文章编号：1003-0077（2017）00-0000-00

融入注意力机制的越南语组块识别方法

王闻慧[1] 毕玉德[2] 雷树杰[1]
（1.信息工程大学 洛阳校区，河南省 洛阳市 471003；2.复旦大学 外国语言文学学院，上海市 200433）

摘要：对于越南语组块识别任务，在前期对越南语组块内部词性构成模式进行统计调查的基础上，该文针对 Bi-LSTM+CRF 模型提出了两种融入注意力机制的方法：一是在输入层融入注意力机制，从而使得模型能够灵活调整输入的词向量与词性特征向量各自的权重；二是在 Bi-LSTM 之上加入了多头注意力机制，从而使模型能够学习到 Bi-LSTM 输出值的权重矩阵，进而有选择地聚焦于重要信息上。实验结果表明，在输入层融入注意力机制后，模型对组块识别的 F 值提升了 3.08%，在 Bi-LSTM 之上加入了多头注意力机制之后，模型对组块识别的 F 值提升了 4.56%，证明了这两种方法的有效性。

关键词：越南语；组块识别；Bi-LSTM+CRF 模型；注意力机制

中图分类号：TP391 **文献标识码：**A

Vietnamese Chunk Identification Incorporating Attention Mechanism

Wang Wenhui[1], Bi Yude[2], and Lei Shujie[1]
(1. Luoyang Division, Information Engineering University, Luoyang, Henan 471003, China; 2. College of Foreign Language and Literature, Fudan University, Shanghai 200433, China)

Abstract: For the Vietnamese chunk identification task, based on the statistical investigation of the internal part-of-speech (POS) patterns of the Vietnamese chunks in the early stage, this paper proposes two ways to integrate the attention mechanism with the Bi-LSTM+CRF model: the first is to integrate the attention mechanism at the input layer, which allows the model to flexibly adjust weights of word embeddings and POS feature embeddings; the second is to add a multi-head attention mechanism on the top of Bi-LSTM, which enables the model to learn weight matrix of the Bi-LSTM outputs and selectively focus on important information. Experimental results show that after integrating the attention mechanism at the input layer, the F-value of Vietnamese chunk identification is increased by 3.08% and after adding the multi-head attention mechanism on the top of Bi-LSTM, the F-value of Vietnamese chunk identification is improved by 4.56%, which demonstrates the effectiveness of the two methods.

Key words: Vietnamese; Chunk Identification; Bi-LSTM+CRF; Attention Mechanism

0 引言

句法分析在自然语言处理任务中占据着重要的位置，是机器翻译（Machine Translation）、自动问答（Automatic Question Answering）等更复杂任务的基础。由于语言自身的复杂性，尤其对于像越南语这样缺乏形态标记、以字为单位的孤立语而言，实现完全的句法分析十分困难。为此，Abney[1]提出了组块分析理论，该理论采取先将句子中的组块识别出，再寻找组块之间关系的方法，降低了句法分析的复杂度。自此，组块识别成为研究者长期关注的重要课题。

对于越南语组块识别而言，其主要面临着以下三大难题：一是越南语缺乏形态标记，并与汉语一样主要通过虚词和词序来表示语法信息，这使得在越南语组块识别中可利用的标记信息较少；二是越南语存在定语后置的现象，这增加了越南语名词组块内部构成的复杂性，同时也加大了越南语名词组块识别的难度；三是在越南语中，动词作定语与动词作谓语在形式上完全一样，这增加了名词组块与动词组块之间的辨识难度。

收稿收稿日期：2019-04-29；**定稿日期**：2019-

对于组块识别而言，早期的识别方法主要基于规则，如基于有限状态机的方法[2]、基于转换学习与错误驱动的方法[3-4]等。从 21 世纪初开始，基于 MBL[5]、SVM[6]、CRF[7]等传统统计模型以及规则与统计模型相结合的方法[8-10]被广泛应用在组块识别任务中。近年来，随着深度学习的兴起，该方法也开始应用于组块识别任务中[11]。而对于越南语的组块识别而言，主要有 Lê Minh Nguyên[12] 等人采用 CRF、SVM、Online Passive-Aggressive Learning 等模型对越南名词组块进行识别，实验结果显示 CRF 模型的识别效果最好。Nguyen Thi Huong Thao[13]等人将词性特征、词汇正字法特征融入到 CRF 模型中对越南语名词短语进行识别，实验结果显示词性、词汇正字法特征对越南语名词短语的识别效果均有提升作用。郭剑毅[14]等人分析总结出了越南语名词组块词性组合特征，并将其作为约束条件融入到 CRF 模型中，得到了较好的识别效果。李佳[11]使用字符级的词向量作为输入，并将词性特征融入到 Bi-LSTM+CRF 模型中对越南语组块进行识别，取得了较好的识别效果。

综合来看，目前对越南语组块识别的研究还较少，识别效果还有很大的提升空间，所使用的模型也主要集中在 CRF 等传统统计模型上。而在深度学习方法的应用方面，目前所采用的模型也较为单一，主要为 Bi-LSTM+CRF 模型，缺乏对如注意力机制等深度学习技术最新发展的应用。此外，在深度学习方法中，当前研究所采用的融入特征的方法也较为机械，大多采用向量之间直接串联拼接的方法，不能够根据输入灵活确定词向量与特征向量各自的权重，这些都限制了对越南语组块的识别效果。为此，本文主要针对深度学习方法进行改进：一是将注意力机制引入神经网络的输入层，使得模型能够灵活决定词向量与特征向量各自的权重；二是将注意力机制融入到 Bi-LSTM+CRF 模型中，从而使模型能够有选择地聚焦于对识别有效的信息上。

1 越南语组块内部结构

1.1 越南语组块

关于越南语组块的界定，从目前来看并没有形成统一的标准，本文以 VLSP（Vietnamese Language and Speech Processing，越南语及语音处理会议）网站公布的越南语组块语料为调查语料库，将越南语组块定义为内部可以嵌套同类型组块的词语序列。在 VLSP 语料中，涉及到的组块类型有八类，如表 1 所示。

表 1 本文组块类型及示例

标记	类型	示例
NP	名词组块	gần 90% gia_đình（近 90%的家庭）
VP	动词组块	t ình_nguyện thu_hẹp（期望收缩）
PP	介词组块	Ngay từ（从……）
AP	形容词组块	mới nhất（最新的）
QP	数量组块	98%
TP	时间组块	hiện_nay（现在）
WH	疑问组块	v ì sao（为什么）
O	其他	v à（和）

1.2 越南语组块内部词性组合模式

以 VLSP 公布的组块标注语料（语料已经进行了词性标注）为调查语料库，本文对各类型组块的内部词性组合模式进行了统计。在表 1 所示的八种越南语组块类型中，名词组块、动词组块、介词组块和形容词组块所占比率最高，共占到了语料中全部组块的 99.94%，为此，本文主要对调查语料库中的名词组块、动词组块、介词组块和形容词组块四种类型组块的内部词性组合模式进行调查统计。其中，对名词组块、动词组块、介词组块与形容词组块频数排名前十位的内部词性组合模式的统计结果分别如表 2、表 3、表 4 和表 5 所示。

在表 2、表 3、表 4 和表 5 中，以“+”作为词性之间的连接符。从四种组块类型的内部词性组合模式来看，介词组块内频数排名前十位的词性组合模式所对应的组块占到了全部介词组块的 99%以上，动词组块与形容词组块在该项统计指标上也分别达到了 93.56%与 96.06%，而名词组块中频数排名前十位的词性组合模式所对应的组块占全部名词组块的比例最低，为 81.36%。

从以上数据可看出，越南语组块内部词性构成模式规律性明显且分布较为集中，因此将词性特征融入到组块识别任务中能够为组块识别提供更多的信息。这是本文在模型中融入词性特征的语言学依据。

从模型的角度讲，由于多头注意力机制能够更好地捕获输入序列中各输入值之间的内在联系[15]，因此将多头注意力机制应用于越南语组块识别任务能够使模型更有效地利用组块的内部构成信息，并通过赋予其相应的权重，有效提升模型对组块的识别效果。这是本文将多头注意力机制

融入 Bi-LSTM+CRF 模型的语言学基础。

从对未登录越南语组块识别的角度讲，使模型能够在遇到未登录越南语组块时相应地增加词性特征信息的权重，并相应地减少词汇信息的权重，则能够提升模型对未登录越南语组块的识别效果。这是本文在深度学习模型输入层融入注意力机制的语言学依据。

表 2 名词组块词性组合模式统计

词性组合模式	数目	累计占比
N（普通名词）	98838	45.84%
P（代词）	20118	55.17%
Np（专有名词）	14182	61.75%
M（数词）	11232	66.96%
Nc（量词）	10451	71.80%
L（限定词）+N	6077	74.62%
N+N	5392	77.12%
N+Np	4273	79.11%
M+N	2470	80.25%
N+V（动词）	2382	81.36%

表 3 动词组块词性组合模式统计

词性组合模式	数目	累计占比
V	101356	83.95%
V+V	3679	87.00%
R（副词）+V	3233	89.68%
V+N	1033	90.53%
V+A（形容词）	950	91.32%
V+R	887	92.06%
R+V+V	620	92.57%
A+V	444	92.94%
R+V+R	424	93.29%
V+V+V	324	93.56%

表 4 介词组块词性组合模式统计

词性组合模式	数目	累计占比
E（介词）	40469	97.53%
E+Cc（并列连词）+E	140	97.87%
R	133	98.19%
R+E	99	98.43%
E+E	86	98.64%
T（助动词）	73	98.81%
Cc+E	63	98.97%
Cc	59	99.11%
T+E	59	99.25%
N	52	99.38%

表 5 形容词组块词性组合模式统计

词性组合模式	数目	累计占比
A	26065	89.24%
R+A	711	91.67%
A+A	431	93.15%
A+N	239	93.97%
A+R	224	94.73%
A+T	98	95.07%
A+C	98	95.41%
T+A	74	95.66%
R+R+A	58	95.86%
A+P	58	96.06%

2 融入注意力机制的 Bi-LSTM+CRF 模型

2.1 越南语词向量与词性特征向量获取

词的分布式表示[16]是一种将词向量化的有效方法，其能够一定程度上表示词的语义信息，是深度学习技术应用于自然语言处理领域的基础。本文通过 Word2Vector 开源工具获取词向量，其包含有 CBOW 与 Skip-gram 两种模型，其中 CBOW 模型通过上下文来预测当前词，Skip-gram 模型则通过当前词来预测上下文。本文选取 CBOW 模型作为词向量的训练模型，对于 CBOW 模型而言，其训练目标是最大化如下函数：

$$\tau = \sum_{w \in C} \log p(w|Context(w)) \quad (1)$$

式（1）中，C表示语料中所有词的集合，w表示属于C的某个词，$Context(w)$表示词w的上下文。

本文使用 VnCoreNLP[17]工具对来自维基百科的大规模无监督越南语语料进行分词和词性标注，分别形成与维基语料相对应的分词语料与词性语料。其中，分词语料为维基语料所对应的词序列，而词性语料为分词语料所对应的词性序列。通过使用 Word2Vector 模型分别对分词语料与词性语料进行训练，本文获取了预训练的越南语词向量与词性特征向量。

2.2 注意力机制

自 2017 年 Bahdanau 等人[18]在英法机器翻译任务中应用注意力机制以来，注意力机制被广泛使用在自然语言处理的各项任务中。虽然注意力机制通常使用在 Seq2Seq 模型中，并作为 Encoder-Decoder 的一种机制来使用，但注意力机制作为一种思想，可以用来支持各种类型的自然语言处理任务。注意力机制的核心思想在于通过计算权重矩阵使得模型有选择地聚焦于重要信息上，其本质是一个查询到一系列（键-值）对的映射，其计算公式如式（2）、（3）、（4）所示。

$$f(\boldsymbol{Q}, \boldsymbol{K}_i) = \boldsymbol{Q}^T \boldsymbol{K}_i \quad (2)$$

$$\alpha_i = soft\max\big(f(\boldsymbol{Q}, \boldsymbol{K}_i)\big) = \frac{\exp(f(\boldsymbol{Q}, \boldsymbol{K}_i))}{\sum_i \exp(f(\boldsymbol{Q}, \boldsymbol{K}_i))} \quad (3)$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sum_i \alpha_i \boldsymbol{V}_i \quad (4)$$

式（2）、（3）、（4）中，$\boldsymbol{Q}$ 表示查询，$\boldsymbol{K}$ 与 $\boldsymbol{V}$ 组成

（键-值）对。式（2）用来计算 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的相似度，其中，相似度的获取除了式（2）中所示的点乘法以外，还可以通过余弦相似性或引入额外的神经网络来获取。一般而言，式（2）、（3）、（4）中的 $\boldsymbol{K}$ 与 $\boldsymbol{V}$ 相等，而在自注意力机制中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 均相等。

作为一种较为成熟的序列标注模型，Bi-LSTM+CRF 被广泛地应用在各种自然语言处理任务中。针对 Bi-LSTM+CRF 模型，本文使用了两种融入注意力机制的方法：一是在 Bi-LSTM 层上添加了一层多头注意力机制，详见 2.3；二是将注意力机制融入到 Bi-LSTM+CRF 模型的输入层中，以获取加入了相应权重的联合向量表示，详见 2.4。

2.3 Bi-LSTM + Multi-Head Attention + CRF

LSTM（Long-Short-Term Memory，长短时记忆网络）是 RNN（Recurrent Neural Network，循环神经网络）的一种变体，其通过加入门限机制一定程度上缓解了 RNN 面临的梯度弥散和梯度爆炸问题。Bi-LSTM 层利用了 LSTM 正向与反向两个序列方向上的信息来对输入信息进行处理，而 CRF 层则通过计算输出值之间的转移概率，进而将输出值间的转移信息融入到模型中，从而提升模型的效果。Bi-LSTM+CRF 模型的整体架构如图 1 所示。

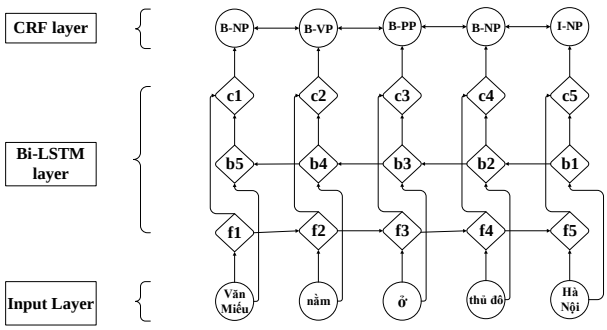

图 1 Bi-LSTM + CRF 模型框架

多头注意力机制是由 Vaswani 等人[15]在 2017 年提出，其由多个放缩点积注意力机制（Scaled Dot-Product Attention）组成，其内部结构如图 2 所示。

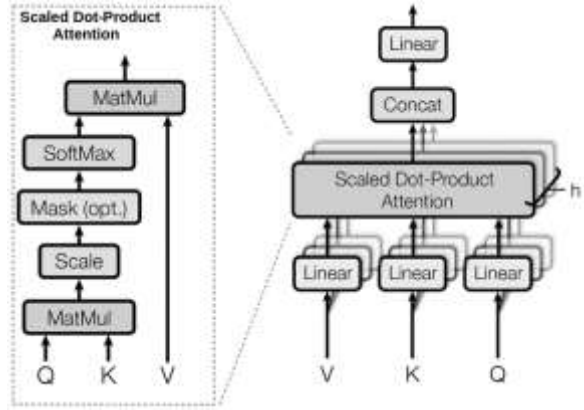

图 2 多头注意力机制

由图 2 可知，在放缩点积注意力机制中，通过对查询 Q 与（键-值）对中的键 K 进行相似度运算等一系列操作，可以获得权重矩阵，进而使模型有选择地聚焦于重要信息上。而在多头注意力机制中，在对输入进行线性变换以后，要进行 h 次放缩点积注意力操作。之后，将 h 次放缩点积注意力操作后的向量进行串联拼接，并进行线性变换后作为多头注意力机制的输出。根据 Vaswani 等人的研究成果，进行多次放缩点积操作的好处在于可以使模型在不同的表示子空间里学到更多的信息[15]。

由 1.2 可知，越南语组块内部构成的规律性较为明显，而多头注意力机制有着较强的利用输入序列中各输入值间规律和关系的能力，因此将多头注意力机制加入识别模型可以增强模型利用其内部构成信息的能力。为此，本文在 Bi-LSTM+CRF 模型的基础上加入了多头注意力机制。融入了多头注意力机制的 Bi-LSTM+CRF 模型的整体架构如图 3 所示。

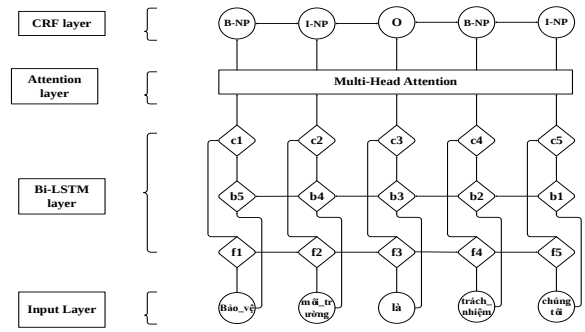

图 3 Bi-LSTM + Multi-Head Attention + CRF 模型

在图 3 中，模型由输入层、Bi-LSTM 层、Attention 层与 CRF 层组成。其中，输入层将输入的词与词性特征转化为相应的向量化表示，并采用首尾串联拼接的方式组合为联合向量输入到 Bi-LSTM 层中。Attention 层在接收 Bi-LSTM 层

的输出后，通过计算权重矩阵，增强了模型利用重要信息的能力，从而获得识别效果的提升。

2.4 融入注意力机制的联合向量表示

在以往基于深度学习的序列标注任务中，特征向量的加入一般通过与词向量的首尾串联拼接获得，如图 4 所示。

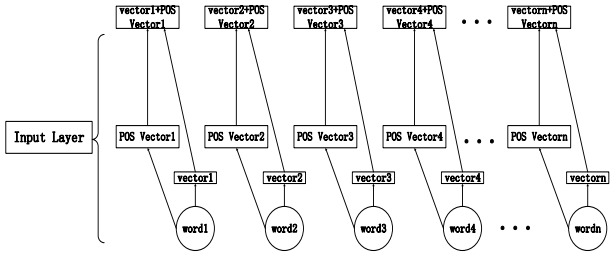

图 4 直接串联的联合向量表示

在图 4 中，通过将预训练的词向量与预训练的词性特征向量首尾串联拼接，得到了融入词性信息的联合向量表示，并作为模型的输入层参与到序列标注任务中。但这种获取联合向量表示的方式较为机械，且不能够对词向量与特征向量在联合向量中的权重进行灵活调整。受 Rei 等人[19]工作的启发，本文提出了融入注意力机制的联合向量表示方法，其计算方法如式（5）、（6）、（7）所示。

$$\boldsymbol{e}_{word} = \alpha \cdot \boldsymbol{v}_{\text{word}} + \beta \cdot \boldsymbol{v}_{\text{pos}} \quad (5)$$

$$\alpha = \sigma(\boldsymbol{W}_a^3 \tanh(\boldsymbol{W}_a^1 \boldsymbol{v}_{\text{word}})) \quad (6)$$

$$\beta = \sigma(\boldsymbol{W}_a^4 \tanh(\boldsymbol{W}_a^2 \boldsymbol{v}_{\text{pos}})) \quad (7)$$

式（5）中，$\boldsymbol{v}_{\text{word}}$表示词向量，$\boldsymbol{v}_{\text{pos}}$表示词性特征向量，$\alpha$为词向量的权重系数，$\beta$为词性向量的权重系数。在式（6）与式（7）中，$\boldsymbol{W}_a^1$、$\boldsymbol{W}_a^3$与$\boldsymbol{W}_a^2$、$\boldsymbol{W}_a^4$是分别用来计算$\alpha$与$\beta$的权重矩阵，$\sigma$表示 sigmoid 函数，其与 tanh 均为激活函数。

通过在输入层加入注意力机制，可以使模型灵活地调整输入的词向量与词性特征向量的权重，进而能够更好地处理序列标注任务，如图 5 所示。

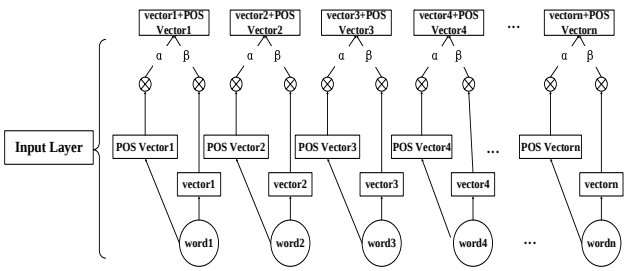

图 5 基于注意力机制的联合向量表示

在图 5 中，预训练的词向量与预训练的词性特征向量被分别输入一层神经网络，并在激活函数的激活下获得各自的权重（词向量的权重为α，词性特征向量的权重为β）。之后，词向量与词性特征向量分别与各自的权重相乘，相乘获得的两个向量通过首尾串联拼接的方式组合为联合向量输入 Bi-LSTM+CRF 模型。与 Rei 等人的方法不同，本文的方法不要求词性特征向量的维度必须与词向量相同，也不要求α与β的和为 1，这进一步增强了本文模型的灵活性。

3 实验及结果分析

3.1 实验数据

本文使用 VLSP 网站公布的组块标注语料为实验数据，语料总规模超过 70 万词。语料中包含 8 种类型的组块，其中名词组块 215620 个、动词组块 120733 个、介词组块 41492 个、形容词组块 29208 个，其余四种组块共 641 个。本文按照 5：1 的比例将语料划分为训练集与测试集。在测试语料中，含有各类型组块 68988 个，其中未登录组块 14108 个，未登录组块占比为 20.45%。

本文使用 IOB2 标注规范，每一类型组块包含“B-组块类型”与“I-组块类型”两种标注类别，其中“B-组块类型”用来标注该类型组块的开头部分，“I-组块类型”则用来标注该类型组块的中间部分与结尾部分，而对于非组块组成成分，统一标注为“O”。本文所使用语料共包含八种组块类型，共计 17 种标注类别。

3.2 评测指标

为了全面评价模型对组块识别的情况，本文设置了 6 个评价指标，如表 6 所示。

表 6 评测指标

评价指标	计算公式
准确率	$P=\frac{\text{all-right-tags}}{\text{all-tags}}$ （4）
越南语组块识别准确率	$P_C=\frac{\text{all-correctly-recognized-Chunks}}{\text{all-recognized-Chunks}}$ （5）
越南语组块识别召回率	$R_C=\frac{\text{all-correctly-recognized-Chunks}}{\text{all-Chunks}}$ （6）
越南语组块识别 F 值	$F_C=\frac{(\beta^2+1)\,P_C R_C}{\beta^2\,(P_C+R_C)}$ （7）
未登录越南语组块识别召回率	$R_{UKC}=\frac{\text{all-correctly-recognized-unknown-Chunks}}{\text{all-unknown-Chunks}}$ （8）

未登录越南语组块类别召回率	$R_{UKTC}=\frac{all-correctly-recognized-unknown-Chunk-Types}{all-unknown-Chunk-Types}$ （9）

在表 6 中，准确率 P 是指标签标注准确率，用来评价整体识别情况；越南语组块识别准确率 P_C 是指对越南语组块整体的识别准确率，只有对整个越南语组块内的所有组成词标注正确才算对该组块识别正确；越南语组块识别召回率 R_C 是对越南语组块整体识别的召回率；越南语组块识别 F-value 则综合评价对越南语组块整体的识别效果；未登录越南语组块识别召回率 R_{UKC} 则用来评价模型对未登录组块的识别效果，是评价模型泛化能力的重要指标，由于对越南语组块的识别的难点和关键点都在于对未登录组块的识别，该指标也是反映模型识别效果的重要指标；未登录越南语组块类别召回率 R_{UKTC} 则排除了对同一未登录越南语组块的反复识别造成的 R_{UKC} 虚高的情况，从类别的角度评价模型对未登录越南语组块的识别效果，该指标同样也是评价模型泛化能力的重要指标。

此外，本文还分别对测试语料中含有的名词组块、动词组块、介词组块和形容词组块的识别情况进行了统计。为了在文中更加清晰直观地反映模型对不同类型组块的识别情况，并对识别情况进行全面的评价，本文对各类型组块识别情况的评价指标设为 F 值，以名词组块为例，其评价指标表示为 F_{NP}。在计算各类型组块的相应指标时，只有对组块整体包含的各个组成词都标注正确才算作对组块识别正确。

3.3 模型设置

本文的模型在训练过程中全部使用自适应学习率优化函数 Adam 作为模型用优化函数。为了避免学习率过高导致的损失值 loss 出现大幅度的震荡，本文在多次试验调整后将模型的 learning rate 设置为 0.001。此外，本文也多次调整 batch size 的大小以达到效果的最优，最终将 batch size 设置为 128。为防止模型出现过拟合现象，本文采用了 Dropout 的方法，并将 dropout 值设置为 0.5，即在每一个迭代训练过程中随机去除 50%的数据量。

为了避免参数设置不同对模型识别效果造成的影响，在本文所进行的实验中，模型的上述超参数设置完全一致，从而验证本文提出的两种将注意力机制融入 Bi-LSTM+CRF 模型方法的有效性。

3.4 实验设计

本文使用了 VLSP 网站公布的 VietChunker[13] 作为本文实验的基准模型，使用其在本文测试集上的测试结果作为本文实验的基线标准。

本文的实验分为五个部分，第一部分使用 VietChunker 进行测试；第二部分使用 Bi-LSTM+CRF 模型，并采用预训练的词向量作为输入；第三部分使用 Bi-LSTM+CRF 模型，并采用预训练的词向量与词性特征向量首尾串联拼接形成的联合向量作为模型输入；第四部分使用 Bi-LSTM + Multi-Head Attention + CRF 模型，采用预训练的词向量与词性特征向量首尾串联拼接形成的联合向量作为模型输入；第五部分使用 Bi-LSTM+CRF 模型，并采用融入注意力机制的联合向量作为模型输入，形成 Attention-over-Input Layer + Bi-LSTM + CRF 架构。

通过五部分实验结果的对比，可以验证本文提出的两种融入注意力机制方法的有效性。

3.5 实验结果与分析

本文在五种实验条件下对全部越南语组块的识别情况如表 7 所示。

表 7 全部组块识别情况统计

	P	P_C	R_C	F_C
VietChunker	85.41%	86.89%	74.66%	80.32%
Bi-LSTM + CRF (词向量)	88.26%	86.06%	79.35%	82.57%
Bi-LSTM + CRF (词向量+词性特征向量)	93.80%	93.03%	87.26%	90.05%
Bi-LSTM + Multi-Head Attention + CRF (词向量+词性特征向量)	96.69%	95.72%	93.51%	94.61%
Attention-over-Input-Layer + Bi-LSTM + CRF (词向量+词性特征向量)	95.96%	94.70%	91.61%	93.13%

由表 7 可知，在本文使用的模型中，所有模型的效果都要好于 VietChunker，这体现了本文方法的有效性。

在 Bi-LSTM+CRF 内部，在加入词性特征向量后，模型对越南语组块的识别效果有了显著提升。其中，在准确率 P 上提升了 5.54%，在越南语组块识别准确率 P_C 上提升了 6.97%，越南语组块识别召回率 R_C 上提升了 7.91%，F_C 上提升了

7.48%，可以看出词性特征对越南语组块识别的提升作用非常明显。

相对于加入词性特征向量的 Bi-LSTM+CRF 模型，在加入多头注意力机制后，模型的识别效果得到了进一步的提升，在准确率 P 上提升了 2.89%，在越南语组块识别准确率 P_C 上提升了 2.69%，越南语组块识别召回率 R_C 上提升了 6.25%，F_C 上提升了 4.56%。这些数据表明，多头注意力机制的加入显著提升了模型对越南语组块的识别效果。

而对于 Attention-over-Input-Layer + Bi-LSTM + CRF 方法而言，相对于加入词性特征向量的 Bi-LSTM+CRF 模型，其在准确率 P 上提升了 2.16%，在越南语组块识别准确率 P_C 上提升了 1.67%，越南语组块识别召回率 R_C 上提升了 4.35%，F_C 上提升了 3.08%，这证实了在输入层融入注意力机制方法的有效性。但相对于融入多头注意力机制的方法而言，在输入层融入注意力机制的方法在越南语组块的识别效果上要相对差一些，其在准确率 P 上要低于前者 0.73%，在 F_C 上低于前者 1.48%。

本文在五种实验条件下对越南语名词组块、动词组块、介词组块与形容词组块的识别效果如表 8 所示。

表 8 各类型组块识别情况统计

	F_{NP}	F_{VP}	F_{PP}	F_{AP}
VietChunker	72.38 %	87.74 %	98.95 %	84.89 %
Bi-LSTM + CRF (词向量)	77.77 %	89.28 %	91.71 %	79.85 %
Bi-LSTM + CRF (词向量+词性特征向量)	84.68 %	96.24 %	99.82 %	93.03 %
Bi-LSTM + Multi-Head Attention + CRF (词向量+词性特征向量)	91.68 %	98.01 %	99.89 %	96.66 %
Attention-over-Input-Layer + Bi-LSTM + CRF (词向量+词性特征向量)	88.57 %	98.92 %	99.89 %	95.79 %

由表 8 可知，在五种实验条件下，模型对四种越南语组块的识别情况与表 7 中所示的对全部越南语组块的识别情况大体一致。而从四种组块类别的角度分析，在五种实验条件下，模型对介词组块的识别效果最好，对名词组块的识别效果最差。这一定程度上反映出这四种不同组块类别内部构成的复杂性不同，其中，名词组块因其内部构成最为复杂、歧义性最为显著，从而使得模型对其识别效果最差。从统计学的角度分析，由 1.2 可知，在这四种越南语组块类型中，内部词性组合模式规律性最为明显的就是介词组块，其前十位词性组合模式所对应的组块就占到了全部介词组块的 99.38%，而名词组块的前十位内部词性组合模式所对应的组块仅占到全部名词组块的 81.26%，这一定程度上解释了表 8 所示的实验结果。

作为评价模型识别效果的重要指标，未登录组块识别召回率能够一定程度上反映模型的泛化能力，本文在五种实验条件下对未登录越南语组块的识别效果如表 9 所示。

表 9 未登录组块识别情况统计

	R_{UKC}	R_{UKTC}
VietChunker	46.56%	53.89%
Bi-LSTM + CRF (词向量)	48.61%	55.21%
Bi-LSTM + CRF (词向量+词性特征向量)	75.50%	82.89%
Bi-LSTM + Multi-Head Attention + CRF (词向量+词性特征向量)	82.69%	86.87%
Attention-over-Input-Layer + Bi-LSTM + CRF (词向量+词性特征向量)	82.71%	89.29%

从表 9 中可以看到，相对于 VietChunker，本文所使用模型在对未登录越南语组块识别方面的表现都要更加优异。而在 Bi-LSTM+CRF 内部，在加入词性特征向量后，Bi-LSTM+CRF 模型对未登录越南语组块的识别效果有了极大的提升，其在未登录越南语组块识别召回率 R_{UKC} 上提升了 26.89%，在未登录越南语组块类型识别召回率 R_{UKTC} 上提升了 27.68%，这反映了词性信息对未登录越南语组块识别的重要性。

相对于加入词性特征向量的 Bi-LSTM+CRF 模型，在加入多头注意力机制后，模型对未登录越南语组块的识别效果有了进一步提升，其在未登录越南语组块识别召回率 R_{UKC} 上提升了 7.19%，在未登录越南语组块类型识别召回率 R_{UKTC} 上提升了 3.98%，这些数据表明多头注意力机制能够提升模型的泛化能力。

与表 7 和表 8 中所示的识别效果不同，Attention-over-Input-Layer + Bi-LSTM + CRF 模型在对未登录越南语组块的识别效果方面要优于 Bi-LSTM + Multi-Head Attention + CRF 模型，其在未登录越南语组块识别召回率 R_{UKC} 上高于后者 0.02%，在未登录越南语组块类型识别召回率 R_{UKTC} 上高于后者 2.42%。这表明，在输入层融入

注意力机制的方法能够更好地调整词向量与词性特征向量在识别过程中所占的比重，使得模型在遇到未登录越南语组块时能够加大词性特征向量所占的权重。考虑到词性信息在模型对未登录组块的预测上的重要作用，这样可以使得模型更好地处理未登录越南语组块，从而增强模型的泛化能力。

5 结论

针对越南语组块识别任务，本文在前期对越南语组块内部词性构成模式进行统计调查的基础上，发现其内部词性构成模式具有很强的规律性，因此提出了融入注意力机制的思路，从而使得模型能够更多地聚焦于组块的内部构成信息。在 Bi-LSTM+CRF 模型的基础上，本文使用了两种融入注意力机制的方法，一是在 Bi-LSTM 之上加入多头注意力机制，二是在输入层融入注意力机制。实验结果表明，两种融入注意力机制方法都能够有效提升模型对越南语组块的识别效果，且两种方法有着各自的优势和特点。其中，在对越南语组块的整体识别情况上，加入多头注意力机制的方法要好于在输入层融入注意力机制的方法，但在对未登录越南语组块的识别情况上，在输入层融入注意力机制的方法要好于在 Bi-LSTM 之上加入多头注意力机制的方法。

参考文献

[1] Abney S P. Parsing By Chunks[J]. Principle-Based Parsing: Computation and Psycholinguistics, 1991, 1(44):257-278.

[2] Abney S. Partial Parsing via Finite-State Cascades[J]. Natural Language Engineering, 1996, 2(4):399-399.

[3] Ramshaw L A, Marcus M P. Text Chunking using Transformation-Based Learning[C]//Proceedings of the 3rd ACL/SIGDAT workshop. Cambridge, MA: Association for Computational Linguistics, 1995:222-226.

[4] Ngai G , Florian R . Transformation-based learning in the fast lane[C]//Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. Pittsburgh, PA, USA, 2001:1-8.

[5] 张昱琪, 周强. 汉语基本短语的自动识别[J]. 中文信息学报, 2002, 16(6):1-8.

[6] 李珩, 朱靖波, 姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报, 2004, 18(2):2-8.

[7] 徐中一, 胡谦, 刘磊. 基于CRF的中文组块分析[J]. 吉林大学学报(理学版), 2007, 45(3):416-420.

[8] 刘芳, 赵铁军等. 基于统计的汉语组块分析[J]. 中文信息学报, 2000, 14(6):28-32.

[9] 张芬, 曲维光, 赵红艳, 等. 基于CRF和转换错误驱动学习的浅层句法分析[J]. 广西师范大学学报(自然科学版) , 2011, 29(3):147-150.

[10] 李素建. 汉语组块计算的若干研究[D]. 北京:中国科学院计算技术研究所学位论文, 2002.

[11] 李佳. 融入依存句法分析的汉越组块对齐研究[D]. 昆明理工大学. 2018.

[12] Le Minh Nguyen, Huong Thao Nguyen, Phuong Thai Nguyen, et al. An empirical study of Vietnamese noun phrase chunking with discriminative sequence models[C]//Proceedings of Workshop on Asian Language Resources, 2009:9-16.

[13] Thao N T H, Thai N P, Minh N L, et al. Vietnamese Noun Phrase Chunking Based on Conditional Random Fields[C]// International Conference on Knowledge and Systems Engineering. IEEE Computer Society, 2009:172-178.

[14] 郭剑毅, 李佳, 余正涛, 等. 基于约束条件随机场的越南语名词组块识别方法:中国, CN 107797994 A[P]. 2018-03-13.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]// The 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017:1-15.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality[C]//Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013), 2013(2): 3111-3119.

[17] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras and Mark Johnson. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations(NAACL 2018), 2018: 56-60.

[18] Dzmitry Bahdanau, Kyung Hyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate[C]// International Conference on Learning Representations (ICLR 2015), 2015: 1-15.

[19] Marek Rei, Gamal K.O. Crichton, Pyysalo Sampo. Attending to Characters in Neural Sequence Labeling Models[C]// Proceedings of COLING 2016, 2016: 309-318.

[20] 王路路, 艾山·吾买尔, 吐尔根·依布拉音, 买合木提·买买提, 卡哈尔江·阿比的热西提. 基于深度神经网络的维吾尔文命名实体识别研究[J]. 中文信息学报, 2019, 33(3): 64-70.

[21] ZhixingTan, MingxuanWang, JunXie, YidongChen, XiaodongShi. Deep Semantic Role Labeling with Self-Attention[C]// The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018: 4929-4936.

[22] Peter Shaw, Jakob Uszkoreit, Ashish Vaswani. Self-Attention with Relative Position Representations[C]// NAACL 2018.

[23] Patrick Verga, Emma Strubell, Andrew McCallum. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction[C]// NAACL 2018.

[24] Zhouhan Lin, Minwei Feng, Cicero Nogueirados Santos, MoYu, Bing Xiang, Bowen Zhou and Yoshua Bengio. A Structured Self-Attentive Sentence Embedding[C]// ICLR 2017.

[25] 刘艳超. 越南语浅层句法分析方法的研究[D]. 昆明理工大学. 2017.

王闻慧（1995—），硕士研究生，主要研究领域为计算语言学。
E-mail：823467350@qq.com

毕玉德（1967—），通信作者，教授、博士生导师，主要研究领域为计算语言学。
E-mail：biyude@fudan.edu.cn

雷树杰（1995—），硕士研究生，主要研究领域为计算语言学。
E-mail：328037935@qq.com