

软注意力掩码嵌入下中文命名 实体识别算法

王秀慧¹, 徐永波²

(1. 山西大同大学 计算机与网络工程学院, 山西 大同 037009; 2. 河南大学 人工智能学院, 郑州 450046)

摘要: 中文词汇的语义存在一定的模糊性, 在中文文本中, 存在一些与命名实体识别相关性较低的特征, 同一个词汇在不同语境中具有不同的含义, 不同的词汇和短语对命名实体的识别具有不同的贡献度, 若不进行加权或掩码操作, 这些特征则会干扰模型的识别准确率。为此, 本文提出一种软注意力掩码嵌入的中文命名实体识别 (CNER) 算法。首先, 建立多层次 CNER 模型, 在模型的词向量表示层, 借助 jieba 技术对输入层传递过来的中文文本进行分词处理, 并利用 Word2Vec 方法获取各词汇的词向量, 形成词向量序列。其次, 在 BiLSTM 层对词向量序列进行双向长短期记忆处理, 得到每个词向量对应的融合了前后文信息的特征向量。再次, 在 BiLSTM 层后嵌入一个软注意力掩码模块, 利用该模块的软注意力机制对 BiLSTM 层输出的特征向量进行加权和掩码操作, 关注对实体识别有重要贡献的特征, 去除和抑制不重要的特征, 提高识别的精度。最后, 在条件随机场 (CRF) 层对经过软注意力掩码模块处理后的特征向量进行标签标注与解码, 从而得到最佳实体标签序列, 该序列即为中文命名实体识别结果。实验结果表明, 该算法可以精准识别中文命名实体, 在实体标签标注覆盖性和 F_1 值方面均有较好的表现。

关键词: 中文命名; 软注意力; 实体识别; 掩码操作; Word2Vec; BiLSTM 模型

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1671-5497(2026)01-0231-08

DOI: 10.13229/j.cnki.jdxbgxb.20241174

Chinese named entity recognition algorithm with soft attention mask embedding

WANG Xiu-hui¹, XU Yong-bo²

(1. School of Computer and Network Engineering, Shanxi Datong University, Datong 037009, China; 2. College of Artificial Intelligence, Henan University, Zhengzhou 450046, China)

Abstract: The semantics of Chinese vocabulary have a certain degree of ambiguity. In Chinese text, there are some features that have low relevance to named entity recognition. The same vocabulary has different meanings in different contexts, and different vocabulary and phrases have different contributions to named entity recognition. If weighting or masking operations are not performed, these features will interfere with the recognition accuracy of the model. To this end, a Chinese named entity recognition (CNER) algorithm

收稿日期: 2024-10-10.

基金项目: 国家自然科学基金项目 (12301138).

作者简介: 王秀慧 (1981-), 女, 副教授, 硕士. 研究方向: 机器学习. E-mail: zenlw@163.com

with soft attention mask embedding is studied. Establish a multi-level CNER model, in the word vector representation layer of the model, use jieba technology to perform segmentation processing on the Chinese text passed from the input layer, and use Word2Vec method to obtain the word vectors of each vocabulary, forming a sequence of word vectors. In the BiLSTM layer, bidirectional long short-term memory processing is applied to the sequence of word vectors to obtain feature vectors that fuse contextual information for each word vector. Embedding a soft attention mask module after the BiLSTM layer, using the soft attention mechanism of this module to perform weighted and masked operations on the feature vectors output by the BiLSTM layer, focusing on features that contribute significantly to entity recognition, removing and suppressing unimportant features, and improving recognition accuracy. Label and decode the feature vectors processed by the soft attention mask module in the CRF layer to obtain the optimal entity label sequence, which is the Chinese named entity recognition result. The experiment shows that the algorithm can accurately recognize Chinese named entities, and has good performance in entity label annotation coverage and F1 value.

Key words: Chinese naming; soft attention; entity recognition; mask operation; Word2Vec; BiLSTM model

0 引言

中文命名实体识别(Chinese named entity recognition, CNER)是自然语言处理(Natural language processing, NLP)领域的核心技术之一,它能够从中文文本中准确识别出人名、地域名、组织名等关键实体^[1,2]。对信息提取、知识关系挖掘等方面有重要意义,同时也是实现智能对话系统、搜索引擎优化等应用的基础^[3-5],对推动NLP技术的发展和應用有较强的必要性。但由于中文词汇的语义具有模糊性,同一个词汇在不同的上下文具有不同的含义,中文命名实体的识别依赖于上下文信息,增加了识别的复杂性。为此,需要研究一种有效的中文命名实体识别算法。

关于命名实体识别问题,诸多学者进行了大量研究,李健等^[6]通过Transformer模型对中文文本序列进行特征抽取,并将得到的特征序列作为输入信息传递给隐马尔科夫模型进行标记预测,进而实现CNER自动识别。但该方法在处理长文本时,由于Transformer模型受计算效率和存储空间限制,导致隐马尔科夫模型在处理其传递的特征时会出现信息损失的情况,影响最终的识别效果。Jeon等^[7]结合建筑结构词库中的先验知识,通过TrAdaBoost算法对构建好的命名实体识别(Named entity recognition, NER)模型进行微调,使其能够适应噪声文本的特点,从而准确识别

出与建筑结构相关的命名实体。尽管TrAdaBoost算法可以提高NER模型的性能,但其却过度依赖建筑结构词库的先验知识,会导致NER模型在处理新型命名实体时泛化能力不足,从而降低识别准确率。方红等^[8]依据专有名词词典,并结合句法依存树捕捉词汇间的句法依赖结构,构建文本信息图结构,将其作为输入信息传递至图神经网络模型中,利用该模型实现命名实体识别。该方法中专有名词词典虽然提供了一定的命名实体识别线索,但是其覆盖范围和更新速度无法满足实际应用的需求,从而导致识别结果产生偏差。廖梦等^[9]利用双向编码表征模型(Bidirectional encoder representations from transformers, BERT)获取中文文本字符嵌入,通过Transformer解码器使字符与标签向量交互学习,增强字符特征,并引入多任务学习模式,对建立的循环神经网络(Recurrent neural network, RNN)模型进行优化训练,利用该模型实现CNER的识别。该方法虽对RNN模型进行了优化训练,但其本身在处理长序列文本时容易出现梯度消失的问题,从而影响识别精度。

软注意力是深度学习注意力机制的关键技术,通过动态调整输入信息的权重,使模型在处理数据时能够聚焦于重要信息。掩码则是一种用于数据筛选和过滤的技术手段,可以应用于不同的数据形式,如字符、数字或向量等,通过设定特定

的规则或模式,将不需要的信息进行屏蔽或忽略。在上述研究的基础之上,本文研究软注意力掩码嵌入下中文命名实体识别算法,旨在提高算法对 CNER 的识别能力,增强其对复杂语言结构和语义关系的理解能力,为 NLP 等相关领域提供技术支持。

1 中文命名实体识别

1.1 构建中文命名实体识别模型

本文设计的 CNER 模型如图 1 所示。

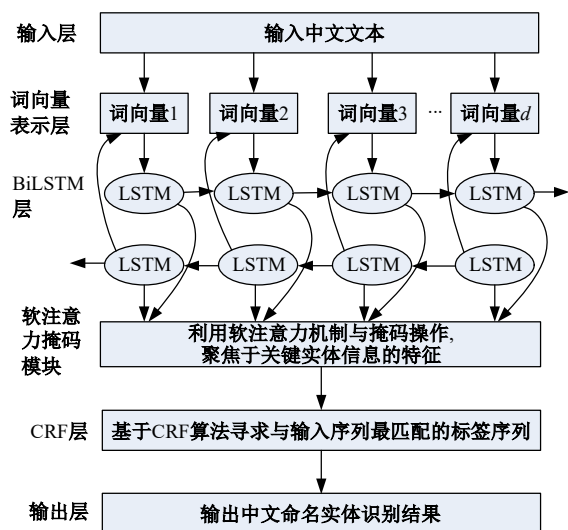


图 1 CNER 模型设计

Fig. 1 CNER model design

由图 1 可知,本文设计的 CNER 模型具有多个层次,其中,利用词向量表示层对输入层传递过来的中文文本进行分词并获取其词向量,形成词向量序列;利用双向长短期记忆(Bi-directional long short-term memory, BiLSTM)层对获取的词向量序列进行双向长短期记忆处理,得到每个时间步的隐藏状态序列(即中文词向量序列中每个词向量对应的融合了前后文信息的特征向量);在 BiLSTM 层后嵌入一个软注意力掩码模块,利用该模块的软注意力机制对 BiLSTM 层的输出进行加权处理,以强调对命名实体识别有重要贡献的特征,通过掩码操作抑制不重要的特征,该模块输出的是经过软注意力机制加权和掩码处理后的中文词特征向量序列^[10,11];最后利用条件随机场(Conditional random field, CRF)层执行标签标注与解码操作,在得到的所有实体标签序列集合中,找到一个与输入序列最匹配的标签序列,由此得出最终的命名实体识别结果。

1.2 CNER 模型各层的实现

1.2.1 基于词向量表示层获取词向量序列

词向量是将词汇映射到一个高维连续空间中的向量表示,这种表示方式能够捕捉词汇之间的语义关系^[12]。在中文命名实体识别中,通过词向量表示,可以将中文词汇的语义特征有效地提取出来,为后续 CNER 任务提供丰富的信息。

在 CNER 模型的词向量表示层中,首先借助 jieba 技术,将连续的中文文本切割成独立的词汇单元,形成一系列的词汇序列。其次,利用 Word2Vec 方法对这些词汇序列进行词向量表示。

本文选择 Word2vec 方法中生成词向量的神经网络(Continuous bag of words, CBOW)模型对分词后的词汇序列进行训练学习,以此生成中文文本的词向量。CBOW 模型主要包括输入、投影以及输出 3 个层次,基本结构如图 2 所示。

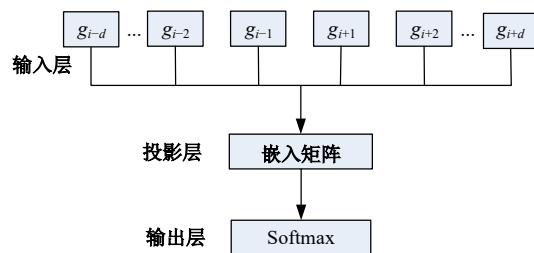


图 2 CBOW 模型结构图

Fig. 2 Structure diagram of CBOW model

为了符合 CBOW 模型输入形式,需将待输入的分词后的中文词汇进行 one-hot 编码。图 2 中, g_{i-d} 、 g_{i-2} 、 g_{i-1} 、 g_{i+1} 、 g_{i+2} 和 g_{i+d} 分别代表待输入词汇对应的 one-hot 编码向量, d 为编码向量数量。通过输入层接收 one-hot 编码向量,当输入窗口大小设定为 7 时,目标词前后的各 3 个词汇的 one-hot 编码向量,会被作为前后文输入 CBOW 模型中。

one-hot 编码向量传递到投影层后,通过与该层的嵌入矩阵执行乘法操作,会被转换成 1 个密集向量,也就是词向量。这些词向量经过聚合操作后,会形成一个单一的上下文向量表示,作为投影层的输出。

在输出层,利用 Softmax 函数将接收到词向量表示转换为一个概率分布,该分布对应于词汇表中每个词汇成为预测目标词的概率。

CBOW 模型在训练过程中,使用反向传播算法更新相关参数(包括嵌入矩阵和输出层的权

重),并以最大化对数似然函数 ϕ 为学习目标, ϕ 描述为:

$$\phi = \log p(x_i | x_{i-d}, \dots, x_{i-2}, x_{i-1}, \dots, x_{i+1}, x_{i+2}, \dots, x_{i+d}) \quad (1)$$

式中: p 为概率密度函数; x_i 为模型获取的词向量。 λ 为给定参数; R 为参数空间,当模型最终收敛时,嵌入矩阵中的参数就被固定下来。从训练好的CBOW模型中访问并保存嵌入矩阵,该矩阵中,每行均对应一个中文词汇的词向量序列。

1.2.2 基于BiLSTM层提取特征向量

在该层中,利用BiLSTM模型捕捉目标中文词向量序列的前后文信息,为后续的命名实体识别任务提供更加准确和全面的特征表示。

BiLSTM模型由两个长短期记忆网络(Long short-term memory, LSTM)单元组成,它们呈相互相反的方向存在。LSTM单元内部包含与之相关的遗忘、输入以及输出3个门控机制^[13]。这些门控机制使用Sigmoid函数控制信息流动的程度,从而实现对长期依赖关系的有效建模和信息的选择性传递^[14]。其中,单一LSTM单元相关计算公式如下:

$$s_i = \delta(\omega_s[h_{i-1}, x_i] + b_s) \quad (2)$$

$$u_i = \delta(\omega_u[h_{i-1}, x_i] + b_u) \quad (3)$$

$$c_i = \delta(\omega_c[h_{i-1}, x_i] + b_c) \quad (4)$$

$$\bar{z}_i = \tanh(\omega_z[h_{i-1}, x_i] + b_z) \quad (5)$$

$$z_i = u_i \times z_{i-1} + s_i \times \bar{z}_i \quad (6)$$

$$h_i = \phi c_i \times \tanh(z_i) \quad (7)$$

式中: u_i 为遗忘门; δ 为Sigmoid函数; s_i 为输入门; ω 为权重矩阵; c_i 为输出门; b 为偏置向量; z 和 $\bar{z}_i$ 分别为记忆单元当前状态及临时状态; h_{i-1} 和 h_i 为上一个和当前中文词向量序列的隐藏状态(中文词特征向量)。

由于BiLSTM模型由两个方向相反的LSTM单元组成,所以当中文词向量序列被输入到BiLSTM模型中时,其中的每个词向量都会经过双向的LSTM单元的处理,由此得到两个隐藏状态 $\vec{h}_i$ 和 $\tilde{h}_i$,将二者结合在一起,便可以得到当前中文词向量对应的融合了前后文信息的特征向量 h'_i ,用公式描述为:

$$h'_i = \vec{h}_i \oplus \tilde{h}_i \quad (8)$$

经过上述操作便可得到输入的当前中文词向量序列对应的特征向量序列 $H = (h'_1, h'_2, \dots, h'_m)$, m 为特征向量数量。

1.2.3 基于软注意力掩码模块的加权和掩码处理

在深度学习领域中,软注意力机制凭借其可嵌入神经网络的可微性、考虑全局信息的全局性以及可动态调整权重的灵活性,在多个领域均有广泛的应用。在CNER任务中,软注意力机制可以帮助CNER模型更好地聚焦中文文本中的关键实体信息;而掩码则是一种将特定信息嵌入模型中的方法,通过掩码操作可以限制对非关键信息的关注。为此,本文将软注意力掩码模块嵌入CNER模型,将其放置在BiLSTM层之后,以此进一步增强模型对中文关键实体信息的捕捉能力,同时减少非关键信息的干扰,提高中文命名实体识别的准确性和效率。以下是对软注意力掩码模块的具体描述:

(1) 计算软注意力权重 w_i

将由BiLSTM层输出的特征向量 h'_i 作为软注意力特征,并对这些软注意力特征进行非线性变换,将其映射到一个新的向量空间,由此得到软注意力权重值 w_i ,用公式描述为:

$$w_i = z_i \tanh(h'_i) \quad (9)$$

(2) 计算软注意力权重概率向量 $p_i(w_i)$

为了计算每个 w_i 的相对重要性,即软注意力权重概率向量 $p_i(w_i)$,以确保所有权重之和为1并反映每个中文词向量对应特征向量的相对重要性,这一步可通过Softmax函数对 w_i 实施归一化处理实现,描述为:

$$p_i(w_i) = \frac{\exp(w_i)}{\sum_{i=1}^m \exp(w_i)} \quad (10)$$

(3) 获取加权特征向量 η_i

使用软注意力权重概率向量 $p_i(w_i)$ 加权特征向量 h'_i ,从而得到特征向量的软注意力权重配置,即加权后的特征向量 η_i ,描述为:

$$\eta_i = \sum_{i=1}^m p_i(w_i) h'_i \quad (11)$$

(4) 构建联合特征向量序列 $\hat{H}$

获取的每个 η_i 都是一个融合上下文特征和重要程度的联合特征向量。将这些向量按顺序排列,便可得到一个包含整个中文文本序列特征的联合特征向量序列 $\hat{H}$ 。

(5) 掩码操作

为了进一步限制对非关键信息的关注,对联

合特征向量序列 $\hat{H}$ 实施掩码操作,这一环节通常需要引入一个掩码矩阵实现。掩码矩阵可以根据任务需求进行设计。对CNER任务而言,可以基于已知的中文命名实体位置信息设置掩码矩阵 O ,矩阵中的元素 o_{ij} 表示第 i 个中文词的第 j 个加权特征是否应该被掩码(0表示掩码,1表示不掩码)。对于掩码矩阵中为0的位置,可以将对应的加权后的特征向量置零或进行其他形式的抑制。对于掩码矩阵中为1的位置,特征向量保持不变。得出的执行掩码处理后的特征向量序列 $\dot{H}$,用公式描述为:

$$\dot{H} = \hat{H} \odot O \eta_i \quad (12)$$

通过以上步骤,得到的特征向量序列 $\dot{H} = (h'_1, h'_2, \dots, h'_v)$ (v 表示该序列中特征向量数量),不仅包含中文文本中每个词的前后文信息,而且通过软注意力机制对所有的特征向量执行了加权操作,使关键命名实体的特征向量得到更大的权重,从而更加突出。此外,掩码操作进一步限制了模型对非关键信息的关注,使模型能够更加专注于识别关键命名实体。因此,经过软注意力机制加权和掩码操作的特征向量序列,为后续的命名实体识别任务提供了更准确和有效的输入,有助于提高CNER模型的识别性能和效率。

1.2.4 基于CRF层的中文命名实体识别结果

在软注意力掩码模块后,设置CRF层进一步优化和解析特征向量序列 $\dot{H}$,从而得出最终的命名实体识别结果。CRF是一种用于建模序列数据的统计建模方法,特别是在序列标签标注项目中得到了大力推广^[15]。将该方法应用于CNER时,CRF模型可以充分考虑中文文本前后文的关联性,为实体标签预测加入一定的约束,以此保证中文实体标签标注的合理性。

将由1.2.3小节输出的经过软注意力机制加权和掩码处理后的中文词特征向量序列 $\dot{H} = (h'_1, h'_2, \dots, h'_v)$,作为CRF层的输入,其输出的是中文命名实体标签序列 $Y = (y_1, y_2, \dots, y_v)$,则关于CRF测评分数 $E(h', y)$ 可以定义为:

$$E(h', y) = \dot{H} \sum_{i=0}^v G_{y_i, y_{i+1}} + \sum_{i=1}^v P_{i, y_i} \quad (13)$$

式中: $G_{y_i, y_{i+1}}$ 为CRF预测的中文命名实体标签 i 到标签 $i+1$ 的转移分数; P_{i, y_i} 为中文词对应的特征向量输出分数。CRF模型预测输出中文命名实

体标签序列 Y 对应的概率 $\rho(Y|\dot{H})$,可以描述为:

$$\rho(Y|\dot{H}) = \frac{e^{E(h', y)}}{\sum_{Y \in \hat{Y}_{\dot{H}}} E(h', y)} \quad (14)$$

式中: $\hat{Y}_{\dot{H}}$ 为输入序列 $\dot{H}$ 得出的所有中文命名实体标签序列。最后,采用维特比算法进行解码操作,由此可得到所有中文命名实体标签序列中,条件概率最大时的标签序列 $\tilde{Y}$,该序列即为CRF模型预测的最优标签序列,也就是中文命名实体识别的最终结果,用公式描述为:

$$\tilde{Y} = \arg \max (\rho(Y|\dot{H})) \quad (15)$$

2 实验分析

本文研究软注意力掩码嵌入下中文命名实体识别算法,以某医院的中文病例为研究对象展开实验。该医院是一所多科室的综合性医疗机构,拥有完善的医疗服务体系,能够满足不同患者的就医需求。

实验主要参数设置如表1所示,搭建的中文病例数据采集平台如图3所示。

表1 实验主要参数

Table 1 Main parameters of experiment

名 称	数 值
词向量维度	200
CBOW窗口大小	7
LSTM单元隐含层	60
隐藏层神经元数量	256
学习率	0.1
Dropout	0.5
数据批次大小	32
最大迭代数量	60

2.1 中文命名实体识别效果分析

利用搭建的中文病例数据采集平台分别从该医院的外科、中医科和口腔科搜集相关数据,组建中文病例数据集。其中,命名实体主要包括人名(Name)、症状(Symptom)、部位(Body)、检查(Test)和药物(Drug)5个类别,并以7:3的比例将数据分成训练与测试两种样本集,中文病例实体分布情况如表2所示。另外,实验使用BIO标注准则(表3)对实体统一进行标签标注。利用本文算法对中文病例数据集进行实体识别后,得出的结果如图4所示。

由图4可以看出,本文算法对中文病例实体

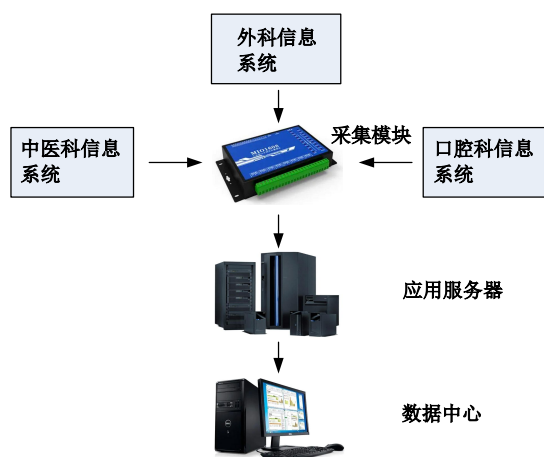


图3 中文病例数据采集平台

Fig. 3 Chinese case data collection platform

表2 中文病例实体分布情况

Table 2 Distribution of Chinese case entities

实体	训练集	测试集
人名(Name)	967	290
症状(Symptom)	3 684	1 105
部位(Body)	613	184
检查(Test)	2 846	854
药物(Drug)	1 057	317

表3 BIO标注准则

Table 3 BIO labeling criteria

标注名称	描述
B	实体的开始
I	实体的内部
E	实体的结束
O	非实体

进行精准识别,包括患者姓名、症状描述、患病部位以及检查项目和所用药物等关键信息,均被准确标注和提取。由此说明,本文方法在中文病例实体识别领域具有较高的准确性和实用性,能够有效地辅助医生进行病例分析和诊断,提高医疗工作的效率和准确性。

2.2 文本实体标签标注结果分析

对于CNER问题,实体标签标注的全面性至关重要,它反映了算法识别并标注出文本中所有相关实体的能力,这是评估算法性能的一个重要方面。为此,实验在不同信噪比条件下,输入不同大小的病例文本,利用覆盖指数指标对文本实体标签标注结果进行评估,以此衡量本文算法在实体标签标注覆盖性上的表现,得出的结果如图5所示。

从图5可以看出,随着病例文本大小的增加,

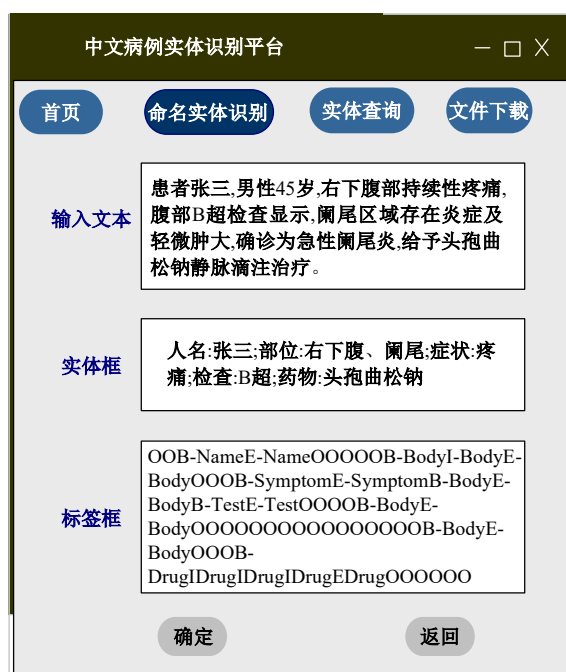


Fig. 4 Chinese case entity recognition results

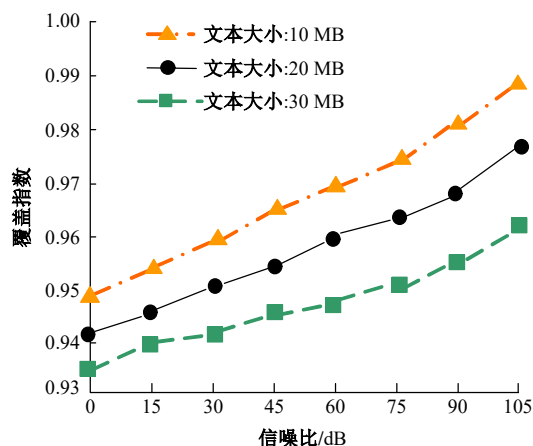


图5 实体标签标注覆盖情况

Fig. 5 Coverage of entity label annotations

尽管实体标签标注的覆盖指数呈现一定程度的下降,但是整体仍然维持在较高水平。即便是在信噪比较低的不利条件下,覆盖指数也依然能够保持在0.93以上。由此说明,本文算法受文本大小或信噪比变化影响较小,标注的实体标签能够较好地涵盖文本中的相关实体,具有较强的鲁棒性和可靠性。

2.3 基于软注意力掩码的消融实验分析

本文算法中软注意力掩码模块是一个较重要的组成部分。为了衡量此模块的有效性,实验进行了消融测试。分别利用本文算法(包含软注意力掩码模块)和消除软注意力掩码模块的算法对

不同大小的中文文本进行了命名实体识别,并利用 F_1 值指标(F_1 值是衡量算法识别精度的常见指标,其值在0~1,越接近1说明算法的性能越好,识别结果既精确又全面;反之则说明算法的性能有待提高,在精确度方面存在不足)对识别结果进行了评估。通过对比分析两种算法在相同数据集上的表现,得出了结果,如图6所示。

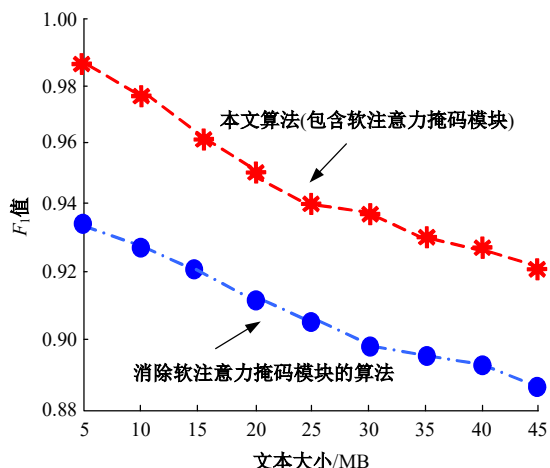


图6 本文方法消融测试

Fig. 6 Ablation test method described in this paper

由6图可知,应用本文算法(包含软注意力掩码模块)时,命名实体识别的 F_1 值较高,表现出优秀的识别性能;而在本文算法基础上消除软注意力掩码模块后, F_1 值显著下降,模型的性能明显减退,严重影响了识别结果的准确性。由此说明,软注意力掩码模块在本文算法中起到关键的作用,它能够有效地提升命名实体识别的准确性和全面性。

3 结束语

近年来,CNER方法取得了显著进步。然而,在实际应用中,由于中文文本的复杂性,如多义词、同义词以及上下文依赖等问题,使中文命名实体识别的准确性仍然面临挑战。为此,本文研究软注意力掩码嵌入下CNER算法,该算法通过引入软注意力机制对输入特征向量进行动态加权,并结合掩码策略有效过滤非关键信息,显著提升了算法对中文文本中命名实体的识别能力,最终实现了高精度、高效率的命名实体标注。未来,随着自然语言处理技术的持续进步以及多模态信息的融合应用,本文算法的泛化能力和适应性将得到进一步增强,将会综合利用文本、图像、声音等信息源,提升命名实体识别的准确性和鲁棒性。

参考文献:

- [1] 王颖洁,张程烨,白凤波,等. 中文命名实体识别研究综述[J]. 计算机科学与探索, 2023, 17(2): 324-341.
Wang Ying-jie, Zhang Cheng-ye, Bai Feng-bo, et al. Review of Chinese named entity recognition research[J]. Journal of Frontiers of Computer Science & Technology, 2023, 17(2): 324-341.
- [2] 赵继贵,钱育蓉,王魁,等. 中文命名实体识别研究综述[J]. 计算机工程与应用, 2024, 60(1): 15-27.
Zhao Ji-gui, Qian Yu-rong, Wang Kui, et al. Survey of Chinese named entity recognition research[J]. Computer Engineering and Applications, 2024, 60(1): 15-27.
- [3] 卢青华,袁丽娜. 基于组合神经网络的软件命名实体识别仿真[J]. 计算机仿真, 2023, 40(1): 489-492, 509.
Lu Qing-hua, Yuan li-na. Software named entity recognition simulation based on combined neural network[J]. Computer Simulation, 2023, 40(1): 489-492, 509.
- [4] 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 Yi-lin, Sun Lu-bing, Zhu Rong-bo, et al. Survey on Chinese named entity recognition with deep learning [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2022, 50(11): 44-53.
- [5] 张昀,黄橙,张玉瑶,等. 面向少量标注数据的中文命名实体识别[J]. 中文信息学报, 2023, 37(3): 101-111.
Zhang Yun, Huang Cheng, Zhang Yu-yao, et al. Chinese named entity recognition with few labeled data[J]. Journal of Chinese Information Processing, 2023, 37(3): 101-111.
- [6] 李健,熊琦,胡雅婷,等. 基于Transformer和隐马尔科夫模型的中文命名实体识别方法[J]. 吉林大学学报: 工学版, 2023, 53(5): 1427-1434.
Li Jian, Xiong Qi, Hu Ya-ting, et al. Chinese named entity recognition method based on Transformer and hidden Markov model [J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(5): 1427-1434.
- [7] Jeon K, Lee G, Yang S, et al. Named entity recognition of building construction defect information from text with linguistic noise[J]. Automation in Construc-

- tion, 2022, 143: No. 104543.
- [8] 方红, 苏铭, 冯一铂, 等. 结合 gazetteers 和句法依存树的中文命名实体识别[J]. 计算机工程与应用, 2022, 58(18): 227-232.
- Fang Hong, Su Ming, Feng Yi-bo, et al. Chinese named entity recognition combined with gazetteers and syntactic dependency tree[J]. Computer Engineering and Applications, 2022, 58(18): 227-232.
- [9] 廖梦, 贾真, 李天瑞. 基于标签信息融合与多任务学习的中文命名实体识别[J]. 计算机科学, 2024, 51(3): 198-204.
- Liao Meng, Jia Zhen, Li Tian-rui. Chinese named entity recognition based on label information fusion and Multi-task learning[J]. Computer Science, 2024, 51(3): 198-204.
- [10] 陈威达, 王林飞, 陶大鹏. 融合软注意力掩码嵌入的场景文本识别方法[J]. 中国图象图形学报, 2024, 29(5): 1381-1391.
- Chen Wei-da, Wang Lin-fei, Tao Da-peng. SAME-net: scene text recognition method based on soft attention mask embedding[J]. Journal of Image and Graphics, 2024, 29(5): 1381-1391.
- [11] 廖列法, 谢树松. 基于注意力机制特征融合的中文命名实体识别[J]. 计算机工程, 2023, 49(4): 256-262.
- Liao Lie-fa, Xie Shu-song. Chinese named entity recognition based on attention mechanism feature fusion[J]. Computer Engineering, 2023, 49(4): 256-262.
- [12] 占文韬, 吴晓鸽, 凌捷. 基于多窗口注意力机制的中文命名实体识别[J]. 小型微型计算机系统, 2024, 45(6): 1325-1330.
- Zhan Wen-tao, Wu Xiao-ling, Ling Jie. Chinese named entity recognition based on multi-window attention mechanism[J]. Journal of Chinese Computer Systems, 2024, 45(6): 1325-1330.
- [13] 赵丹丹, 黄德根, 孟佳娜, 等. 多头注意力与字词融合的中文命名实体识别[J]. 计算机工程与应用, 2022, 58(7): 142-149.
- Zhao Dan-dan, Huang De-gen, Meng Jia-na, et al. Chinese named entity recognition by integrating multi-heads attention mechanism and character and words fusion[J]. Computer Engineering and Applications, 2022, 58(7): 142-149.
- [14] 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]. 计算机工程, 2022, 48(6): 89-94, 106.
- Li Jun-huai, Chen Miao-miao, Wang Huai-jun, et al. Chinese named entity recognition method based on ALBERT-BGRU-CRF[J]. Computer Engineering, 2022, 48(6): 89-94, 106.
- [15] 张栋, 陈文亮. 基于上下文相关字向量的中文命名实体识别[J]. 计算机科学, 2021, 48(3): 233-238.
- Zhang Dong, Chen Wen-liang. Chinese named entity recognition based on contextualized char embeddings[J]. Computer Science, 2021, 48(3): 233-238.