计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900247

融合词性与声调特征的越南语语法错误检测

张 洲 朱俊国 余正涛
昆明理工大学信息工程与自动化学院 昆明 650500
昆明理工大学云南省人工智能重点实验室 昆明 650500
(zhangzhoukust@foxmail.com)

摘 要 BERT(Bidirectional Encoder Representation from Transformers)预训练语言模型在对越南语分词时会去掉越南语音节的声调,导致语法错误检测模型在训练过程中会丢失部分语义信息。针对该问题,提出了一种融合越南语词性和声调特征的方法来补全输入音节的语义信息。由于越南语的标注语料稀缺,语法错误检测任务面临训练数据规模不足的问题。针对该问题,设计了一种由正确语料生成大量错误文本的数据增强算法。在越南语维基百科和新闻语料上的实验结果表明,所提方法在测试集上取得了最高的 $F_{0.5}$ 和 F_1 分数,证明该方法可提高检测效果,并且随着生成数据规模的扩大,该方法与基线模型方法的效果都得到了逐步提升,从而证明了所提数据增强算法的有效性。

关键词:预训练语言模型;越南语语法错误检测;特征融合;数据增强
中图法分类号 TP391

Incorporating Part of Speech and Tonal Features for Vietnamese Grammatical Error Detection

ZHANG Zhou, ZHU Jun-guo and YU Zheng-tao
School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China

Abstract The BERT pre-trained language model removes the tones of the syllables when segmenting Vietnamese words, which leads to the loss of some semantic information during the training process of grammatical error detection model. To address this problem, an approach combining part of speech and tonal features is proposed to complete the semantic information of the input syllables. Grammatical error detection task confronts the problem of insufficient training data due to the scarcity of labeled Vietnamese data. To address this problem, a data augmentation algorithm is designed to generate a large number of error texts from the correct corpus. Experimental results on Vietnamese Wikipedia and news corpus show that the proposed method achieves the highest $F_{0.5}$ and F_1 score on the test set, which proves it improves the detection performance. Both the proposed method and the baseline model method have a gradual improvement with the increasing scales of the generated data, which proves that the proposed data augmentation algorithm is effective.

Keywords Pre-trained language model, Vietnamese grammatical error detection, Feature fusion, Data augmentation

1 引言

语法错误检测[1](Grammatical Error Detection, GED)指利用计算机程序自动识别一段输入文本中的错误字词及其错误类型。例如,输入一段含有错误的文本"It make me feel optimistic",GED系统检测后输出错误的词"make"和它对应的标签"Form"。这项工作对于拼写检查和文档校对等自然语言处理应用有很大的实用价值。

越南语是一种典型的单音节、不变形、有声调的语言,在语法上主要靠词序和虚词(连词、介词、关联词)来表达语法关系。

由于越南语的这种语言特点,在实际工程应用中容易产生虚词误用、词序错误、音节混淆等语法错误,表1列出了这些错误类型的例子。为了能够纠正这些语法错误,开发一个

到稿日期:2021-09-28 返修日期:2022-03-22
基金项目:国家自然科学基金(62166022,61732005,61866020);云南省重大科技专项计划(202002AD080001,202103AA080015);云南省科技厅面上项目(202101AT070077);云南省人培项目(KKSY201903018)
This work was supported by the National Natural Science Foundation of China(62166022,61732005,61866020), Yunnan Provincial Major Science and Technology Special Plan(202002AD080001,202103AA080015), General Project of Yunnan Provincial Department of Science and Technology (202101AT070077) and People Training Project of Yunnan Province(KKSY201903018).
通信作者:朱俊国(jg.zhu.hit@qq.com)

能自动识别越南语文本中的错误音节及其错误类型的软件很有必要。

表 1 越南语错误类型示例

Table 1 Examples of Vietnamese error types

Error Type	Correct Sentence	Wrong Sentence
Conjunction	nếu cô nổi điên , anh sẽ giải quyết việc ấy vào buổi sáng .	quá cô nổi điên , anh sẽ giải quyết việc ấy vào buổi sáng .
Preposition	các bác đã có ai dùng đến lúc sập nguồn chưa ?	các bác đã có ai dùng đang lúc sập nguồn chưa ?
Coordination	chị ấy và tôi uống cà phê .	chị ấy hay tôi uống cà phê .
Word Order	ư và được sử dụng công nghệ có thể truy cập báo cáo trực tuyến .	ư và được sử dụng công nghệ có thể truy cập báo cáo tuyến trực .
Syllables	đèn trong phòng khá sáng .	đèn trong phòng khá xẳng .

语法错误检测方法一般分为以下 3 种。

(1)基于语法规则的方法。即利用正确语法或者错误语法编写算法或程序来判断文本中出错字词的位置。该方法可解释性强且效果好,但是人工成本极高而且检测的错误类型有限。

(2)基于统计机器学习的方法。最典型的是利用最大熵模型(Maximum Entropy Model)和 *N*-gram[2] 语言模型的方法,该方法通过计算大规模单语语料中的字词出现概率,建立概率模型来获取文本的语言特征。该方法不需要语言学专家提供语法规则,虽然节约了人工成本,但其效果受语料规模和质量的影响,无法得到深层次的语义信息表示。

(3)基于深度学习的方法。该方法利用深度神经网络对句子中的每个词及其对应的标签进行建模,从大规模单语语料中自动学习深层次的语法信息,是目前广泛使用的一种方法。近年来,预训练语言模型[3-4]被应用于各项自然语言处理任务中,最典型的 BERT[4] 预训练语言模型在各项自然语言处理任务中均取得了最优的效果。GED 任务也开始使用 BERT 模型来进一步提升效果,但是 BERT 模型的分词器在对越南语进行分词时会去掉部分音节的声调,分词结果如例 1 所示。

例 1 输入句子:mẹ ơi mẫu thiết kế đầu đẹp| thật nà . |分词结果为:['[CLS]','me','o','##i','mau','thi','##et','ke','đau','đe','##p','that','na','.','[SEP]']。

从例 1 可以看出,BERT 模型的分词器通过字节对编码(Byte Pair Encoding)的方式将某些音节的声调去除,导致在建模时会丢失部分语义信息,因为声调是区分正确和错误音节的重要特征。除此之外,GED 任务的相关研究主要集中于英语和汉语等语种上,针对越南语的相关研究较少且越南语的标注语料稀缺,从而导致越南语的 GED 任务面临训练数据规模不足的问题。

针对上述问题,本文提出了一种基于 BERT 预训练语言模型并且融合越南语词性和声调特征的方法,通过额外的音节特征来加强建模时的语义信息。本文还设计了一种基于语法错误类型的数据增强算法,通过给大量正确的单语语料注入错误来得到训练样本。实验结果表明,所提方法能够有效提升语法错误检测的效果。

2 相关工作

语法错误检测早期的工作都是基于语法规则来设计算法和程序实现的。例如,Macdonald 等[5] 设计了一个自动语法检查器来检测英语文本中的拼写和词汇等错误;Foster 等[6] 提出一种基于错误语法规则的检错和纠错算法,根据错误语法来识别文本中的错误单词。基于规则的方法虽然可解释性强,但是人工成本太高而且不具备通用性,英语的检测方法无法适用于汉语和越南语等其他语种。

为了节省人工成本,提高方法的通用性,基于统计机器学习的方法被应用于 GED 任务。例如,Tetreault 等[7] 训练了一个最大熵分类器来检测英语中的介词错误。由于越南语在书写形式上和英语类似,因此存在和英语类似的拼写错误。*N*-gram 语言模型可以用于越南语的单个或者多个音节的拼写检查(Spell Checking)。例如,Nguyen 等[8] 通过音节的 Bi-gram 得分和滑动窗口机制来计算邻近音节的共现概率,再根据阈值来判断当前音节是否出错;Huong 等[9] 利用大规模越南语语料和 Word2Vector[10] 模型来训练越南语的 *N*-gram 语言模型,在运用滑动窗口机制的同时考虑中心词左右两边出现词的概率,从而提升了检测效果。

随着计算资源的丰富和计算机计算能力的提升,基于深度学习的方法逐渐流行起来,语法错误检测被视为是一个序列标注任务[11],通过端到端的方式对句子中的每个词和标签建模。第一个有重要意义的研究工作是 Rei 等[12] 利用双向 LSTM[13](Long Short-Term Memory,LSTM)网络来构建语法错误检测模型。Pislar 等[14] 提出了一种基于多头注意力(Multi-head Attention)机制的序列标注模型,通过联合学习的方式同时建模整个句子和每个词的表示,从而使模型学到不同层次的语义信息。随着 BERT 模型的广泛应用,如何有效利用 BERT 模型的词嵌入表征(Word Embedding)也被广泛研究。例如,Masahiro 等[15] 在序列标注模型训练时利用了 BERT 模型所有中间层的隐状态表示,进一步提升了英语语法错误检测的效果;Zhang[16] 提出了一个融合图卷积网络[17](Graph Convolution Network,GCN)和 BERT 模型的中文语法错误检测模型,在 2020 年中文语法错误诊断评测任务[18] 中取得了最好的效果。

3 融合词性与声调特征的语法错误检测

3.1 基于多语言 BERT 的序列标注模型

多语言 BERT[19](Multi-lingual BERT,mBERT)模型是由谷歌公司发布的一种特定的预训练语言模型。该模型通过自监督学习的方式,在 100 多种语言的大规模维基百科数据上以掩码语言模型(Masked Language Model)的方式进行训练,得到了 100 多种语言的词嵌入表征,因此,它可以同时支持 100 多种语言的微调,其中就包括越南语。它的基本结构和 BERT 模型并无太大区别,可以适应各种自然语言处理下游任务,其核心组件是一个多层的双向 Transformer[20] 编码器,其网络结构如图 1 所示。

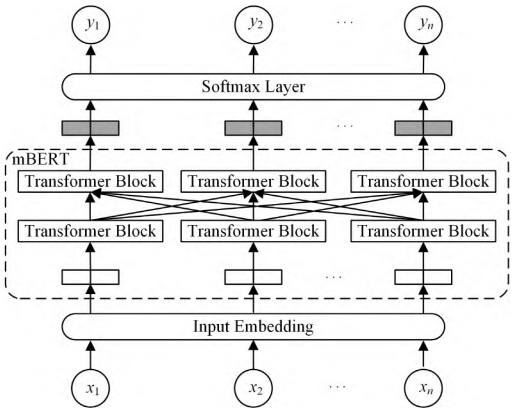

图 1　mBERT 序列标注模型

Fig. 1　mBERT sequence labeling model

本文将 mBERT 模型用于序列标注任务，通过序列标注的方式对越南语文本进行语法错误检测。对于一个输入的句子 $S(x_1, x_2, \cdots, x_n)$，mBERT 序列标注模型首先计算每个音节的词嵌入表征，再将词嵌入表示向量送入一个多层的 Transformer 网络中计算中间层的隐状态表示，最后将最后一层 Transfomer 网络输出的隐状态向量送入一个 softmax 层，计算每个音节对应标签的预测概率序列 $y(y_1, y_2, \cdots, y_n)$。

3.2　越南语的词性和声调特征表示

3.2.1　词性特征表示

一个合理的越南语词性标注集对于语法错误检测模型的训练十分重要。词性标注集过大会增加文本标注的难度，词性标注集过小又无法反映足够的语义信息。考虑到以上两点，本文选用由越南学者开发的一个开源的词性标注工具 Underthesea[1)] 进行越南语的词性标注任务。该工具提供了越南语的 24 种词性标签，本文主要使用其中的 10 种标签，词性标注集如表 2 所列。

表 2　越南语词性标注集

Table 2　Vietnamese POS tagging set

POS Tag	Meaning of Tags	Examples
N	Common Noun	tiếng, thủ đô
V	Verb	viết, đá, đặt
A	Adjective	tốt, xấu, đẹp
P	Pronoun	tôi, hắn, nó, y
R	Adverb	sẽ, đang, xong
E	Preposition	trên, dưới, trong
C	Conjunction	vì vậy, tuy nhiên
Cc	Coordination	và, hoặc, với
CH	Punctuation	, . ; ? ! :
F	Others	á, ạ, ấy, chắ

越南语句子进行分词后，每个音节可以对应 10 种标签中的一种，因此，可以使用一个 10 维的独热码向量（One-hot Embedding）来表示每个音节的词性特征向量。

3.2.2　声调特征表示

越南语是一种带有声调的、基于音节表达语义的语言。越南语的音节有一个稳定的结构，一个音节由 5 个部分组成，即音节＝音节起始音＋中间音＋主元音＋尾音＋声调。例如，音节“chính”＝[ch]＋[ ]＋[元音 i]＋[nh]＋[锐声声调]。其中，音节的声调一共有 6 种，表 3 所列为音节不同声调的示例。

表 3　越南语声调标注集

Table 3　Vietnamese tone tagging set

Tone Tag	Meaning of Tags	Examples
0	No Tone	trong, sao, ma
1	Down Tone	làm, phần, liền
2	Up Tone	áp, máy, có
3	Question Tone	của, đẩy, để
4	Wavy Tone	đã, lỗi, những
5	Dot Tone	sự, mạnh, tuyệt

越南语的每个音节都只有一种声调，每一种声调都有一个固定的字符集编码，可以通过字符集编码来判断某个音节属于哪种声调，然后使用一个 6 维的独热码向量来表示每个音节的声调特征向量。

3.3　特征融合方法

mBERT 模型在对越南语句子进行预处理时会去掉某些音节的声调，导致在模型后续的训练过程中会丢失部分语义信息。为了补全越南语音节的语义信息，本文在基于 mBERT 的序列标注模型的基础上，提出了一种融合词性与声调特征的方法，采用字符编码的方式，将越南语音节的词性和声调特征向量作为额外的信息融入到越南语句子的词嵌入表征中。

输入一个句子 $S(x_1, \cdots, x_i, \cdots, x_n)$，融合词性与声调特征的过程可分为 5 个步骤。1）对输入的单个音节 x_i 进行字符拆分，得到当前音节的每个字符 c_j，再用一个一维的线性表 A 存储每个字符；2）对整个句子进行词性标注，得到每个音节对应的词性标签 p_i，然后在线性表 A 尾部插入 p_i；3）根据当前音节所属的字符集进行声调判断，得到其对应的声调标签 t_i，然后在线性表 A 尾部插入 t_i；4）根据线性表 A 的每个元素进行 one-hot 编码，分别得到每个音节对应的字符特征向量 $\boldsymbol{C}_i$，词性特征向量 $\boldsymbol{P}_i$ 和声调特征向量 $\boldsymbol{T}_i$；5）以上向量按照最大维度进行填充操作，确保它们的维度一致，最后将词性特征向量和声调特征向量进行拼接操作，得到每个音节对应的特征向量 $\boldsymbol{F}(\boldsymbol{x}_i)$。整个计算流程以及模型的网络结构如图 2 所示。

整个特征融合的计算过程可由式(1)－式(3)来表示：

$$\boldsymbol{C}_i = \sum_{j=1}^{M}\sum_{k=1}^{D_C} f_k(c_j),\ \boldsymbol{P}_i = \sum_{k=1}^{D_P} f_k(p_i),\ \boldsymbol{T}_i = \sum_{k=1}^{D_T} f_k(t_i) \tag{1}$$

$$\boldsymbol{C}_i, \boldsymbol{U}_i, \boldsymbol{V}_i = Pad(\boldsymbol{C}_i, \boldsymbol{P}_i, \boldsymbol{T}_i) \tag{2}$$

$$\boldsymbol{F}(\boldsymbol{x}_i) = \boldsymbol{C}_i \oplus \boldsymbol{U}_i \oplus \boldsymbol{V}_i \tag{3}$$

其中，$f_k(x)$ 表示第 k 维的 one-hot 编码，D_C 表示字符特征维度，D_P 表示词性特征维度，D_T 表示声调特征维度，M 表示每个音节对应的字符总数，函数 $Pad(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ 表示对输入向量按照最大维度进行填充操作，$\oplus$ 表示向量的拼接操作。

得到特征向量后，将其和 mBERT 模型计算得到的每个音节对应的隐状态向量进行再次拼接，得到整个句子的隐状态表示。再将整个句子的隐状态表示输入到一个双向的

1) https://github.com/undertheseanlp/underthesea

LSTM 网络中，最后经过一个 softmax 层计算得到每个音节对应标签的预测概率。整个计算过程如式(4)—式(7)所示：

$$\boldsymbol{X}_i=\boldsymbol{F}(\boldsymbol{x}_i)\oplus\boldsymbol{h}_i^L \tag{4}$$

$$\overrightarrow{H_i}=\mathrm{LSTM}(\boldsymbol{X}_i,\overrightarrow{H_{i-1}}),\overleftarrow{H_i}=\mathrm{LSTM}(\boldsymbol{X}_i,\overleftarrow{H_{i-1}}) \tag{5}$$

$$\boldsymbol{H}_i=\overrightarrow{H_i}\oplus\overleftarrow{H_i} \tag{6}$$

$$Y_i=\mathrm{softmax}(\boldsymbol{W}_O\boldsymbol{H}_i+\boldsymbol{b}_O) \tag{7}$$

其中，$\boldsymbol{h}_i^L$ 表示 mBERT 模型最后一层的输出，$\boldsymbol{H}_i$ 表示双向 LSTM 网络的隐状态输出，$\boldsymbol{W}_O$ 表示最后输出层的权重矩阵参数，$\boldsymbol{b}_O$ 表示输出层偏置项矩阵参数。模型训练时使用交叉熵损失对模型参数进行优化，其计算式如式(8)所示：

$$loss=-\frac{1}{N}\sum_i\sum_{c=1}^{C}g_c(x_i)\log(P(Y_i\mid\boldsymbol{H}_i)) \tag{8}$$

其中，N 表示样本总数；C 表示类别总数；$g_c(x)$ 表示符号函数，如果观测样本 x_i 的真实类别为 c 则取 1，否则取 0；Y_i 是最后 softmax 层计算得到的每个音节标签类别的预测概率。

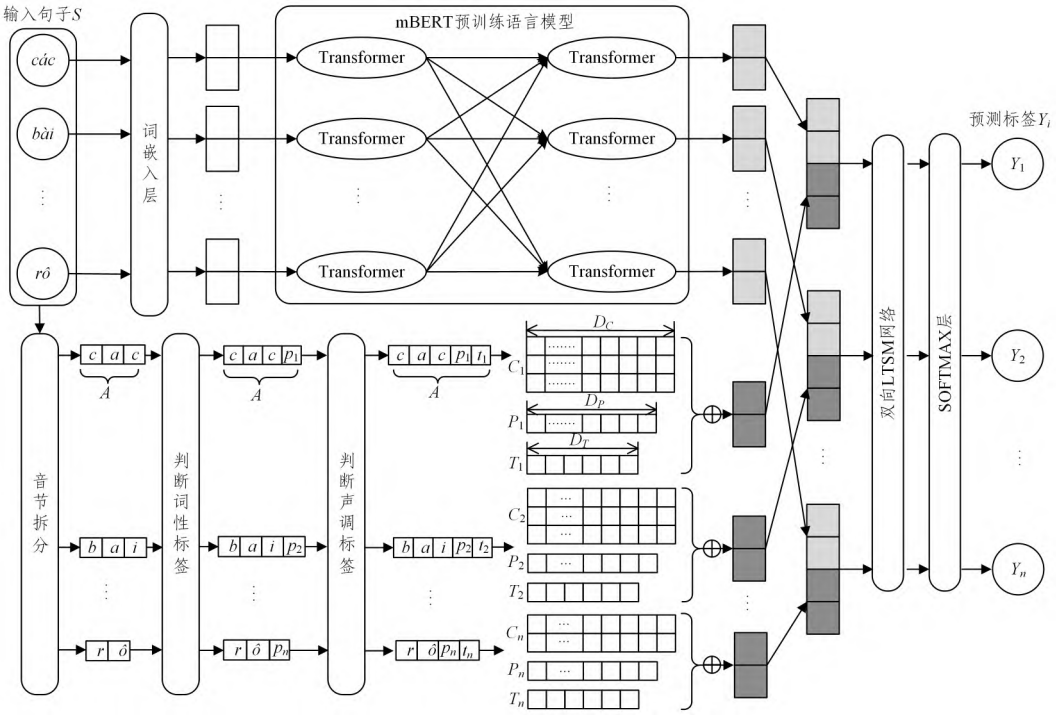

图 2 越南语语法错误检测模型结构图

Fig. 2 Architecture of Vietnamese grammatical error detection model

4 基于错误类型的错误数据生成

4.1 音节混淆集词典构建

本文根据越南语学习教材《基础越南语》[21]中的音节混淆规则来构建音节混淆集词典。首先，对大规模越南语单语语料进行预处理，去除特殊字符和重复句子；然后，对单语语料进行音节切分，以每个音节为键(key)，其出现的词频为值(value)来构建音节词典；最后，对音节词典按照字典序排序，首字母相同的音节为一组，再按照音节混淆规则进行重排序，将易混淆的音节组成同一个集合得到最终的音节混淆集词典。整个混淆集词典的构建过程如图 3 所示。

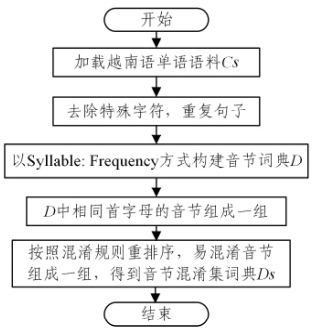

图 3 越南语音节混淆集词典构建流程

Fig. 3 Process of building Vietnamese syllable confusion set dictionary

4.2 错误数据生成算法

本文针对越南语中的虚词误用、词序错误和音节混淆等错误类型，通过替换和改变词序等方式向正确语料中注入错误。在生成音节混淆错误时，利用 4.1 节中构建的音节混淆集词典提供候选项，越南语语法错误数据生成算法如算法 1 所示。

算法 1 错误数据生成算法

输入：单语语料 Cs，词典 D，音节混淆集词典 Ds，出错概率 Pe%，替换概率 Pu%，词序打乱概率 Pm%，音节混淆概率 Ps%

输出：含有错误的语料 Ce

```
1. Begin:
2.     for all sentences S in Cs do
3.     L=POS_tag(S)/*对句子 S 进行词性标注*/
4.       生成 1100 的随机数 N
5.       if N<Pe then/*不产生错误*/
6.          continue
7.       end if
8.       生成 1100 的随机数 M
9.      if M>Pu do/*进行词替换*/
10.        L 中所有连词加入词集 Wc
11.        L 中所有介词加入词集 Wp
12.        L 中所有关联词加入词集 Wr
13.        if Wc !=NULL do
14.           用 D 中的词 k1 替换 Wc 中的词 w1
15.        end if
16.        if Wp !=NULL do
```

```
17.          用 D 中的词 k2 替换 Wp 中的词 w2
18.        end if
19.        if Wr !=NULL do
20.          用 D 中的词 k3 替换 Wr 中的词 w3
21.      end if
22.      else if M>Pu+Pm do/*进行词乱序*/
23.          L 中选择一个词 w4
24.          if length(w4)>2 do
25.          打乱 w4 的词序
26.          end if
27.        else if M>Pu+Pm+Ps do/*音节混淆*/
28.          从 S 中选择一个音节 s
29.          用 Ds 中的易混淆词 s′替换 s
30.        end if
31.        将错误句子 S′加入语料 Ce 中
32.    end for
33. End
```

5　实验与分析

5.1　实验数据及其预处理

目前还没有公开的越南语标注数据集可用于越南语的语法错误检测任务。为了获取用于实验的数据集，本文从越南语版本的维基百科网站[1)]上利用 Python 爬虫技术爬取了 702 MB 的通用领域文本数据，又从越南快讯官方门户网站[2)]和越南青年报[3)]等新闻门户网站上爬取了 300 MB 的新闻语料，然后对得到的文本数据进行预处理。

首先，过滤掉原始文本中包含特殊字符和非越南语字符的句子，删除音节数超过 40 的句子和音节重复率超过 90% 的句子，得到用于实验的 1 088 146 句正确的越南语句子。再从预处理后的句子中随机采样 12 500 个正确句子，将这些句子进行划分，选取其中 10 000 句作为训练集的一部分，1 500 句作为验证集的一部分，剩余 900 句作为测试集的一部分。然后，利用第 4.2 节中的错误数据生成算法对语料库中剩余的句子进行造错，并采用 BIO 标注方式对每个音节进行标注，选取其中的 15 000 句并入训练集，2 000 句并入验证集，1 100 句并入测试集。实验数据统计信息如表 4 所列。

表 4　实验数据统计信息

Table 4　Experimental data statistics

		Train Set	Dev Set	Test Set
Right Sentences		10 000	1 500	900
Wrong Sentences		15 000	2 000	1 100
Tokens		609 829	85 763	48 097
Average Length		25.64	24.50	24.04
Error Types/count	Conjunction	3 762	410	228
	Preposition	4 461	505	283
	Coordination	4 362	545	306
	Word Order	5 629	668	377
	Syllable	4 946	615	351

5.2　模型参数设置

本文选择 bert-base-multilingual-cased 模型参数作为 mBERT 序列标注模型的词嵌入初始化参数，该模型由 12 层 Transformer 编码器组成，每个 Transformer 编码器包含 12 个注意力头，隐层大小为 768，词向量维度为 512。整个模型在一块 RTX2080ti 上训练 20 个轮次。在训练微调阶段，设置训练数据批次大小为 32，学习率设置为 1×10^{-5}，预热比例为 0.1，句子最大长度为 128，使用 Adam 优化器对模型进行优化。

5.3　评价指标

语法错误检测任务的评价指标与分类任务的评价指标相同，通过计算分类标签的精确率（Precision，P）、召回率（Recall，R）和 F 值（F-score）来评价模型。值得注意的是，由于忽略一处错误的影响比将正确标签误分类为不正确的标签要小，该任务一般使用 $F_{0.5}$ 值来衡量模型的整体性能，因为它在精确率上的权重更大，各评价指标的计算式如下所示：

$$P=\frac{\text{预测正确的标签数}}{\text{预测出的标签数}}\times100\% \tag{9}$$

$$R=\frac{\text{预测正确的标签数}}{\text{数据集中的标签总数}}\times100\% \tag{10}$$

$$F_\beta=(\beta^2+1)PR/(\beta^2P+R) \tag{11}$$

5.4　基线方法

为了验证本文模型在实验数据集上的有效性，将本文方法将与以下 3 个方法进行对比。

Bi-LSTM[12]：基于双向 LSTM 网络的序列标注模型，实验中对比了使用 Word2Vector[10] 和 GloVe[22] 两种词嵌入表征的效果。

IDCNN[23]：一种基于膨胀卷积（Dilated Con-volution）网络的序列标注模型，实验中同样对比了使用 Word2Vector[10] 和 GloVe[22] 两种词嵌入表征的效果。

MHAL（Multi-head Attention Labeling）[14]：基于多头注意力机制的序列标注模型，通过联合学习的方式同时对单词和整个句子进行建模。

5.5　对比实验结果分析

表 5 列出了本文方法和基线方法在验证集和测试集上的结果对比，其中验证集的结果是训练过程中模型表现最好的。

表 5　不同方法在验证集和测试集上的比较

Table 5　Performance comparison of different methods on validation set and test set

（单位：%）

Data Set	Methods	P	R	$F_{0.5}$	F_1
Dev Set	Bi-LSTM+Word2Vector	53.91	41.71	50.93	47.03
	Bi-LSTM+GloVe	59.62	39.08	53.95	47.21
	IDCNN+Word2Vector	59.25	30.11	49.65	39.93
	IDCNN+GloVe	60.76	32.52	51.77	42.37
	MHAL（Pislar et al. 2020）	67.87	54.22	64.36	60.28
	mBERT+LSTM+feature（proposed）	**69.63**	**75.39**	**70.63**	**72.39**
Test Set	Bi-LSTM+Word2Vector	56.31	44.79	53.55	49.89
	Bi-LSTM+GloVe	60.72	44.53	56.61	51.38
	IDCNN+Word2Vector	61.45	32.31	52.06	42.35
	IDCNN+GloVe	62.43	33.98	53.47	44.01
	MHAL（Pislar et al. 2020）	64.04	57.99	62.57	60.86
	mBERT+LSTM+feature（proposed）	**70.49**	**75.49**	**71.36**	**72.91**

1) http://www.wikipedia.org

2) http://www.vnexpress.net

3) https:// thanhnien.vn

从表 5 的实验结果可以看出，与其他方法相比，本文方法在测试集和验证集上均取得了最好的结果。在基线方法中，MHAL 的效果最好。本文方法在测试集上的精确率、召回率、$F_{0.5}$值和 F_1值比 MHAL 方法分别高出 6.45%，17.15%，8.79%和 12.05%，其中召回率提升明显，是 F 值提升的主要原因，表明本文方法能够检测到基线方法没有检测到的错误。

5.6 消融实验结果分析

为了验证融合词性和声调特征的方法能够改进原先的 mBERT 序列标注模型的错误检测效果，本文进行了一组消融实现来验证不同组件对 mBERT 模型的影响。本文一共进行了 4 次不同的实验：第一次实验只使用 mBERT 模型进行训练，没有在编码端融入任何的额外特征；第二次实验在 mBERT 模型的输出层加上 LSTM 网络后进行模型的训练；第三次实验在 mBERT 模型的编码端融合额外的特征后进行模型的训练；第四次实验在 mBERT 模型的编码端融合额外的特征，同时在模型的输出层加上 LSTM 网络。表 6 列出了采用不同方式训练在测试集上的结果。

表 6 测试集上 mBERT 模型使用不同组件的结果

Table 6 Results of mBERT model using different components on test set

（单位：%）

Methods	P	R	$F_{0.5}$	F_1
mBERT	65.21	72.67	66.49	68.74
mBERT+LSTM	65.78	71.69	66.82	68.61
mBERT+feature	67.25	72.24	68.19	69.66
mBERT+LSTM+feature	**70.49**	**75.49**	**71.36**	**72.91**

由表 6 的前 3 行可以看出，双向 LSTM 网络和额外的字符级特征嵌入对 mBERT 模型的作用效果明显不同。只增加双向 LSTM 组件时，mBERT 模型的精确率提升了 0.57%，召回率下降了 0.98%，$F_{0.5}$值和 F_1值的变化情况与之吻合且变化幅度较小。只增加额外的字符级特征嵌入时，mBERT 模型的 $F_{0.5}$值和 F_1值分别提升了 1.7%和 0.92%，证明额外的字符级特征嵌入对 mBERT 序列标注模型效果的提升更加明显。

5.7 数据增强方法效果验证

为了验证 4.2 节中提出的数据增强算法的有效性，本文分别在 25K，50K，75K，100K 和 125K 的训练数据规模下，在测试集上比较了 MHAL 基线方法和本文方法的效果，图 4 给出了实验结果的示例图。

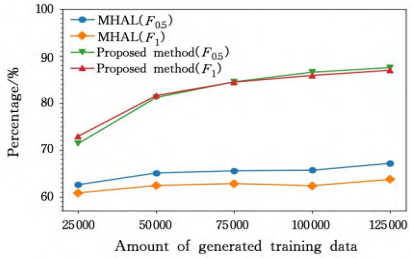

图 4 不同训练数据规模下模型性能的变化情况

Fig. 4 Changes of model performance with different training data scales

从图 4 可以看出，随着训练数据规模的增大，本文方法和基线方法的 $F_{0.5}$值和 F_1值都在逐步提升，但是基线方法的提升较为缓慢。当训练数据规模从 25K 增加到 50K 时，本文方法的 $F_{0.5}$值和 F_1值提升明显。随着训练数据规模的继续增大，本文方法和基线方法的 F 值的变化逐步趋缓，但总体上仍然呈现上升的趋势，由此可以证明本文提出的错误数据生成算法是有效的。

结束语 语法错误检测在自然语言处理的实际工程应用中具有重要价值，如何有效地利用 BERT 等预训练语言模型成为近年来的研究热点，但是 BERT 模型在处理越南语句子时会去掉部分音节的声调，导致部分语义信息丢失。本文针对这一问题，提出了一种融合词性与声调特征的方法，用额外的字符级编码特征来增强模型学到的语义信息；针对越南语标注语料较少的问题，本文提出一种错误数据生成算法来扩充训练集的规模。实验结果表明：与基线方法相比，本文方法在维基百科（越南语版）和新闻语料上取得了最高的 $F_{0.5}$和 F_1分数，而且随着训练数据规模的增大，基线方法和本文方法的效果都在逐步提升，证明了本文提出的方法和算法是有效的。在接下来的工作中，我们会考虑如何将模型与语法错误纠正任务相结合，将检测到的错误予以纠正。

参 考 文 献

[1] MADI N, AL-KHALIFA H S. Grammatical Error Checking Systems: A Review of Approaches and Emerging Directions [C]// Proceedings of the Thirteenth International Conference on Digital Information Management. 2019: 142-147.

[2] YIN C, WU M. Survey on N-gram Model [J]. Computer Systems and Applications, 2018, 27(10): 33-38.

[3] PETERS M, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 2227-2237.

[4] 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. 2019: 4171-4186.

[5] MACDONALD N, FRASE L, GINGRICH P, et al. The Writer's Workbench: Computer Aids for Text Analysis[J]. IEEE Transactions on Communications, 1982, 30(1): 105-110.

[6] FOSTER J, VOGEL C. Parsing Ill-Formed Text Using an Error Grammar[J]. Artificial Intelligence Review, 2004, 21(3): 269-291.

[7] TETREAULT J R, CHODOROW M. The Ups and Downs of Preposition Error Detection in ESL Writing[C]// Proceedings of the 22nd International Conference on Computational Linguistics. 2008: 865-872.

[8] NGUYEN P H, NGO T D, PHAN D A, et al. Vietnamese Spelling Detection and Correction Using Bi-gram, Minimum Edit Distance, SoundEx Algorithms with Some Additional Heuristics [C]// Proceedings of 2008 International Conference on Research, Innovation and Vision for the Future. IEEE, 2008: 96-102.

[9] HUONG N,DANG T T,NGUYEN T T,et al. Using Large N-gram for Vietnamese Spell Checking[C]// Proceedings of the Sixth International Conference on Knowledge and Systems Engineering. 2015:617-627.

[10] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estimation of Word Representations in Vector Space[C]// Proceedings of International Conference on Learning Representations. 2013:1-12.

[11] REI M. Semi-supervised Multitask Learning for Sequence Labeling[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:2121-2130.

[12] REI M,YANNAKOUDAKIS H. Compositional Sequence Labeling Models for Error Detection in Learner Writing[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:1181-1191.

[13] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.

[14] PISLAR M,REI M. Seeing Both the Forest and the Trees: Multi-head Attention for Joint Classification on Different Compositional Levels[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:3761-3775.

[15] MASAHIRO K,MAMORU K. Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection[J]. Computación y Sistemas,2019,23(3):883-891.

[16] ZHANG J H. Combining GCN and Transformer for Chinese Grammatical Error Detection[J]. arXiv:2105.09085,2021.

[17] KIPF T N,WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// Proceedings of 5th International Conference on Learning Representation. 2017:1-14.

[18] RAO G Q,YANG E H,ZHANG B L. Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnosis [C]// Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. 2020:25-35.

[19] PIRES T,SCHLINGER E,GARRETTE D. How Multilingual is Multilingual BERT? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4996-5001.

[20] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is All You Need[C]// Proceedings of Advances in Neural Information Processing Systems. 2017:5998-6008.

[21] TAN Z C,XU F Y,LIN L. Basic Vietnamese[M]. Beijing: World Publishing Corporation,2013:3-95.

[22] PENNINGTON J,SOCHER R,MANNING C D. GloVe:Global Vectors for Word Representation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[23] STRUBELL E,VERGA P,BELANGER D,et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017:2670-2680.

ZHANG Zhou, born in 1992, postgraduate. His main research interests include natural language processing and grammatical error correction.

ZHU Jun-guo, born in 1982, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include natural language processing and machine translation.

(责任编辑:杨雪敏)