来提取局部特征[22],从而实现对命名实体识别有较大影响权重特征的重点关注。同时,考虑实现多层或双向的 LSTM-SNP 模型,以提高模型提取特征的能力。

参 考 文 献

[1] PĂUN G. Computing with membranes[J]. Journal of Computer and System Sciences, 2000, 61(1): 108 − 143.

[2] 张葛祥, 潘林强. 自然计算的新分支—膜计算[J]. 计算机学报, 2010, 33(2): 208 − 204.

ZHANG G X, PAN L Q. A new branch of natural computing-membrane computing[J]. Journal of Computer Science, 2010, 33(2): 208 − 204.

[3] 黄亮. 膜计算优化方法研究 [D]. 杭州: 浙江大学, 2007.

HUANG L. Study on optimization method of membrane calculation [D]. Hangzhou: Zhejiang University, 2007.

[4] 潘林强, 张兴义, 曾湘祥, 等. 脉冲神经膜计算系统的研究进展及展望 (英文)[J]. 计算机学报, 2008, 31(12): 2090 − 2096.

PAN L Q, ZHANG X Y, ZENG X X, et al. Research progress and prospect of pulse neural membrane computing system[J]. Journal of Computer Science, 2008, 31(12): 2090 − 2096.

[5] PENG H, LV Z, LI B, et al. Nonlinear spiking neural P systems[J]. International Journal of Neural Systems, 2020, 30(10): 2050008.

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 − 1780.

[7] LIU Q, LONG L, YANG Q, et al. LSTM-SNP: A long short-term memory model inspired from spiking neural P systems[J]. Knowledge-Based Systems, 2022, 235: 107656.

[8] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the web: An experimental study[J]. Artificial intelligence, 2005, 165(1): 91 − 134.

[9] WANG Z, LI J, WANG Z, et al. XLore: A Large-scale English-Chinese bilingual knowledge graph[C]/Proceedings of the 12th International Semantic Web Conference (Posters & Demonstrations Track) .[S.l.]: CEUR-WS org, 2013: 121−124.

[10] 康怡琳, 孙璐冰, 朱容波, 等. 深度学习中文命名实体识别研究综述[J]. 华中科技大学学报 (自然科学版), 2022, 50(11): 44 − 53.

KANG Y L, SUN L B, ZHU R B, et al. Overview of the research on Chinese named entity recognition for in-depth learning[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2022, 50(11): 44 − 53.

[11] MAKINO T, OHTA Y, TSUJII J. Tuning support vector machines for biomedical named entity recognition[C]//Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain.Pennsylvania, USA: Association for Computational Linguistics, 2002: 1− 8.

[12] KRISHNAN V, MANNING C D. An effective two-stage model for exploiting non-local dependencies in named entity recognition[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.Pennsylvania, USA: Association for Computational Linguistics, 2006: 1121 − 1128.

[13] LUO L, YANG Z, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381 − 1388.

[14] LI L, GUO Y. Biomedical named entity recognition with CNN-BLSTM-CRF[J]. Journal of Chinese Information Processing, 2018, 32(1): 116 − 122.

[15] LI J Y, FEI H, LIU J A, et al. Unified named entity recognition as word-word relation classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10965 − 10973.

[16] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning.San Francisco CA, USA: Morgan Kaufmann Publishers Inc, 2001: 282 − 289.

[17] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1532 − 1543.

[18] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357 − 370.

[19] DeepAI. CoNLL 2003 (English) dataset download[EB/OL]. (2003-11-12)[2022-11-26].https://deepai.org/dataset/conll-2003-english.

[20] Linguistic Data Consortium. OntoNotes Release 5.0Download. [EB/OL].(2013-10-16)[2022-11-26].https://www.ldc.upenn.edu/.

[21] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012.

LI H. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012.

[22] 李明扬, 孔芳. 融入自注意力机制的社交媒体命名实体识别[J]. 清华大学学报 (自然科学版), 2019, 59(6): 461 − 467.

LI M Y, KONG F. Social media named entity recognition with self attention mechanism[J]. Journal of Tsinghua University (Natural Science Edition), 2019, 59(6): 461 − 467.

(编校: 饶莉)