

基于跨度与类别增强的中文新闻命名实体识别

祁瑞艳¹, 李龙杰¹, 徐世琚¹, 马笠恭², 马志新¹

(1. 兰州大学信息科学与工程学院, 甘肃 兰州 730000;

2. 兰州市公安局, 甘肃 兰州 730000)

摘要: 在新闻领域, 识别命名实体涉及复杂的语法结构和长名称, 这为确定实体边界带来了挑战, 同时也引发了序列标注方法在预测长实体时的提前中断问题。为了应对这些挑战, 提出了一种基于跨度与类别增强的中文新闻命名实体识别模型——SpaCE。该模型基于Transformer结构的双向编码器表示预训练模型(BERT), 通过跨度预测和类别描述增强, 提升了识别性能。在编码新闻文本信息的过程中, 模型结合类别描述以增强语义知识, 并采用基于跨度的解码方式来解决长实体预测中断问题。另外, 通过精确标记的方法引入词边界信息, 并优化实体匹配策略, 有效减少了由跨度解码引起的非实体匹配情形。与基线模型相比, SpaCE在3个数据集上的性能均有所提升。另外, 在无序文本上, SpaCE仍表现出了较强的命名实体识别能力, 具有很好的鲁棒性。

关键词: 新闻命名实体识别; BERT; 跨度; 类别增强; 词边界信息

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202437

Named entity recognition based on span and category enhancement for Chinese news

QI Ruiyan¹, LI Longjie¹, XU Shicheng¹, MA Ligong², MA Zhixin¹

1. School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China

2. Lanzhou Public Security Bureau, Lanzhou 730000, China

Abstract: In the field of news, the identification of named entities is complicated by complex syntactic structures and long entity names, which pose challenges for determining entity boundaries and lead to interruptions in predicting long entities using sequence labeling methods. To address these challenges, a model named SpaCE (span and category enhancement for Chinese news named entity recognition) was proposed. This model was developed based on the bidirectional encoder representation pre-trained model with a Transformer structure (BERT) and was enhanced by span prediction and category description to improve recognition performance. During the encoding of news text information, category descriptions were incorporated to enhance semantic knowledge, and a span-based decoding method was adopted to address interruptions in predicting long entities. Furthermore, word boundary information was introduced through precise labeling, and the entity matching strategy was optimized, effectively reducing non-entity matching caused by span decoding. Compared to baseline models, SpaCE demonstrated improved performance on three datasets. Furthermore, SpaCE exhibits strong named entity recognition capabilities on disordered texts, indicating its robustness.

Key words: news named entity recognition, BERT, span, category enhancement, word boundary information

收稿日期: 2024-08-13; 修回日期: 2024-09-24

通信作者: 马志新, mazhx@lzu.edu.cn

基金项目: 甘肃省科技计划项目 (No. 23YFGA0005, No. 21ZD8RA008)

Foundation Items: Science and Technology Program of Gansu Province (No. 23YFGA0005, No. 21ZD8RA008)

0 引言

命名实体识别 (named entity recognition, NER) 是自然语言处理中的重要任务, 旨在识别文本中具有特定意义的实体, 如人名、地名、组织机构名等。在中文新闻领域, 通过命名实体识别技术, 可以自动抽取非结构化文本中的实体, 并将它们划分到预设的不同类别中^[1]。这样一来, 便能更方便地对新闻文本进行分析和挖掘, 以更好地理解新闻事件的背景和相关实体之间的关系。

在政治新闻语料“中华人民共和国外交部和阿尔及利亚民主人民共和国外交部联合声明”中, “中华人民共和国外交部”可被视为一个组织机构实体, 也可分为“中华人民共和国”和“外交部”两个实体。考虑政治新闻信息的特殊性, 有必要将“中华人民共和国外交部”作为一个整体进行识别, 以确保实体含义的完整性。此外, 一些新闻报道中, 地名类型的实体可能以省市县镇村的格式显示。在识别这些实体时, 可将单独的省名视为一个地名实体, 但考虑上下文语境, 将整个地名作为一个整体识别可更全面地了解事件发生地。这些情况导致需识别的实体长度增加, 使用序列标注模型对长实体进行预测可能会出现中断, 需要一种更有效的方法来解决这一问题。

在NER任务中, 经典的解决方案是序列标注^[2-3], 即将整个输入文本视为一个序列, 为序列中的每个字进行标注, 标签分为实体和非实体两种。通过这种方式, 实体识别任务转化为对每个字进行分类。然而, 序列标注只能给每个字一个标签, 其在处理嵌套实体时存在一定的局限性, 因此, 近期的研究工作将命名实体识别从序列标注转向了跨度预测^[4-5]。跨度预测将包含实体信息的片段视为一个跨度, 不再单独考虑片段中每个字的标签, 而是将实体抽取任务视为识别开始索引、结束索引的问题, 并为二者之间形成的跨度分配预定义的实体类别。相较于序列标注, 跨度预测能够准确地定位实体在文本序列中的位置, 进而有助于提高实体识别的性能。这一转变不仅克服了序列标注的局限性, 也使得实体识别更为准确和高效。

从目前实体识别技术的发展来看, 跨度预测已被许多方法所采用, 并显著提升了实体识别的性能。目前, 由跨度的开始位置和结束位置确定实体的方法主要有两种, 其中较为流行的是“最近匹配

原则”。然而, 为了更合理地抽取跨度并考虑到长实体的情况, 本文采用了文献[6-7]提出的“启发式匹配原则”。该原则首先获取某一类别的候选开始位置和结束位置, 然后通过匹配概率最高的开始位置和结束位置来确定预测的实体片段。然而, 在某些情况下, 候选位置的概率值相差不大, 若每次都选择概率最高的匹配, 可能会导致解码出非实体的片段。以图1为例, “林”被预测为候选开始位置 s_1 , “李”被预测为候选开始位置 s_2 , “伟”被预测为候选结束位置 e_1 。若 s_1 的概率值略高于 s_2 , 那么“林”和“伟”之间的子序列——“林丹对阵李宗伟”, 就可能被解码为一个人名类型的实体。在这种情况下, 采用“启发式匹配原则”可能会面临一些挑战, 尤其是在处理概率接近的情况时。概率相近可能会导致实体识别的准确性和鲁棒性下降。

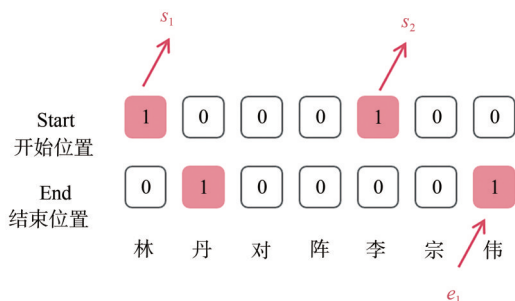


图1 跨度匹配实例分析

注: 值为1表示该字符被预测为实体开始或结束位置, 值为0表示该字符不属于实体边界, 下同。

针对上述问题, 本文提出了一种基于跨度与类别增强的中文新闻命名实体识别 (named entity recognition based on span and category enhancement for Chinese news, SpaCE) 模型。首先, 所提模型通过精确标定引入词的边界信息, 并将其与编码后的文本嵌入相结合, 形成结合向量。其次, 将结合向量输入双向长短期记忆网络 (bidirectional long short term memory, BiLSTM), 以增强模型对词边界信息的学习, 从而有助于精确定位一些实体边界模糊的位置。最后, 预先定义的实体类别描述语句经过共享编码器编码, 并融入携带词边界信息的文本表示中。这一步引入了有关实体类型的解释信息, 以辅助模型预测实体, 同时减少对训练数据的依赖。此外, 所提模型基于跨度实体识别, 通过优化实体匹配策略, 有效减少了跨度解码引起的非实

体匹配情形的发生。为了验证所提模型在中文命名实体识别任务中的有效性，在两个公开数据集 OntoNotes 4.0 和 MSRA，以及一个自建的数据集 GsrNews 上进行了实验。结果显示，在新闻领域的命名实体识别任务中，所提模型表现优异。

1 相关工作

2015年，Huang等^[3]提出了双向长短期记忆网络-条件随机场（bidirectional long short term memory - conditional random field, BiLSTM-CRF）模型，通过引入条件随机场（conditional random field, CRF）模块以优化输出序列标签，取得了显著的识别效果。因此，BiLSTM-CRF模型逐渐成为解决NER任务时基于序列标注的主流框架。基于序列标注的命名实体识别框架如图2所示，包括嵌入层、序列编码层和标签解码层^[8]。嵌入层的作用是将 token（处理文本的基本单位）级别的特征转化为向量表示，序列编码层进一步捕获文本序列的上下文特征，标签解码层获取序列内部的规则特征，并对编码层输出的特征向量进行分类。

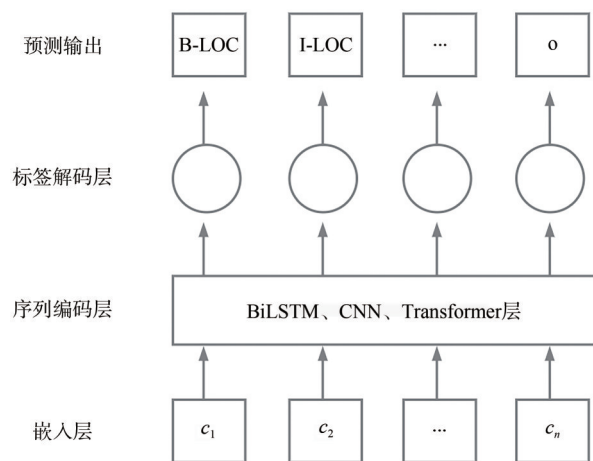


图2 基于序列标注的命名实体识别框架

注：BiLSTM、CNN、Transformer 是常见的编码层结构，B-LOC、I-LOC、O 是序列标注中常用的标签输出格式。

过去，针对命名实体识别任务的研究主要侧重于序列标注。为适应新闻领域数据的特点，研究人员通常会对现有的识别模型进行改进。在应对搜狐 coreEntityEmotion_train 语料中的核心实体识别和核心实体情感分析任务时，胡甜甜等^[9]在传统的 BiLSTM-CRF 模型基础上引入了注意力机制，提出了一种结合注意力机制的长短期记忆神经网络和条件随机场的模型，即 AttBi-LSTM-CRF。另外，

郑彦斌等^[10]在编码层选择使用双向门控循环单元（gated recurrent unit, GRU）替代 BiLSTM，设计了一种基于双向 GRU-CRF 的神经网络模型，用于处理中国驻东盟十国大使馆的中文新闻数据的命名实体识别任务。

在序列标注模型中，标签解码层通常采用 CRF 作为主要组件。这是因为 CRF 能够从训练数据集中学习到一些约束条件，以确保生成的实体标签序列是合理有效的。例如，标注序列开头是“B-”或“O”，“I-”不作为序列的起始标签；又或者标签“O”后面不会立即出现“I-”标签^[11]。通过考虑标签之间的转移概率，CRF 有助于生成更符合实际语言规律的标签序列。然而，由于 CRF 只考虑相邻标签之间的转移概率，对于一些较长的命名实体，可能出现预测中断的情况。

传统序列标注模型存在长实体预测中断和无法处理实体嵌套等问题。因此，近期的一些研究将命名实体识别方法从序列标注转向了跨度预测。基于跨度预测的命名实体识别方法通常包括两个步骤：跨度识别和跨度分类。Yu 等^[4]采用 BiLSTM 来获取序列中 token 的向量表示，接着利用两个前馈神经网络分别表示实体的头部 token 和尾部 token，然后通过双仿射注意力机制对句子中的开始位置和结束位置对进行打分，最终为这些跨度分配预定义的实体类别。Li 等^[5]首次提出了基于机器阅读理解框架来处理嵌套和非嵌套命名实体识别问题。对于嵌套实体，通过回答不同问题来识别不同类型的实体，但这会导致模型训练速度较慢。相比之下，Yang 等^[6]提出了一种新方法，该方法将文本序列和与类别相关的查询语句通过共享编码器产生嵌入，然后利用注意力机制将查询语句信息融入文本表示中。与之前 Li 等^[5]提出的将实体标签描述与原始文本进行拼接的方法相比，这种分开编码再融合的方法不仅解决了效率低下的问题，而且避免了被文本自身分散注意力，能更充分地利用标签知识。

与英文不同，中文句子中没有空格符来分隔单词。根据模型输入的不同粒度，在中文 NER 任务中，文本处理的基本单位通常是字级表示或词级表示。根据文献[2,12]，基于字级表示的方法在中文命名实体识别系统中的表现比基于词级表示的方法更佳。针对中文分词和命名实体识别任务，He 等^[12]提出了一种基于字符的条件随机场模型，对文本进行处理时分别比较了字级表示、词级表示以

及结合词性特征的词级表示。实验结果显示,基于字符的条件随机场模型展现了最佳的识别效果。然而,使用字向量作为神经网络的输入无法利用单词的语义信息,同时丧失了词的确切边界信息。中文句子中词语边界通常是实体边界^[13],这对于中文NER任务至关重要。

2 本文方法

本文提出一种名为SpaCE的新闻实体命名识别方法,SpaCE框架如图3所示,主要包括3个关键部分:语义编码模块、语义融合模块和跨度解码模块。在语义编码模块中,首先采用两个基于Transformer的双向编码器表示(bidirectional encoder representations from Transformers, BERT)预训练模型^[14],分别对新闻文本和类别描述信息进行编码,并通过精确标记引入词的边界信息,将其与编码后的文本嵌入相结合,形成结合向量。在语义融合模块中,将结合向量输入BiLSTM中,以加强模型对词边界信息的学习,并通过注意力机制将携带词边界信息的文本与类别描述信息融合,获得文本的增强嵌入表示。跨度解码模块则利用增强嵌入来预测每个字是否是某类别实体的开始位置或结束位置。

2.1 语义编码模块

语义编码模块的目标是将新闻文本和类别描述信息编码为嵌入表示。由于实体类别数量有限且类别描述文本都较短,在使用单独的编码器编码类别描述信息时,不能通过微调很好地训练模型。为了达到融合类别描述信息的目的,和文献[4]一样,两个编码器处理各自输入时共享模型权重,使得原始新闻文本在编码过程中就受到实体类别描述信息的影响,同时减少权重参数。

受到文献[5]的启发,针对预先定义各个类别实体集合 $T=\{t_1, \dots, t_{|T|}\}$,为每种类别提供了相应的类别描述信息 $Q^i=(q_1^i, \dots, q_m^i)$ 。例如,对于实体类别“会议”,其类别描述信息可以是“政府部门、专业协会等举办的具体会议名称”。将新闻文本($C=c_1, \dots, c_n$)和类别描述信息($Q=Q^1, \dots, Q^{|T|}$)预处理后分别输入两个BERT编码器中,获得它们各自的嵌入表示 $h_C \in R^{n \times d}$ (R 表示张量维度)和 $h_Q \in R^{|T| \times m \times d}$,这里 n 和 m 分别表示新闻文本和实体标签描述语句的序列长度, d 是隐藏层的特征维

度, $|T|$ 是实体类别的数量。

$$h_C = \text{BERT}(C) \quad (1)$$

$$h_Q = \text{BERT}(Q) \quad (2)$$

在新闻领域,命名实体识别涉及复杂的语法结构和长名称,这使得实体边界的确定变得非常困难,亟须提高识别性能。值得注意的是,在中文句子中,词语边界通常隐含着实体边界信息。为了确保所提模型能够准确学习词的边界信息,首先使用jieba分词的精确模式对文本进行分词,确定每个词在序列中的开始位置,然后根据每个词的长度,从其开始位置开始并延伸至词长度末尾,采用准确标记的方式构建包含这些位置间字序列的词边界向量 $v_{wb} \in R^{n \times 3}$ 。在词边界向量中,标记为1表示某个字是词的开始或结束,对于长度超过2的词,位于中间的字的向量将被标为(0,0,0),以示其既非词的开始字也非结束字。表1为一条新闻文本分词后通过精确标记构建词边界向量的示例。

表1 构建词边界向量的示例

新闻文本	新华社北京...意义			
	$c_1, c_2, c_3, c_4, c_5, \dots, c_{n-1}, c_n$			
分词	新华社	北京	...	意义
	c_1, c_2, c_3	c_4, c_5	...	c_{n-1}, c_n
词开始	1, 0, 0	1, 0	0	1, 0
词中间	0, 0, 0	0, 0	0	0, 0
词结尾	0, 0, 1	0, 1	0	0, 1

在新闻报道中,当介绍人物时,通常会将其姓名作为一个独立的小标题,如图4第二行的“钟南山”,并在后续文本的开头再次提及该人名(见图4第三行)。然而,在数据库存储中,它们被视为一个连续段落,因此同一个人名可能会连续出现而中间没有明显分隔。通过后续实验发现,针对这种情况只记录第一次出现的起始索引用于构建词边界向量,相较于每次出现都记录,实体识别的性能表现更出色。

2.2 语义融合模块

语义融合模块旨在将类别描述与携带词边界信息的文本嵌入相融合,以增强语义知识的表征。首先,将新闻文本的字向量 h_C 和词边界向量 v_{wb} 进行拼接,并将拼接后的向量 v_{C_wb} 输入BiLSTM中进行深入交互学习,以获得携带词边界信息的文本嵌入 h_{C_wb} 。

$$h_{C_wb} = \text{BiLSTM}(v_{C_wb}) \quad (3)$$

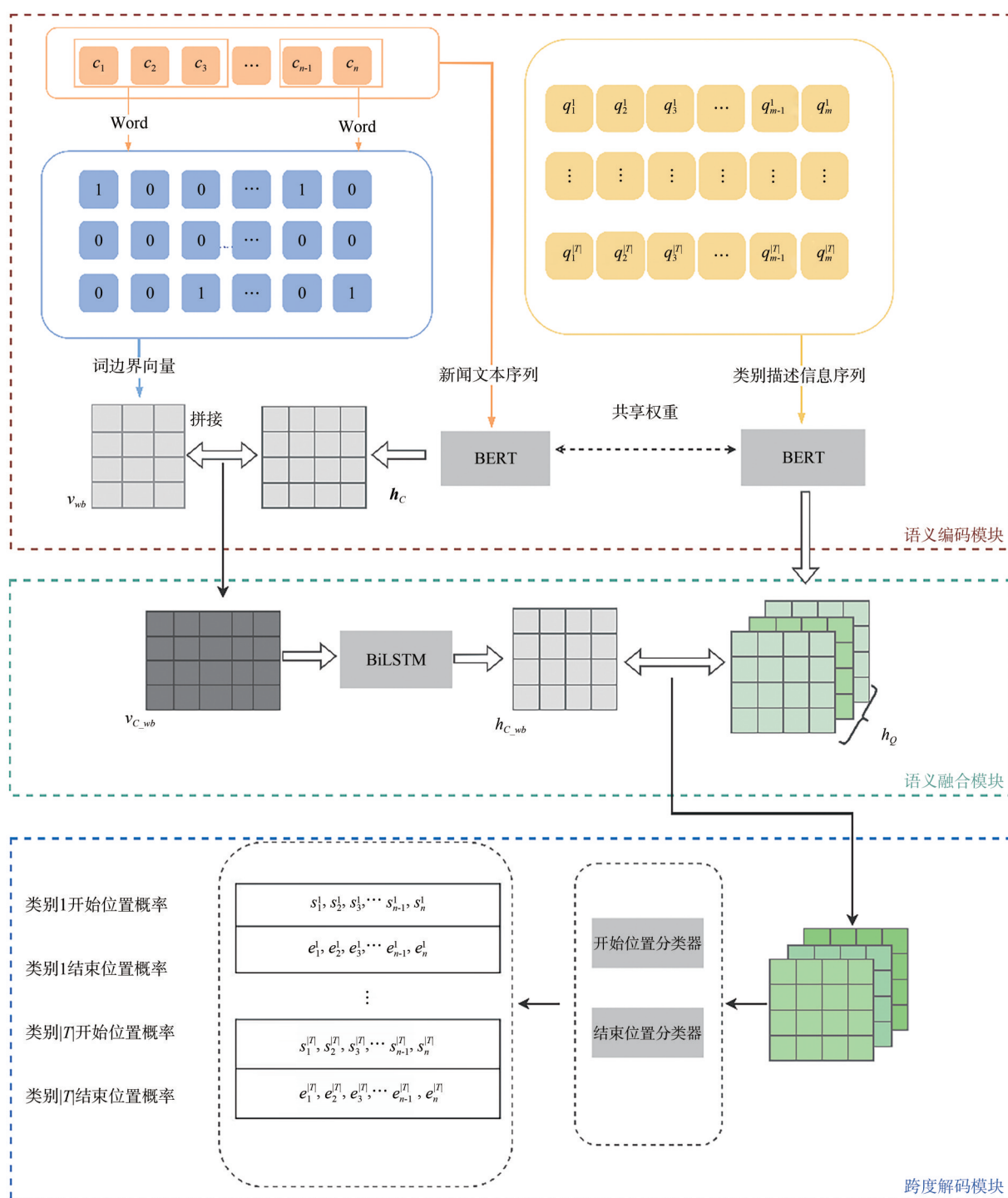


图3 SpaCE框架

通过注意力机制将携带词边界信息的文本和类别描述信息进行融合。其过程如下：首先，将文本中每个字的嵌入向量与类别描述中每个字的嵌入向量进行运算，得到对应的注意力分数。对于某一类别描述，文本中的每个字都会与该类别描述中的所有字产生一个注意力分数，从而形成该类别下的一

组注意力分数。通过对该类别下注意力分数进行归一化处理，得到注意力权重^[15]。其次，基于得到的注意力权重，计算文本中某个字对该类别描述的加权关注度。具体来说，将类别描述中每个字的向量与文本字向量根据对应的注意力权重相乘，并在类别维度上求和，从而得到针对某种实体类型描述

“共和国勋章”建议人选

钟南山

钟南山，男，汉族，中共党员，1936年10月生，福建厦门人，广州医科大学附属第一医院国家呼吸系统疾病临床医学研究中心主任，中国工程院院士，第十一、十二届全国人大代表，第八、九、十届全国政协委员。他长期致力于重大呼吸道传染病及慢性呼吸系统疾病的研究、预防与治疗，成果丰硕，实绩突出。

图4 人物介绍示例¹

所要进行融合的信息。最后，将融合信息与携带词边界信息的文本字向量相加，完成该类型描述信息的融合。具体计算如式(4)~式(6)所示。

$$\alpha_{c_{i_wb}, q_j^t} = \frac{\exp(\mathbf{h}_{c_{i_wb}} \cdot \mathbf{h}_{q_j^t})}{\sum_{j=1}^m \exp(\mathbf{h}_{c_{i_wb}} \cdot \mathbf{h}_{q_j^t})} \quad (4)$$

$$\mathbf{h}_{c_{i_wb}}^t = \mathbf{h}_{c_{i_wb}} + \sum_{j=1}^m \alpha_{c_{i_wb}, q_j^t} \cdot \mathbf{h}_{q_j^t} \quad (5)$$

$$\bar{\mathbf{h}}_{c_{i_wb}}^t = \tanh(V \cdot \mathbf{h}_{c_{i_wb}}^t + b) \quad (6)$$

在式(4)中， c_{i_wb} 表示新闻文本 C 中的第 i 个字(其中 $1 \leq i \leq n$)， $\mathbf{h}_{c_{i_wb}}$ 是编码层输出的隐藏向量。类似地， q_j^t 表示实体类别描述语句中的第 j 个字(其中 $1 \leq j \leq m$)， $t \in T$ (实体类别数量)， $\mathbf{h}_{q_j^t}$ 是输出的第 j 个字的隐藏向量。通过对这两者进行点积运算，并使用softmax函数获取注意力分数 $\alpha_{c_{i_wb}, q_j^t}$ 。

式(5)计算文本中每个字对类别描述中字信息的加权关注度和^[17]，将其加到原始编码层的语义表示中，从而获得融合了实体标签语义信息的表示 $\mathbf{h}_{c_{i_wb}}^t$ 。式(6)对融合后的嵌入表示进行tanh(·)运算得到 $\bar{\mathbf{h}}_{c_{i_wb}}^t$ ，其中， $V \in R^{d \times d}$ 和 $b \in R^d$ 均为可学习参数。

最后对字 c_{i_wb} 和所有类别的描述语句重复上述计算过程，以获得融合了所有类别描述信息的向量表示

$$\bar{\mathbf{h}}_{c_{i_wb}} = (\bar{\mathbf{h}}_{c_{i_wb}}^1, \dots, \bar{\mathbf{h}}_{c_{i_wb}}^{|T|})$$

2.3 跨度解码模块

将实体类别解释的语义信息融合到字级表示之后，SpaCE通过计算句子中每个字成为某一类实体的开始或结束位置的概率，来确定特定类别实体在句子中的跨度。与Li等^[5]所采用的方法相似，SpaCE通过训练获得两个分类器，即开始位置分类器和结束位置分类器。开始位置分类器用于计算第 i 个字的向量表示 $\bar{\mathbf{h}}_{c_{i_wb}}$ 为类别集合 T 中开始位置的概率，如(7)所示。

$$\mathbf{start}_{c_{i_wb}} = \text{sigmoid}\left(f\left(M_s \otimes \bar{\mathbf{h}}_{c_{i_wb}} + b_s\right)\right) \quad (7)$$

其中， $M_s \in R^{|T| \times d}$ 和 $b_s \in R^d$ 均为可学习参数， $\otimes$ 代表逐元素乘法， f 代表对输入矩阵进行行求和运算。 $\mathbf{start}_{c_{i_wb}}$ 是一个 $n \times |T|$ 大小的矩阵，即包含 $(\mathbf{start}_{c_{i_wb}}^1, \mathbf{start}_{c_{i_wb}}^2, \dots, \mathbf{start}_{c_{i_wb}}^{|T|})$ ， $\mathbf{start}_{c_{i_wb}}^1$ 表示 c_{i_wb} 为类别 T 集合中第一个类别的开始位置的概率。

结束位置分类器的原理与开始位置分类器相同，如(8)所示。

$$\mathbf{end}_{c_{i_wb}} = \text{sigmoid}\left(f\left(M_e \otimes \bar{\mathbf{h}}_{c_{i_wb}} + b_e\right)\right) \quad (8)$$

在当前的研究工作中，判断文本序列中的字是否为某一类别实体的开始或结束位置，存在两个主要的匹配原则：最近匹配原则和启发式原则。最近匹配原则^[18-19]是指将某一类别实体的开始位置和模型预测出同类别实体的结束位置中最接近的位置进行匹配，从开始位置到结束位置的子序列形成该类别的实体片段。启发式原则^[6-7]是指在某一类别的候选开始位置和结束位置中，仅匹配该类别实体概率最高的开始位置和结束位置。考虑到长实体的情况，本文采用了启发式原则。然而，由于候选位置的概率值在某些情况下相差不大，仅匹配概率最高的开始位置和结束位置可能导致匹配出非实体的情况。图5展示了在概率接近的情况下采用启发式原则可能会导致匹配出非实体的两种情形。在图5(a)中，“林”被预测为候选开始位置 s_1 ，“李”被预测为候选开始位置 s_2 ，“伟”被预测为候选结束位置 e_1 。若 s_1 作为开始位置的概率略高于 s_2 ，则“林丹对阵李宗伟”被解码为一个人名类型的实体。在图5(b)中，“林”被预测为候选开始位置 s_1 ，“丹”被预测为候选开始位置 e_1 ，“伟”被预测为候选结束位置 e_2 。若 e_2 作为结束位置的概率略高于

1 钟南山、张伯礼等4人被推为“共和国勋章”和“国家荣誉称号”建议人选，来源于央广网。

e_1 ，同样地，“林丹对阵李宗伟”被解码为一个人名类型的实体。

为了更有效地处理上述情形，在概率接近的情况下采用启发式原则，并结合最近匹配原则的思想。具体而言，对已有的候选位置和随后出现的符合条件的新候选位置进行比较，若两者的概率差值小于预设阈值，则考虑采用最近匹配原则；若概率差值超过阈值，则选择概率最高的开始和结束位置进行匹配。在图5所展示的情形下，对图5(a)概率值接近的候选位置 s_1 、 s_2 进行比较，如果二者的差值小于预设阈值，采用最近匹配策略，因此确定了 (s_2, e_1) 所形成的跨度“李宗伟”为目标预测实体。类似地，若图5(b)概率接近的候选位置 e_2 、 e_1 的差值也小于预设阈值，则 (s_1, e_1) 中所形成的跨度“林丹”即为所预测的实体。通过对实体匹配策略进行优化，有效减少了跨度解码引起的非实体匹配情形，从而提高了实体识别的准确性和鲁棒性。

3 实验分析

3.1 实验数据

为了验证本文模型在中文新闻命名实体识别任务中的有效性，使用了两个公开数据集——OntoNotes 4.0和MSRA^[20]，以及一个本文自建的数据集GsrBNews进行实验。其中，GsrBNews初始语料来源于甘肃日报社的100篇新闻文章。在同一篇新闻中，相邻的几句话通常会围绕同一个主题或话题展开，这些句子在语义上会存在一定的相似性。将相邻的几句话作为一个文本序列进行命名实体识别通常比单独处理单个句子更为有效，因为这种方式能够更全面地考虑上下文信息，从而提升模型对命名实体的识别能力。因此，本文对提取的新闻数据

进行了一定的处理：遇到句号、感叹号或问号后对句子进行切分，然后按照原始顺序对切分后的句子进行重新组合。如果重组后的句子长度超过254个字符，则会生成新的子数据，并对剩余句子重复执行上述过程，直至所有句子都被组装完成。之后对重组的句子进行人工标注，形成小型NER新闻数据集——GsrBNews。OntoNotes4.0和MSRA同样来自新闻领域，通常用于评估命名实体识别模型的性能。表2展示了3种实验数据集的统计信息，其中OntoNotes4.0和MSRA中的每条数据都是一个单独的句子。

表2 实验数据集的统计信息

数据集	训练集/条	验证集/条	测试集/条	实体类型数/种
GsrBNews	319	103	102	7
OntoNotes4.0	15 650	4 301	4 346	4
MSRA	41 729	4 637	4 366	3

通过对数据集进行观察，在标注过程中确定了7种常见的实体类别，并在表3中详细说明了每种类别的解释信息以及实体数量。基于表3中的类别定义，对新闻文本使用JSON格式进行标注，数据标注示例见表4。

3.2 实验设置

本文模型的主要部分包括编码器、词边界信息结合层、开始位置分类器和结束位置分类器。具体的网络结构如表5所示。

考虑到算力的限制，编码器选用了BERT-base模型，其核心结构为包含12层的BertEncoder。通过BiLSTM实现字向量与词边界向量的交互，其中参数embed_dim表示字向量维度与词边界向量维度的总和；bidirectional=True表示使用双向BiLSTM，其在时间步长 t 上的输出为 $h_t=[\vec{h}_t; \overleftarrow{h}_t]$ ，其中 $\vec{h}_t$ 是

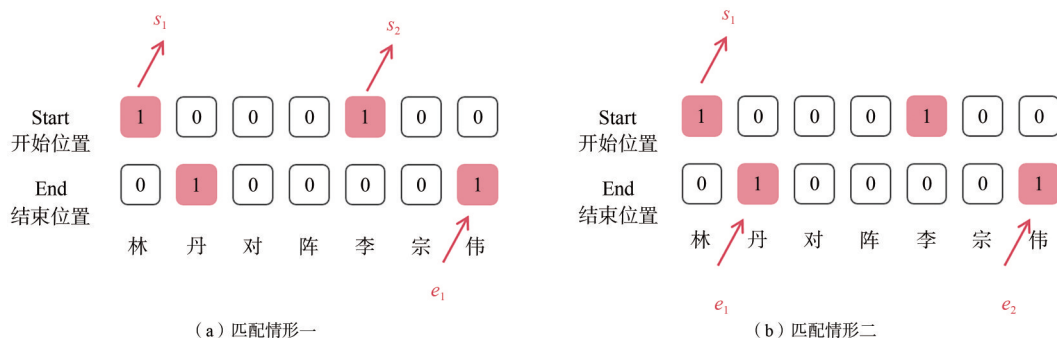


图5 导致匹配出非实体的两种情形

表3 GsrbNews类别名与类别解释

类别名	中文简称	类别解释	数量
CON	会议	包括政府部门、专业协会等所召开的具体会议名称	53
LOC	地名	具体的国家、城市、地区、街道等,或地理特征、地理区域、旅游景点、商业区域等	847
ORG	组织机构名	具体的组织机构名称,例如公司、学校、政府机构、新闻或媒体机构、非营利组织等	441
POS	职务、职位	在国家或特定组织机构中具有特定职责和权力范围的工作角色	429
PER	人名	人名和虚构的人物形象	714
TIME	时间	具体的时间点、时间段以及其他与时间相关的概念	539
WORK	作品	创作出来的具体艺术、文学、音乐、影视等作品	101

表4 数据标注示例

新闻文本	汪晓霞货车司机作为新就业群体之一,在经济社会发展中发挥了重要作用。为加强货车司机群体思想引导等工作,推动道路货运领域高水平治理、高质量发展,今年以来,我省聚焦摸底数、建体系、筑阵地、解诉求、送关爱等方面,扎实推进全域深化货车司机群体党建试点工作,使货车司机的心更暖了,服务经济社会发展的劲儿更足了,推动走出了一条党建引领道路货运事业高质量发展的新路子。健全组织体系,夯实党建引领行业发展的基础。我省东西纵横1 600多千米,货车司机单车经营方式居多,长期奔波在货运线上,流动性很强,管理和服务难度大。
所含实体	[{"start_idx": 0, "end_idx": 2, "type": "PER", "entity": "汪晓霞"}, { "start_idx": 3, "end_idx": 6, "type": "POS", "entity": "货车司机" }, { "start_idx": 36, "end_idx": 39, "type": "POS", "entity": "货车司机"}, { "start_idx": 70, "end_idx": 73, "type": "TIME", "entity": "今年以来"}, { "start_idx": 110, "end_idx": 113, "type": "POS", "entity": "货车司机" }, { "start_idx": 124, "end_idx": 127, "type": "POS", "entity": "货车司机"}, { "start_idx": 210, "end_idx": 213, "type": "POS", "entity": "货车司机"}]

表5 本文模型主要部分网络结构

模型组件	网络结构及参数
编码器	BertEmbeddings、BertEncoder、BertPooler
词边界信息结合层	LSTM(embed_dim=771, hidden_dim=384, bidirectional=True)
开始位置分类器	Classifier_Layer(class_num=7, out_features=768)
结束位置分类器	Classifier_Layer(class_num=7, out_features=768)

正向LSTM在时间步长 t 的隐藏状态, $\hat{h}_t$ 是反向LSTM在时间步长 t 的隐藏状态,“;”表示拼接操作,它将正向和反向的隐藏状态在每个时间步上按维度进行连接,从而生成更丰富的表示^[21]。因此隐藏层维度hidden_dim是最终输出维度的一半。在Classifier_Layer中,参数out_features表示最终输出向量的维度;class_num表示实体类别的数量。学习率遵循Yang等^[6]的设置,模型其他部分参数设置如表6所示。

在实体识别任务中,通常使用精确度(Precision)、召回率(Recall)和 F_1 值作为模型识

别效果的衡量标准^[22],其定义如下:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (9)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (10)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (11)$$

其中,TP表示模型识别正确的实体数量,FP表示模型识别错误的实体数量,FN表示模型未识别出的真实实体数量。

3.3 实验结果

为了验证SpaCE的有效性,本文选取了两种基于序列标注的方法——SoftLexicon^[23]和BERT-Tagger^[14],以及两种基于跨度预测的方法——LEAR^[6]和GlobalPointer^[24]。SoftLexicon是一种简单且有效的词汇增强NER方法,它将每个字符的所有匹配词语划分为BMES,共4个词集(即以某个字符为开始、中间、结束、单独的所有匹配词语),从而

表6 模型其他部分参数设置

数据集	最大序列长度	批次大小	训练轮次	学习率	
				编码层	任务层
GsrbNews	256	4	20	3×10^{-5}	6×10^{-5}
OntoNotes4.0	128	32	5	8×10^{-6}	8×10^{-5}
MSRA	128	32	20	3×10^{-5}	6×10^{-5}

将词汇信息融入字符表示中。BERT-Tagger利用BERT模型学习更丰富的词向量表示,以更好地捕捉上下文信息。LEAR为了学习到实体标签本身的语义信息,分别使用BERT编码器对标签描述文本和原始文本进行编码,得到不同的文本表示。为提升训练效率和减小模型尺寸,原始文本和标签描述文本共享BERT编码器的权重。GlobalPointer提出了一种新的基于片段的NER框架。在之前的跨度预测方法中,通常会使用两个独立的分类器分别预测实体的开始和结束位置,然后判断这两个位置之间是否构成一个完整的实体。然而,理论上应该联合考虑实体的开始和结束位置,因为它们共同构成一个完整的实体。基于这一理念,GlobalPointer将实体的首尾视为一个整体进行判别,从而更具全局性。

表7~表9是在3个不同数据集上的实验结果。整体而言,基于跨度预测的方法在3个数据集上表现出显著优于基于序列标注的方法的性能。与两个基于跨度预测的模型相比,SpaCE在3个数据集上均取得了最高的 F_1 值。并且,SpaCE和GlobalPointer均正确预测了较多实体。根据之前的分析,LEAR在跨度解码阶段使用启发式匹配原则,这可能导致非实体匹配的情况。例如,某个实体的开始位置可能与另一个实体的结束位置匹配成功,从而将本应识别为两个实体的情况错误地匹配为一个实体。相比之下,SpaCE通过优化实体匹配方法,减少了上述错误匹配情况的发生,因此其召回率高于LEAR。GlobalPointer则将跨度的开始位置和结束位置作为一个整体来判断是否构成完整实体,因此能够排除那些仅由分开的开始和结束位置组成的非实体。但是,也在一定程度上削弱了其对一些罕见实体的识别能力。表7显示,SpaCE在GsrBNews数据集上的实验结果,分别比其他两种基于跨度预测的方法在 F_1 值上提高了0.85%和1.03%。从表8和表9的结果来看,SpaCE在OntoNotes 4.0和MSRA两个数据集上的结果同样优于4个对比方法,表现出了很好的识别性能。

3个数据集上的识别结果证明了SpaCE的有效性。在SpaCE中引入词的边界信息,有助于确定一些模糊实体边界的位置。在处理候选位置概率接近的情况下,通过在具体情境下选择适宜的匹配策略,提高了判定预测实体为正确实体的准确性。在3个新闻数据集中,GsrBNews数据集的性能提升幅

表7 各模型在GsrBNews数据集上的实验结果

模型	精确率	召回率	F_1 值
SoftLexicon	75.00%	64.34%	69.26%
BERT-Tagger	87.68%	82.20%	84.85%
LEAR	92.10%	79.21%	85.17%
GlobalPointer	87.57%	83.86%	84.99%
SpaCE	91.56%	81.11%	86.02%

注:加粗数字表示最优结果,下同。

表8 各模型在OntoNotes4.0数据集上的实验结果

模型	精确率	召回率	F_1 值
SoftLexicon	77.13%	75.22%	76.16%
BERT-Tagger	78.01%	80.35%	79.16%
LEAR	83.79%	80.49%	82.11%
GlobalPointer	82.76%	82.78%	82.13%
SpaCE	84.06%	80.86%	82.43%

表9 各模型在MSRA数据集上的实验结果

模型	精确率	召回率	F_1 值
SoftLexicon	94.73%	93.40%	94.06%
BERT-Tagger	94.97%	94.62%	94.80%
LEAR	95.93%	94.60%	95.26%
GlobalPointer	94.19%	94.27%	94.09%
SpaCE	95.50%	95.19%	95.34%

度明显高于其他两个数据集。这一现象或许可以归因于另外两个数据集规模较大,这种规模优势促使模型在训练过程中更深入地学习数据特征,进而使每个模型均取得出色的实体识别效果,尤其在MSRA数据集上,性能提升幅度相对较小。相比之下,SpaCE在数据量较少的情况下与其他模型相比表现出显著的性能优势。

前面的数据处理阶段将每篇新闻视为一个长序列,并将其分割成多个片段,这种处理方式有助于保留原始新闻文档中句子之间的顺序结构。此外,为了验证SpaCE的鲁棒性,本文通过实验分析其在无序文本数据上的实体识别效果。为此,本文以句子结束符(句号、感叹号或问号)为分割标志,将GsrBNews数据集切割成独立句子,然后对切割后的句子进行随机排序以打乱其逻辑顺序。最后,遵循之前的处理策略,构建无序数据测试集。无序文本示例见表10。

表11显示了不同模型在无序文本数据上的实验结果。对比表11和表7的结果可以发现,在无序文本数据上,所有模型的性能都有不同程度的下

表10 无序文本及其包含的实体

新闻文本	亚吉铁路运营部门迅速开行了2列临时旅客列车,义务提供应急运输服务,在做好车辆消毒清洁的前提下,将学生们顺利送至目的地。习近平总书记强调,全党同志要永葆自我革命精神,增强全面从严治党永远在路上的政治自觉,决不能滋生已经严到位的厌倦情绪。她和老伴商量后决定,老两口分开住,分别照看孙子和外孙。远程办公、线上教育等新模式,为许多人居家工作和学习提供了支撑。
所含实体	[{"start_idx": 0, "end_idx": 7, "type": "ORG", "entity": "亚吉铁路运营部门"}, {"start_idx": 59, "end_idx": 61, "type": "PER", "entity": "习近平"}, {"start_idx": 62, "end_idx": 64, "type": "POS", "entity": "总书记"}]

降。在无序文本上,句子的逻辑顺序被打乱,相邻句子之间无关联性。模型在进行文本处理时无法利用句子的上下文信息,从而导致实体识别性能的下降。尽管如此,与其他模型相比,本文提出的SpaCE仍然保持了较高的识别结果,证明SpaCE具有较强的鲁棒性。

表11 不同模型在无序文本数据上的实验结果

模型	精确率	召回率	F_1 值
SoftLexicon	71.72%	61.68%	66.33%
BERT-Tagger	87.13%	80.98%	83.94%
LEAR	90.53%	76.63%	83.00%
GlobalPointer	88.76%	81.21%	84.20%
SpaCE	91.61%	80.16%	85.51%

此外,本文还对SpaCE在不同数据集上的实体识别能力进行了分析。表12展示了3个数据集上的实体类别,图6显示了SpaCE在各种实体类型上的识别效果。

表12 3个数据集中涉及的各种实体类别

类别名	中文简称	类别名	中文简称
CON	会议	PER	人名
LOC	地点	TIME	时间
ORG	组织机构名	WORK	作品
POS	职务、职位	GPE	地理政治实体

在GsrBNews数据集上,SpaCE在不同实体类型的识别中表现优异,尤其在会议(CON)和作品(WORK)这两类实体的识别中,均达到了100%的准确率。这一现象可以归因于这些实体在语法结构上具有明显的特征性标识。例如,会议类实体通常包含“会议”“全会”等关键词,而作品类实体通常被书名号表示。这些显著的语言特征大大增强了模型的识别能力。相比之下,组织机构(ORG)类实体的识别效果较弱。进一步分析表明,ORG类实体中存在大量简写形式,如“中国建筑”和“铁建国际”,实际上分别指“中国建筑集团有限公司”和“中国铁建国际集团有限公

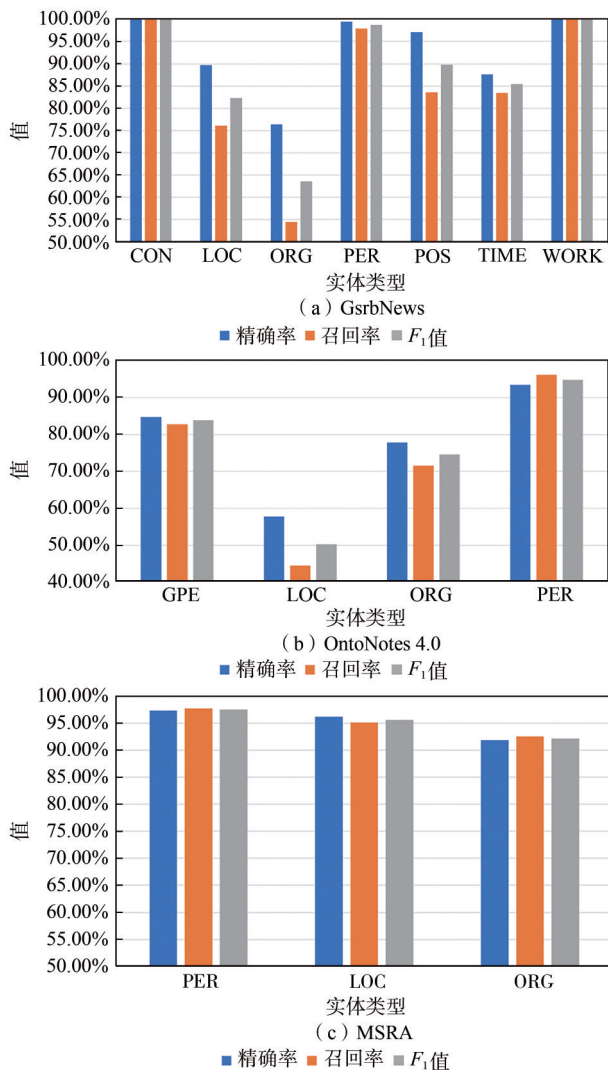


图6 SpaCE于GsrBNews、OntoNotes4.0、MSRA数据集上各实体类型的识别效果

司”。简写往往会引发歧义,例如将“中国建筑”误解为泛指中国的建筑物,进而增加了模型的识别难度,导致识别效果不佳。在OntoNotes 4.0数据集中,地理政治实体(GPE)指国家、城市、州县等行政区域,而地点(LOC)则代表山脉、河流等自然地理景观。通过对比这两类实体的识别情况发现,LOC类的实体通常包含“海”“湾”“岛”等自然地理特征词汇。然而,一些GPE中也可能含

有这些词汇，如“珠海”和“长岛”，这使得模型需要具备更强的分辨能力，才能准确区分这两类实体。特别是在遇到同时具有自然地理特征和城市属性的名称时，模型的识别难度显著增加，这也是部分LOC类的实体识别效果不佳的原因之一。在MSRA数据集中，SpaCE在不同类型的实体识别上表现出色，3种类型的实体均达到了较高的识别精度。

3.4 消融实验

为了进一步验证所提模型的有效性，本文设计了一系列SpaCE的变体进行实验：在语义编码模块中，不考虑新闻文本中词的边界信息（w/o wb）；基于跨度解码的过程中仅使用启发式原则，不优化匹配策略（w/o ms）；不引入词边界信息并且采用单一的启发式原则（w/o wb&ms）。本文在GsrBNews和OntoNotes4.0数据集上进行了实验，实验结果见表13和表14。

在未优化实体匹配策略的情况下，模型在两个数据集上的性能均呈现下降趋势。这表明优化实体

表13 GsrBNews数据集上的消融实验结果

方法	精确率	召回率	F_1 值	ΔF_1
SpaCE	91.56%	81.11%	86.02%	0.00%
w/o ms	91.41%	80.98%	85.88%	↓0.14%
w/o wb	92.42%	79.48%	85.46%	↓0.56%
w/o wb&ms	92.10%	79.21%	85.17%	↓0.85%

表14 OntoNotes4.0数据集的消融实验结果

方法	精确率	召回率	F_1 值	ΔF_1
SpaCE	84.06%	80.86%	82.43%	0.00%
w/o ms	83.95%	80.76%	82.32%	↓0.11%
w/o wb	83.51%	80.89%	82.18%	↓0.25%
w/o wb&ms	83.79%	80.49%	82.11%	↓0.32%

匹配策略对于减少由跨度解码引起的非实体匹配情况至关重要，并在新闻数据集上提升了实体识别性能。对比去除词边界信息的模型实验结果，可以看出，引入词边界信息所蕴含的实体边界信息能够改善网络输入特征，有助于确定某些实体边界模糊的位置，从而进一步提升模型性能。在去除词边界信息且仅使用启发式原则的情况下，模型在两个数据集上的表现同样有所下降，这进一步证实了本研究以词边界信息与优化实体匹配策略为研究出发点的有效性。

3.5 案例分析

针对同一份新闻语料，传统的序列标注模型SoftLexicon的识别结果如图7所示，而SpaCE的识别结果见表15。通过比较两种模型的实体识别结果，观察到在人名（PER）类型实体的识别方面，两种模型均展现出良好的识别能力。然而，在地点（LOC）和时间（TIME）类型实体的识别中，使用基于序列标注的方法时，SoftLexicon在识别“山东省淄博市高青县常家镇”

近	日	,	山东省淄博市高青县常家镇											
O	O	O	B-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC
天鹅湖村白莲藕			种植基地			,	藕	农	采					
I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	O	O	O	O	O	O	O	O
收	白	莲	藕	,	脸	上	洋	溢	着	丰	收	的	喜	悦
O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
。	中秋		节	即	将	来	临	,	市	场	上	对	荷	藕
O	B-TIME	I-TIME	O	O	O	O	O	O	O	O	O	O	O	O
的	需	求	量	大	增	,	藕	农	们	抢	抓	农	时	收
O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
获	白	莲	藕	,	供	应	节	日	市	场	。	董振霞		
O	O	O	O	O	O	O	O	O	O	O	O	B-PER	I-PER	I-PER
王克军			摄	影	报	道								
B-PER	I-PER	I-PER	O	O	O	O								

图7 SoftLexicon的识别结果

地”和“中秋节”这两个实体时存在识别不完整的问题，效果不佳。与之相比，SpaCE能够准确识别这两个实体，表现更出色。

在另一个新闻数据集上，本文对3种基于跨度预测的方法进行了性能评估，具体结果如表16。LEAR在识别结果中将“老挝万象有望夕发朝至。图为建设中的野象谷车站”（红字标记部分）识别为地点（LOC）类型实体；而GlobalPointer则未能正确识别“老挝万象”。这种情况可能是由于GlobalPointer将跨度的起始和结束作为整体来判断是否构成完整实体。相比之下，该模型会排除那些仅由分开的开始和结束位置组成的非实体，但也因此在某种程度上减弱了对一些罕见实体的识别能力。在本文提出的SpaCE中，通过引入词边界信息和优化实体匹配策略，相对于LEAR降低了导致误匹配情况发生的可能性。与GlobalPointer相比，SpaCE能够准确标识出“老挝万象”（绿字标记部分）这一地点类型实体，显示出更为出色的性能

表现。

表17展示了3个基于跨度预测的方法在GsrB-News数据集上漏识别和错误识别的实体。与LEAR和GlobalPointer相比，SpaCE只漏掉了一个实体且不存在错误识别的实体。因此，SpaCE在处理无序文本时具有更强的识别能力。

4 结束语

为了解决中文新闻命名实体识别任务中基于序列标注的方法在预测长实体时可能出现的中断情况，以及获得实体的候选开始位置、结束位置后基于跨度解码时出现的非实体匹配问题，本文基于BERT预训练模型，采用跨度和类别增强的方法，提出了SpaCE。SpaCE对文本进行分词后，采用精确标记的方式引入词的边界信息，将其与文本表示结合，有助于定位一些实体边界模糊的位置。为确保模型充分学习到实体标签本身的语义信息，本文为预定义的实体标签提供了相应的类别描述信息。

表15 SpaCE的识别结果

新闻文本	近日,山东省淄博市高青县常家镇天鹅湖村白莲藕种植基地,藕农采收白莲藕,脸上洋溢着丰收的喜悦。中秋节即将来临,市场上对莲藕的需求量大增,藕农们抢抓农时收获白莲藕,供应节日市场。董振霞、王克军摄影报道
识别结果	[{"text": "山东省淄博市高青县常家镇天鹅湖村白莲藕种植基地", "label": "LOC", "start_idx": 4, "end_idx": 26}, {"text": "王克军", "label": "PER", "start_idx": 91, "end_idx": 93 }, {"text": "董振霞", "label": "PER", "start_idx": 88, "end_idx": 90 }, {"text": "中秋节", "label": "TIME", "start_idx": 47, "end_idx": 49}]

表16 基于跨度预测的方法的识别结果

新闻文本	近期,中老昆(明)万(象)铁路国内段建设进展顺利,位于中老边境的友谊隧道建设进入收尾阶段,各车站土建工程基本完工,正线铺轨完成超过1/3。中老昆万铁路全长1 000多千米,建成通车后,云南昆明至老挝万象有望夕发朝至。图为建设中的野象谷车站。新华社发
正确实体	[{"start_idx": 27, "end_idx": 35, "type": "LOC", "entity": "中老边境的友谊隧道"}, {"start_idx": 91, "end_idx": 94, "type": "LOC", "entity": "云南昆明"}, {"start_idx": 96, "end_idx": 99, "type": "LOC", "entity": "老挝万象"}, {"start_idx": 113, "end_idx": 117, "type": "LOC", "entity": "野象谷车站"}, {"start_idx": 119, "end_idx": 121, "type": "ORG", "entity": "新华社"}]
LEAR	[{"text": "老挝万象有望夕发朝至。图为建设中的野象谷车站", "label": "LOC", "start_idx": 97, "end_idx": 118}, {"text": "云南昆明", "label": "LOC", "start_idx": 92, "end_idx": 95}, {"text": "新华社", "label": "ORG", "start_idx": 120, "end_idx": 122}]
GlobalPointer	[{"start_idx": 119, "end_idx": 121, "type": "ORG", "entity": "新华社"}, {"start_idx": 91, "end_idx": 94, "type": "LOC", "entity": "云南昆明"}, {"start_idx": 113, "end_idx": 117, "type": "LOC", "entity": "野象谷车站"}]
SpaCE	[{"text": "老挝万象", "label": "LOC", "start_idx": 97, "end_idx": 100}, {"text": "野象谷车站", "label": "LOC", "start_idx": 114, "end_idx": 118}, {"text": "云南昆明", "label": "LOC", "start_idx": 92, "end_idx": 95}, {"text": "新华社", "label": "ORG", "start_idx": 120, "end_idx": 122}]

注：（1）蓝色标记部分为3个模型均未识别出的实体；（2）GlobalPointer预测结果索引从0开始，LEAR和SpaCE预测结果索引从1开始。

表17 基于跨度预测模型在无序文本数据上的识别结果

示例文本	亚吉铁路运营部门迅速开行了2列临时旅客列车,义务提供应急运输服务,在做好车辆消毒清洁的前提下,将学生们顺利送至目的地。前些年,浙江省推行最多跑一次改革,很快,这项改革在全国推广。感谢你们的关心,这些清洁的水会让今年的宰牲节过得很好。中老昆万铁路全长1 000多千米,建成通车后,云南昆明至老挝万象有望夕发朝至。加强创新突破,转变发展理念、创新发展模式、增强发展动能,确保高质量发展。2019年年底,城关区被工业和信息化部、民政部、国家卫生健康委员会确定为智慧健康养老示范基地。(图7)
LEAR	"missed": [{"text": "城关区", "label": "LOC", "start_idx": 201, "end_idx": 203}, {"text": "老挝万象", "label": "LOC", "start_idx": 147, "end_idx": 150}, {"text": "亚吉铁路运营部门", "label": "ORG", "start_idx": 1, "end_idx": 8}] "false": [{"text": "老挝万象有望夕发朝至。加强创新突破,转变发展理念、创新发展模式、增强发展动能,确保高质量发展。2019年年底,城关区", "label": "LOC", "start_idx": 147, "end_idx": 203}, {"text": "前些年", "label": "TIME", "start_idx": 60, "end_idx": 62}]
GlobalPointer	"missed": [{"text": "城关区", "label": "LOC", "start_idx": 201, "end_idx": 203}, {"text": "老挝万象", "label": "LOC", "start_idx": 147, "end_idx": 150}, {"text": "亚吉铁路运营部门", "label": "ORG", "start_idx": 1, "end_idx": 8}] "false": [{"start_idx": 146, "end_idx": 147, "type": "LOC", "entity": "老挝"}]
SpaCE	"missed": [{"text": "亚吉铁路运营部门", "label": "ORG", "start_idx": 1, "end_idx": 8}] "false": []

注: missed表示漏掉的实体, false表示错误识别的实体。

SpaCE通过注意力机制将这些信息与携带词边界信息的文本相融合。此外,在跨度解码阶段,通过优化实体匹配策略,有效减少了跨度解码引起的非实体匹配情形。在3个新闻领域数据集中,本文提出的SpaCE显著提升了性能。此外,本文还通过在无序文本上的实验进一步验证了SpaCE的鲁棒性。

参考文献:

- [1] 罗兵,张显峰,段立,等.基于RoBERTa-Span-Attack的标签指针网络军事命名实体识别[J].海军工程大学学报,2024,36(1):76-82,93.
- [2] LIU B, ZHANG X F, DUAN L, et al. Military named entity recognition based on RoBERTa-Span-Attack tag pointer network[J]. Journal of Naval University of Engineering, 2024, 36(1): 76-82, 93.
- [3] LIU X Z, ZHU C H, ZHAO T J. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? [C]//International Conference on Intelligent Computing. Berlin, Heidelberg: Springer, 2010: 634-640.
- [4] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB]. arXiv preprint, 2015, arXiv:1508.01991.
- [5] YU J T, BOHNET B, POESIO M. Named entity recognition as dependency parsing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 6470-6476.
- [6] LI X Y, FENG J R, MENG Y X, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5849-5859.
- [7] YANG P, CONG X, SUN Z Y, et al. Enhanced language representation with label knowledge for span extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 4623-4635.
- [8] YANG S, FENG D W, QIAO L B, et al. Exploring pre-trained language models for event extraction and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5284-5294.
- [9] 郑洪浩,宋旭晖,于洪涛,等.基于深度学习的中文命名实体识别综述[J].信息工程大学学报,2021,22(5):590-596.
- [10] ZHENG H H, SONG X H, YU H T, et al. Survey of Chinese named entity recognition based on deep learning[J]. Journal of Information Engineering University, 2021, 22(5): 590-596.
- [11] 胡甜甜,但雅波,胡杰,等.基于注意力机制的Bi-LSTM结合CRF的新闻命名实体识别及其情感分类[J].计算机应用,2020,40(7):1879-1883.
- [12] HU T T, DAN Y B, HU J, et al. News named entity recognition and sentiment classification based on attention-based bi-directional long short-term memory neural network and conditional random field[J]. Journal of Computer Applications, 2020, 40(7): 1879-1883.
- [13] 郑彦斌,夏志超,郭智,等.东盟十国新闻文本的命名实体识别[J].科学技术与工程,2018,18(35):162-168.
- [14] ZHENG Y B, XIA Z C, GUO Z, et al. Named entity recognition of news texts in ten ASEAN countries[J]. Science Technology and Engineering, 2018, 18(35): 162-168.
- [15] 高国忠,李宇,华远鹏,等.基于BERT-BiLSTM-CRF模型的油气领域命名实体识别[J].长江大学学报(自然科学版),2024,21(1):57-65.
- [16] GAO G Z, LI Y, HUA Y P, et al. Named entity recognition in oil and gas domain based on the BERT-BiLSTM-CRF model[J]. Journal of Yangtze University (Natural Science Edition), 2024, 21(1): 57-65.
- [17] HE J Z, WANG H F. Chinese named entity recognition and word segmentation based on character[C]//Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. [s.l.:s.n.] 2008: 108-4022.
- [18] 李宝昌,郭卫斌.词典信息分层调整的中文命名实体识别方法[J].华东理工大学学报(自然科学版),2023,49(2):276-283.
- [19] LI B C, GUO W B. Chinese named entity recognition based on hierarchical adjustment of lexicon information[J]. Journal of East China University of Science and Technology, 2023, 49(2): 276-283.

- [14] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv:1810.04805.
- [15] 崔文成, 王可丽, 邵虹. 基于稠密块和注意力机制的肺部病理图像异常细胞分割[J]. 智能科学与技术学报, 2023, 5(4): 525-534.
CUI W C, WANG K L, SHAO H. Abnormal cell segmentation for lung pathological image based on denseblock and attention mechanism[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(4): 525-534.
- [16] 李超, 侯霞, 乔秀明. 融合知识的文博领域低资源命名实体识别方法研究[J]. 北京大学学报(自然科学版), 2024, 60(1): 13-22.
LI C, HOU X, QIAO X M. A low-resource named entity recognition method for cultural heritage field incorporating knowledge fusion[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2024, 60(1): 13-22.
- [17] 于媛芳, 张勇, 左皓阳, 等. 基于角色信息引导的多轮事件论元抽取[J]. 北京大学学报(自然科学版), 2023, 59(1): 83-91.
YU Y F, ZHANG Y, ZUO H Y, et al. Multi-turn event argument extraction based on role information guidance[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2023, 59(1): 83-91.
- [18] DU X Y, CARDIE C. Event extraction by answering (almost) natural questions[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 671-683.
- [19] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1476-1488.
- [20] GUI T, ZOU Y, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 1040-1050.
- [21] 角远韬, 李润梅, 王剑. 基于模糊自然语言处理的铁路 CTC 接口文本智能测试方法[J]. 智能科学与技术学报, 2024, 6(2): 201-209.
JIAO Y T, LI R M, WANG J. Intelligent testing method for railway CTC interface data based on fuzzy natural language processing[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(2): 201-209.
- [22] 祁鹏年, 廖雨伦, 覃斌. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统, 2023, 44(9): 1857-1868.
QI P N, LIAO Y L, QIN B. Survey on deep learning for Chinese named entity recognition[J]. Journal of Chinese Computer Systems, 2023, 44(9): 1857-1868.
- [23] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[EB]. arXiv preprint, 2020, arXiv:1908.05969.
- [24] SU J L, MURTADHA A, PAN S F, et al. Global pointer: novel efficient

span-based approach for named entity recognition[EB]. arXiv preprint, 2022, arXiv:2208.03054.

[作者简介]



祁瑞艳 (2001-), 女, 兰州大学信息科学与工程学院硕士生, 主要研究方向为深度学习、自然语言处理等。



李龙杰 (1981-), 男, 兰州大学信息科学与工程学院博士生, 主要研究方向为大数据分析与应用、图机器学习、深度学习等。



徐世铮 (2000-), 男, 兰州大学信息科学与工程学院硕士生, 主要研究方向为深度学习、自然语言处理等。



马笠恭 (1969-), 男, 本科, 兰州市公安局网络安全保卫支队, 主要研究方向为网络安全管理、网络安全防护等。



马志新 (1973-), 男, 博士, 兰州大学信息科学与工程学院教授, 主要研究方向为大数据分析与应用、复杂网络分析、网络空间安全等。