文章编号：1003-0077(2023)11-0038-11

融合字符与词性特征的泰语文本语法错误检测

施灿镇[1,2]，朱俊国[1,2]，余正涛[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500)

摘 要：文本语法错误检测与纠正旨在自动识别并纠正文本中的语法错误。与汉语、英语等语言不同，该任务在一些泰语语言的文本上受制于数据规模问题，仍然只能针对简单规则进行识别和校正。该文结合相应的语言学及错误类型特点，基于人工启发式规则，利用单语数据构建了一定规模的泰语文本语法错误检测与纠正语料库。基于该语料库，该文提出一种融合语言学特征的泰语文本语法错误检测方法，在多语言 BERT 序列标注模型的基础上融合字符、词与词性的深层语义表达。实验结果表明，该文方法的错误检测性能比仅依赖于多语言 BERT 的基线模型提升了 1.37% 的 F_1 值，并且模型性能会随着训练数据规模的增大而提高，证明了该文语料库构建方法的有效性。

关键词：文本语法错误检测；泰语；语料库；特征融合

中图分类号：TP391 **文献标识码**：A

Combining Character and Part-of-Speech Features for Thai Text Grammar Error Detection

SHI Canzhen[1,2], ZHU Junguo[1,2], YU Zhengtao[1,2]

(1. School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Key Laboratory of Artificial Intelligence in Yunnan Province,
Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Text grammatical error detection and correction aims to automatically identify and correct grammatical errors in text. In contrast to Chinese, English and other languages, this task for Thai texts remains rule based method due to the limited data. This paper constructs a large-scale Thai text grammatical error detection and correction corpus based on artificial heuristic rules using monolingual data. Based on this corpus, this paper proposes a grammatical error detection method of Thai text that integrates linguistic features. It integrates the deep semantic expression of characters, words and parts of speech via the multilingual BERT. The results show that the proposed method improves by 1.37% F_1 value than the baseline model that only relies on multilingual BERT.

Keywords: text grammatical error detection; Thai; corpus; feature fusion

0 引言

文本语法错误检测与纠正旨在自动识别并校正文本中的语法错误且不改变其原始语义，用于提高文本的可读性，规范语句表达[1-2]，在搜索引擎、语音识别后处理、机器翻译以及具有评价和反馈功能的计算机辅助语言学习系统中有着广泛的应用[3]。近年来，随着泰国旅游业的飞速发展，越来越多不同语言和知识背景的人都有兴趣将泰语作为一门外语进行学习，而这些第二语言学习者在学习过程中难免会因为语言和文化的差异产生语法错误。而现有的

收稿日期：2022-09-09 **定稿日期**：2022-11-15

基金项目：国家自然科学基金(62166022，61732005)；云南省科技厅面上项目(202101AT070077)；云南省人培项目(KKSY201903018)

泰语文本语法错误检测与纠正的相关工作主要集中在光学字符识别系统的后处理上，还停留在简单的拼写纠错层面[4-6]，缺少对其他语法错误类型（如用词不当错误、语序错误、停顿错误等）的研究，因此对泰语文本语法错误检测与纠正任务的研究具有重要意义。

目前，文本语法错误检测（Grammatical Error Detection，GED）任务在中文和英文等资源丰富的语言中取得了显著进展，与之相对应的，也有一些公开的标注数据集，例如，CoNLL-2014[7]，CGED[8-11]等。然而，公开可用的标注数据集依旧是相对稀少的，数据匮乏问题成为限制该任务发展的主要原因，特别是对于泰语等低资源语言而言，这个问题尤为突出。为了缓解这个问题，有许多研究者提出了不同的数据增广方法，大致可分为两类：一类是基于规则的方法，通过在原句中随机引入噪声来构造错误句子。另一类是基于模型的方法，例如，使用机器翻译模型对其他语言进行回译。与前人方法不同的是，本文首先结合泰语语言特点对泰语文本中的错误类型进行归纳与定义，然后针对不同的错误类型人工制定相应的算法规则，最后利用单语数据生成语料库。

文本语法错误检测通常被视为序列标注任务，使用神经网络建模词与错误类型标签的关系。但是传统序列标注模型缺乏对语言特点的考虑，致使其在识别特定语言中的特有错误时性能不佳。本文在传统序列标注模型的基础上融合泰语语言学特征，增强模型的语义信息，从而提高语法错误检测的性能。实验结果表明，本文方法的语法错误检测性能与仅依赖于多语言 BERT 模型的方法相比提升了 1.37%的 F_1 值。此外，在用词不当错误、词冗余错误和停顿错误的检测上分别提升了 0.82%，1.82%和 3.99%的 F_1 值，并且模型性能会随训练数据规模的增加而提高。

1　相关工作

神经网络模型的训练通常需要大规模的训练数据，而人工构建语料库成本昂贵且耗时，因此有许多学者研究如何模拟真实场景构建伪数据来扩增训练数据，主要方法大致可分为两类：一类是基于规则的方法。Wang 等人[12]借鉴去噪自编码器的思想，首先通过规则（包括随机插入、删除、替换等）对文本增噪，以此生成包含各种错误的句子，再通过序列到序列模型试图恢复原始句子。第二类是基于模型的方法。Lichtarge 等人[13]借助桥梁语言进行往返翻译，构建了大规模的训练数据。Zhou 等人[14]借助神经机器翻译模型显著好于统计翻译模型的特点，用前者生成质量较好的句子，用后者生成质量较差的句子，构成训练句对。但是基于规则的方法生成的训练数据无法完全覆盖真实场景下的错误类型，并且生成的错误和语言学习者所犯的错误差异较大，而基于模型的方法生成的句子可能具有歧义，或者不包含语法错误，又或者与原句在语句的构成上差距太大。针对以上两个问题，本文尝试采用基于人工启发式规则的方法来构建泰语文本纠错语料库，既能规避错误类型覆盖不全的问题，也不会引入过多具有歧义的数据。

国内外检测文本语法错误的主流是序列标注模型，即输入一段文本，输出文本中每个词对应的错误类型标签。Kaneko 等人[15]提出多头多层注意力模型，使用多头注意力机制提取 BERT 模型所有中间层的隐藏表示，从而提升语法错误检测的效果。Pislar 等人[16]使用多头注意力机制将单个单词和完整句子之间的表示联系起来，联合学习序列标注和句子分类任务，其中每个注意力头都被作为一个序列标注器，用来检测每个词对应的标签。Yuan 等人[17]首先将该任务视为一个二进制序列标注任务，通过微调 ELECTRA[18]模型来实现，然后使用 ERRANT框架[19]将二进制检测结果派生到多类错误类型，从而提升语法错误纠正的效果。可见，学者们主要从任务特点的方向开展研究，而结合语言特点开展的工作相对较少，且面向泰语的相关工作更是少之甚少。本文结合泰语文本先由连续的泰文字符按一定组合规则排列成词再构成句子的特点，在传统序列标注模型的基础上融合字符与词性特征，进一步提升语法错误检测性能。

2　方法与模型

2.1　泰语语法错误类型定义与标注方案

泰语属于汉藏语系，和汉语一样没有明确指定单词边界，在进行更深入的分析之前通常需要进行分词。从语法上看，泰语和汉语一样，都属于孤立语，缺乏严格意义上的形态标记和变化，主要靠词序、虚词等语法手段表现词与词的关系或其他语法意义[20]。泰语的基本语序与汉语相同，都是“主-谓-

宾”(Subject-Verb-Object, SVO)结构,区别在于泰语中的修饰语在被修饰语之后,简而言之,泰语形容词应放在名词之后,副词放在动词之后,例如,汉语的“这幅画真美”,在泰语中的语序为“画幅这美真”,因而容易产生语序错误。

与属于语素文字的汉字不同,泰文属于音节音位文字,是以辅音字母为主体、元音以附加符号形式标出的一种表音文字。词汇主要由音节组成,包括单音节词和多音节词,而一个音节一般由辅音、元音和声调等多个字符组成,不同的声调有区分词汇和语法的作用,其中有44个辅音、32个元音和5个声调(第一声调不标符号),如表1所示。

表1 泰文辅音、元音及声调字符

类型	字符
辅音	ก จ ด ต ฎ ฏ บ ป อ ข ฃ ฉ ฐ ถ ผ ฝ ศ ษ ส ห ค ฅ ฆ ช ซ ฌ ญ ฑ ฒ ณ ท ธ น พ ฟ ภ ม ย ร ล ง ว ฬ ฮ
元音	-ะ -า -ิ -ี -ึ -ื -ือ -ุ -ู เ-ะ เ- แ-ะ แ- โ-ะ โ- เ-าะ -อ เ-อะ เ-อ -ัวะ -ัว เ-ียะ เ-ีย เ-ือะ เ-ือ ฤ ฤๅ ฦ ฦๅ -ำ ไ- ใ- เ-า
声调	่ ้ ๊ ๋

需要注意的是,泰语辅音虽然有44个,但实际上只发21个音,并且从表1中可以很容易看出在辅音字符中许多字符的字形是相近的,这极易引起拼写错误,这也是最常见的泰语文本语法错误。除此之外,泰语中存在许多意思相近但用法或词性不同的单词,例如,“ซัก”和“อาบ”,两者都有“洗”的意思,但前者仅与衣物、布料等词汇一起使用,后者仅适用于洗澡场景。并且,受泰国国家制度、宗教文化的影响,泰语分为世俗用语、王族用语和僧侣用语三种,同样是表达一个意思,但三种用语使用的词汇可能各有不同,由此就容易产生用词不当错误,这与中文中的用词不当错误有本质上的区别。另外,泰语中没有明确指定单词边界,一般使用空格(停顿)表示短语或句子的结束,相当于中文里的顿号、逗号和句号。停顿的位置不同,语义也会有所不同,因此极易引发歧义现象,例如“แก้กระหาย”的意思为“把雀斑消掉”,而“แก้กระหาย”的意思为“解渴”。

和汉语一样,泰语在实际使用过程中也存在词汇的冗余和缺失错误,有时虽然可以起到简化表达或强调语句成分的作用,但是在语法上是不规范的。泰语文本词冗余错误主要表现在使用多余的修饰语修饰被修饰语,或者是介词、连词等虚词的冗余。词缺失错误主要表现为动词、副词等实词的缺失,以及连词、介词等虚词的缺失。

综上,本文将泰语文本中存在的错误归为六类,分别是拼写错误、词冗余错误、词缺失错误、用词不当错误、语序错误以及停顿错误(分为冗余和缺失两种)。标注方案遵循BIO规则[21],“B”表示文本语法错误的起始,“I”表示文本语法错误的中间或结尾部分,“O”表示对应单词没有语法错误,“C”表示拼写错误,“R”表示词冗余错误,“M”表示词缺失错误,“S”表示用词不当错误,“W”表示语序不当错误,“PM/PR”表示停顿错误(缺失/冗余)。错误类型及标注示例如表2所示,其中,“_”表示停顿(空格),“[]”中为文本语法错误发生的位置。

表2 泰语文本语法错误类型及标注示例

错误类型	示例
拼写错误(C)	อย่า/ลืม/ซื้อ/ไข่/ไก่/กับ/[เมย→เนย]/กระป๋อง/หนึ่ง O O O O O O O B-C O O O 别忘了买一些鸡蛋和一罐黄油。
词冗余错误(R)	เพื่อ/ให้/สามารถ/สร้าง/ประสบการณ์/ผู้/[การ]/ใช้/ที่/น่า/พอใจ/ด้วย/เดสก์ท็อป O O O O O O O B-R O O O O O O O 能够使用桌面创建令人满意的体验。
词缺失错误(M)	ยิ่ง/มร/สุม/เศรษฐกิจ/หนัก/[ไร]/_/โอกาส/พลาด/ก็/มี/สูง/มาก/เท่า/นั้น O O O O B-M O O O O O O O O O O 经济动荡越大,出错的机会就越大。
用词不当错误(S)	เธอ/[อาบ→ล้าง]/เท้า/ให้/แม่/เป็น/ครั้ง/แรก/ใน/ชีวิต O B-S O O O O O O O O O 这是她有生以来第一次给妈妈洗脚。
语序不当错误(W)	พ่อแม่/มี/ความ/ปิติ/กับ/ความ/สำเร็จ/[ลูก/ของ→ของ/ลูก] O O O O O O O O B-W I-W 父母为孩子的成功感到高兴。
停顿错误(PM/PR)	นั่น/เนื้อ/[_]/สุนัข/ผม/ชอบ/กิน(停顿缺失) O B-PM O O O O O นั่น/เนื้อ/_/สุนัข/[=]/ผม/ชอบ/กิน(停顿冗余) O O O O B-PR O O O 那是肉,我的狗爱吃。

2.2 泰语文本语法错误检测与纠正语料库构建

网络上现有的泰语文本语法纠错语料资源有限,远不足以用于训练神经网络模型,而采用人工的

方式进行语料收集和注释是十分昂贵的。针对这个问题，在第 2.1 节定义的错误类型的基础上，本文设计出一种基于人工规则自动构建泰语文本纠错语料库的方法，对不同错误类型设计有针对性的错误构建规则。

其中对于拼写错误和词选择错误的构建，需要先分别构造字符混淆集和近义词混淆集，前者使用手工构造，后者通过爬取电子词典并筛选构造而成，以下给出定义：

字符混淆集　分为字形相近混淆集和字音相近混淆集，分别记为 $C_s=\{\text{sset}_1,\text{sset}_2,\cdots\}$ 和 $C_p=\{\text{pset}_1,\text{pset}_2,\cdots\}$，其中，元素 $\text{sset}_i=\{c_1,c_2,\cdots\}$，$\text{sset}_i$ 中每一个字符互为其他字符的混淆字符，pset_j 也类似如此。例如，$\text{sset}_i=\{$“ผ”,“ฝ”,“พ”,…$\}$，字符形态相近；$\text{pset}_j=\{$“ข”,“ฃ”,“ค”,…$\}$，国际音标均为“kh”。

近义词混淆集　记为 $U=\{\text{uset}_1,\text{uset}_2,\cdots\}$，其中元素 $\text{uset}_k=\{u_1,u_2,\cdots\}$，$\text{uset}_k$ 中每个词互为其他词的混淆。例如，$\text{uset}_k=\{$“ล้าง”,“อาบ”,“ฟอก”,…$\}$，均有“洗”的意思。

在构建语料库之前，本文首先从网络上收集一定数量的单语语料，使用开源工具 PyThaiNLP① 对文本进行预处理，得到泰语文本语法错误原始语料库，记为 S_o。然后编写算法规则对 S_o 中的每个句子随机生成错误：对于拼写错误，本文采用三种策略生成，分别是字形相近字符替换、发音相近字符替换以及基于编辑距离的随机替换；对于词冗余错误，借助 2-gram 词表在句子中随机插入单词；对于词缺失错误，根据词性标签随机删除句子中的动词、连词、介词和助词等；对于用词不当错误，使用近义词混淆集进行随机替换；对于词序错误，随机打乱句子中连续的 2～5 个词；对于停顿错误，根据词性标签随机插入或删除停顿。具体算法流程如算法 1 所示，其中分词和词性标注均使用 PyThaiNLP 处理，词性标注选用 Sornlertlamvanich 等人[22] 提出的标注方案，包含 47 种词性标签，2-gram 词表来自泰国国家语料库②。

需要注意的是，考虑到实际场景下文本语法错误的复杂性，同时保证模型具有一定的鲁棒性和泛化能力，我们以概率 P_e 令部分语料包含文本语法错误，以概率 P_q 令含有语法错误的句子包含一到两个错误。

算法 1. 泰语文本语法纠错语料自动生成算法

输入：原始语料 S_o，字符混淆集 C_s、C_p，2-gram 词表 B，近义词混淆集 U，出错概率 P_e，单句仅产生一处错误的概率 P_q，拼写错误概率 P_c，词冗余错误概率 P_r，词缺失错误概率 P_m，用词不当错误概率 P_s，词序错误概率 P_w，停顿(空格)错误概率 P_p；

输出：泰语文本语法纠错语料库 thGECC。

```
1. for each s in S_o
2.    T, P←process(s); /* 进行分词、词性标注并分别存入列表 T 和 P 中 */
3.    if random( P_e ) is True; /* 以概率 P_e 生成语法错误 */
4.       Q←getErrorQuantity( P_q ) /* 以概率 P_q 生成一处语法错误 */
5.       for I from zero to Q   /* 循环 Q 次 */
6.          C←getErrorClass(P_c, P_r, P_m, P_s, P_w, P_p);
            /* 按概率随机选择要生成的错误类型存入变量 C */
7.          T′, E←getError (T, P, C, C_s, C_p, B, U);
            /* 根据 C 生成含语法错误的句子并进行标注，分别存入 T′和 E */
8.       end for
9.    end if
10.   thGECC←getErrorPairs(T, T′, E); /* 将“正确-错误-错误类型标签”存入集合 thGECC */
11. end for
```

2.3　语料库数据统计

考虑到数据的领域丰富性，本文的原始语料主要来自中国国际广播电台网泰语版和在线泰语词典网，分为经济与贸易、外交、教育、环境、军事、政治、旅游、国际、社会以及生活与文化等十类约 78 995 条句子。为保证数据分布的一致性，本文首先将原始语料库按照约 8∶1∶1 的比例划分为训练集、验证集和测试集，再使用算法 1 生成语料，其中含有语法错误的句对有 63 006 条，语法正确的句对有 15 989 条，每个包含语法错误的句子中有一到两处错误。需要注意的是，由于我们采用人工规则算法自动生成语料库，每条原始语句产生不同错误类型的概率是相同的，所以不同领域的数据对于错误类型的分布是没有影响的。

数据集来源领域统计数据如表 3 所示，数据集的错误类型分布详细统计数据如表 4 所示。

① https://github.com/PyThaiNLP

② https://www.arts.chula.ac.th

表3 数据集来源领域数据统计

类别	训练集	验证集	测试集
经济与贸易	7 510	933	962
外交	7 953	980	1003
教育	5 748	712	713
环境	3 072	372	410
军事	1 168	150	153
政治	10 545	1 316	1 244
旅游	4 088	544	582
国际	6 370	779	784
社会	7 458	965	943
生活与文化	9 224	1 170	1 145
合计	63 136	7 921	7 939

表4 数据集错误类型分布数据统计

错误类型	训练集	验证集	测试集
拼写错误	23 746	3 048	3 046
词冗余错误	13 374	1 634	1 699
词缺失错误	8 715	1 142	1 093
用词不当错误	18 163	2 225	2 304
语序不当错误	4 305	573	525
停顿错误	2 867	353	368
合计	71 170	8 975	9 035

2.4 泰语文本语法错误检测

本文将文本语法错误检测视为序列标注任务。目前,该任务主流方法为基于双向长短时记忆网络(Bi-directional Long Short Term Memory Network, BiLSTM)的模型和基于BERT(Bi-directional Encoder Representation from Transformers)的模型。由于传统序列标注模型缺乏对语言特点的考虑,导致其在识别特定语言中的特有错误时性能不佳。为此,本文结合泰语文本由连续的泰文字符按一定组合规则排列成词再构成句子的特点,在多语言BERT(multilingual Bi-directional Encoder Representation from Transformers, mBERT)[23]序列标注模型的基础上融合词、字符与词性的深层语义表达进行语法错误检测。首先利用多语言BERT模型提取词的深层语义特征,同时使用词性与字符特征编码器编码词性与字符特征,将三者以拼接的方式融合得到最终的特征向量,然后使用多头注意力机制提取更丰富的特征信息。

模型主要由词嵌入层、词性与字符特征嵌入层、BERT层、词性与字符特征编码层、多头注意力层以及Softmax层组成,整体架构如图1所示。对于每个批次的训练数据,模型的训练分为以下四步:

第一步 对于句子$S=\{w_1,w_2,\cdots,w_n\}$(n表示句子中含有的词数),首先通过词嵌入层得到初始词嵌入向量$\boldsymbol{h}_i^0$,再通过BERT层提取深层语义特征$\boldsymbol{h}_i^L$,如式(1)、式(2)所示。

$$\boldsymbol{h}_i^0=\boldsymbol{W}_e\boldsymbol{w}_i+\boldsymbol{W}_p \tag{1}$$

$$\boldsymbol{h}_i^l=\text{Transformer\_block}(\boldsymbol{h}_i^{l-1}) \tag{2}$$

其中,$\boldsymbol{W}_e$、$\boldsymbol{W}_p$分别表示词嵌入权重和位置嵌入权重,l表示BERT模型层数,L表示BERT模型的最后一层。

第二步 先对输入的句子进行词性标注,再将输入的单词拆分为字符,然后通过词性与字符特征嵌入层得到初始词性特征向量$\boldsymbol{P}_i$和初始字符特征向量$\boldsymbol{T}_i$;接着使用卷积神经网络编码字符特征,然后将两者拼接起来得到初始字符与词性特征向量$\boldsymbol{F}_i$,如式(3)所示。最后,将其输入到一个BiLSTM网络中进行更深层次的编码,将BiLSTM网络的最后一层隐状态作为最终的词性与字符特征向量$\boldsymbol{F}'_i$,如式(3)~式(5)所示。

$$\boldsymbol{F}_i=\text{Conv}(\boldsymbol{T}_i)\oplus\boldsymbol{P}_i \tag{3}$$

$$\overrightarrow{\boldsymbol{h}}_i=\text{LSTM}(\overrightarrow{\boldsymbol{h}}_{i-1},\boldsymbol{F}_i),\overleftarrow{\boldsymbol{h}}_i=\text{LSTM}(\overleftarrow{\boldsymbol{h}}_{i-1},\boldsymbol{F}_i) \tag{4}$$

$$\boldsymbol{F}'_i=\overrightarrow{\boldsymbol{h}}_i\oplus\overleftarrow{\boldsymbol{h}}_i \tag{5}$$

其中,$\oplus$表示向量拼接操作,$\overrightarrow{\boldsymbol{h}}_i$和$\overleftarrow{\boldsymbol{h}}_i$分别表示位置$i$的前、后向LSTM的最后一层隐状态。

第三步 将前两步得到的输出做拼接操作得到最终的特征向量$\boldsymbol{H}_i$,如式(6)所示,作为多头注意力层的输入,使用多头注意力机制在不同的表示子空间中捕获得到更加丰富的特征信息$\boldsymbol{H}'_i$,如式(7)~式(9)所示。

$$\boldsymbol{H}_i=\boldsymbol{h}_i^L\oplus\boldsymbol{F}'_i \tag{6}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

$$\textbf{Head}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{8}$$

$$\boldsymbol{H}'_i=\text{Concat}(\textbf{Head}_1,\cdots,\textbf{Head}_h)\boldsymbol{W}^o \tag{9}$$

其中,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$向量由$\boldsymbol{H}_i$和权重矩阵计算得到,$\sqrt{d_k}$为$\boldsymbol{K}$的维度,$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$分别表示当前注意力头下$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$的权重矩阵;$\boldsymbol{W}^o$为注意力头权重矩阵;$h$表示多头注意力机制的头数。

第四步 对多头注意力层的输出 $\boldsymbol{H}'_i$ 作线性变换后经 Softmax 层得到错误类型的概率分布 $\boldsymbol{Y}_i$，并计算交叉熵损失函数，如式(10)、式(11)所示，最后通过反向传播更新网络参数。

$$\boldsymbol{Y}_i = \text{Softmax}(\boldsymbol{W}_o \boldsymbol{H}'_i + \boldsymbol{b}_o) \tag{10}$$

$$\text{Loss} = -\sum_{i=1}^{C} y_i \log(\boldsymbol{Y}_i) \tag{11}$$

其中，$\boldsymbol{W}_o$，$\boldsymbol{b}_o$ 分别为线性变换的权重矩阵和偏置矩阵，C 表示错误类型标签，y_i 为样本标签的 one-hot 表示，当样本属于第 i 类时取 1，否则取 0。

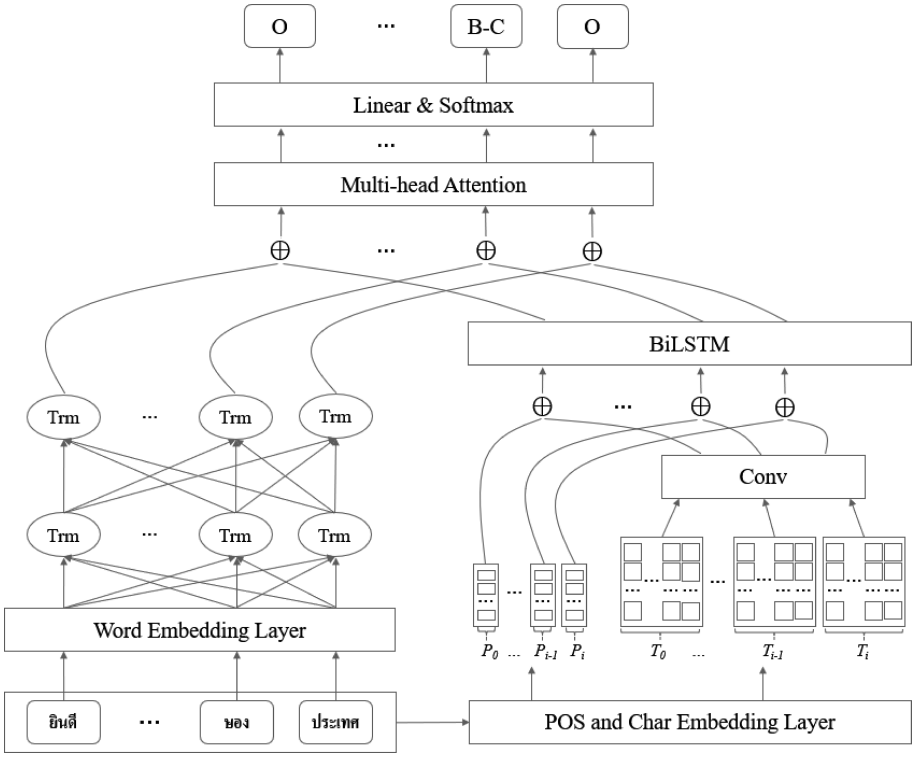

图 1　总体模型架构图

3　实验

3.1　数据集与评测指标

实验数据集为第 2.3 节所述的泰语文本纠错语料库，从中抽取包含语法错误的句子和对应的错误类型标签用于泰语文本语法错误检测任务，详细统计信息见表 3、表 4。

在评估泰语文本错误检测模型时，本文采用精确率 P、召回率 R 以及 F_1 值作为评价指标，如式(12)～式(14)所示。

$$P = \frac{\text{正确识别的错误类型标签总数}}{\text{识别出的错误类型标签总数}} \tag{12}$$

$$R = \frac{\text{正确识别的错误类型标签总数}}{\text{数据集中的错误类型标签总数}} \tag{13}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{14}$$

3.2　实验过程与结果分析

为验证本文所提模型的有效性，本文进行了多组实验。由于当前缺少关于泰语文本语法错误检测任务的研究，没有可作为参照的基线，因此本文在实验前首先设定四个基线模型：

(1) **BiLSTM**　基于 BiLSTM 的序列标注模型，本文对比了使用 GloVe 词嵌入初始化和 Word2Vec 词嵌入初始化的错误检测性能。

(2) **BiLSTM-LAN**[24]　分层细化的标签注意力序列标注模型，该方法使用多头注意力联合编码来自单词表示子空间和标签表示子空间的信息。

(3) **MHAL**[16]　基于多头注意力机制的 BiLSTM 序列标注模型，使用多头注意力机制将单个单词和完整句子之间的表示联系起来，其中每个注意力头都被作为一个序列标注器，用来检测每个词对应的标签。

(4) **mBERT**[23]　基于多语言 BERT 的序列标注模型，使用深度双向 Transformer 编码器建模语

言特征。

实验中，Word2Vec词嵌入和GloVe词嵌入均使用泰语版维基百科语料训练，前者使用Skip-Gram模型，两者的词嵌入维度都为100，上下文窗口大小都为5，最小词频都为5。BiLSTM网络隐藏单元为300，Dropout率为0.5，学习率为0.015，学习率衰减为0.05；BiLSTM-LAN网络和MHAL网络的参数设置与BiLSTM网络基本一致，唯一不同在于MHAL网络中学习率为1.0，学习率衰减为0.9；基于mBERT的模型词嵌入维度为768，隐藏单元大小为768，学习率为1e-5，句子最大长度为150，使用Adam优化器优化模型参数。此外，本文模型中的词性特征嵌入和字符特征嵌入的维度均为64。

在本文构建的测试集上的实验结果如表5所示。从表中可以看出，基于词性与字符特征融合的泰语文本语法检错方法在召回率R和F_1值上都比所有对比模型高，相比mBERT序列标注模型在仅损失0.41%精确率的情况下，在召回率上提高了3.8%，总体效果提升了1.37%的F_1值；在精确率上低于BiLSTM-LAN方法和MHAL方法，可能是由于前者使用了多头注意力联合编码来自单词表示子空间和标签表示子空间的信息，可以更准确地获得单词与标签之间的联系，而后者则使用多头注意力机制将单个单词和整个句子间的表示巧妙地联系在一起，可以捕获到更丰富的上下文信息，从而使得模型预测的精确率更高。

表5 不同方法的性能对比 （单位：%）

方法	精确率	召回率	F_1
Word2Vec-BiLSTM	45.46	32.01	37.57
GloVe-BiLSTM	60.15	32.04	41.81
BiLSTM-LAN	**74.96**	37.12	49.65
MHAL	71.06	41.06	52.05
mBERT	62.91	71.66	67.00
本文方法	62.50	**75.46**	**68.37**

由图2显示，相比mBERT序列标注模型，本文方法对词冗余错误和停顿错误的检测有相对明显的提升，分别提升了1.82%和3.99%的F_1值，在用词不当错误的检错方面也有小幅度提升，为0.82%的F_1值，证明了融合词性与字符特征的有效性。而在拼写错误和词序错误的检测上会有略微的下降，分别为0.32%和0.62%的F_1值。对此，本文通过分析实验数据和测试结果发现，引入词性与字符特征虽然在一定程度上有助于语法错误的检测，但同时也会引入噪声，导致模型预测时额外产生了少量误判，使精确率有略微下降，进而使模型在识别这两种错误类型时性能下降。

此外，无论是在泰语，还是在汉语等资源相对丰富的语言中，对词缺失错误的检测与纠正都是一个具有挑战性的问题。如图2所示，不管是基线模型还是本文模型，检测词缺失错误的性能都不理想，但本文方法与mBERT对于不使用mBERT的模型能够更好地检测词缺失错误。导致这一现象的原因可能是由于词缺失错误的检测相对于其他类型的文本语法错误检测需要依赖更加丰富的语义信息。另外，本文方法在mBERT模型的基础上融合了字符与词性特征，但是这两类特征对缺失词来说是未知的，因此融合这两类特征对检测词缺失错误的作用不大，从而导致两个模型在检测这类错误时的性能基本持平。

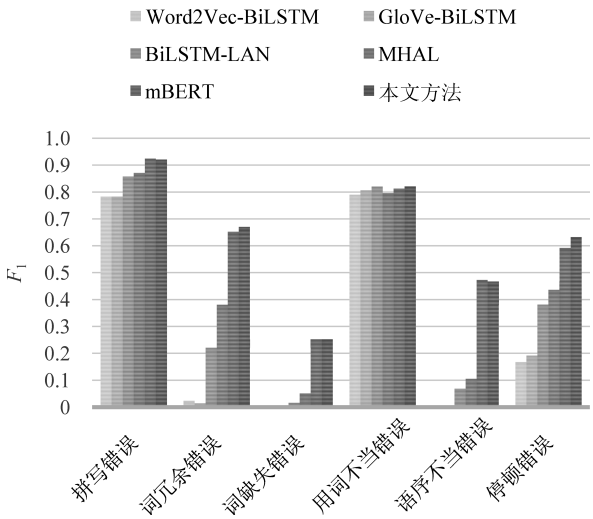

图2 不同方法在不同错误类型上的性能对比

3.3 消融研究

为验证融合词性与字符特征，以及多头注意力机制对多语言BERT序列标注模型性能的影响，本文通过以下7种方式进行消融实验：

（1）**mBERT** 仅依赖于多语言BERT的序列标注模型。

（2）**mBERT＋MHA** 在mBERT模型基础上加入多头注意力机制的模型。

（3）**mBERT＋feature(CharCNN)** 在mBERT模型基础上融合采用卷积神经网络编码的字符特征的模型。

（4）**mBERT＋POS＋MHA** 在mBERT模型

基础上融合词性特征并加入多头注意力机制的模型。

（5）**mBERT＋ CharCNN＋MHA**　在 mBERT 模型基础上融合采用卷积神经网络编码的字符特征，并加入多头注意力机制的模型。

（6）**mBERT＋feature（CharLSTM）＋MHA**　在 mBERT 模型基础上同时融入词性与字符特征并加入多头注意力机制的模型。与本文方法的差别在于字符特征使用双向长短时记忆网络编码。

（7）**本文方法**　在 mBERT 模型基础上同时融入词性与字符特征并加入多头注意力机制的模型，其中字符特征使用卷积神经网络编码。

如表 6 所示，在 mBERT 序列标注模型的基础上，单独加入多头注意力机制或词性与字符特征的性能会有所下降，而同时加入多头注意力机制和词性特征或字符特征的其中一种都会提升一定的性能。对此，可能的原因是 mBERT 模型已经使用深度双向 Transformer 网络很好地提取了特征信息，单独加入多头注意力机制会破坏这些特征信息；而单独加入的额外特征即使不进行更深层的编码，也会对强大的 BERT 模型造成负面影响，综上可以证明本文所提模型中各组件的存在都是有必要的。除此之外，本文还对不同的字符特征提取方式进行了实验（表 6 中最后两行），对比发现使用卷积神经网络的性能更好。

表 6　本文方法中不同组件对模型性能的影响　（单位：%）

模型结构	精确率	召回率	F_1
mBERT	62.91	71.66	67.00
mBERT＋MHA	60.76	74.30	66.85
mBERT＋feature（CharCNN）	**62.99**	70.79	66.66
mBERT＋POS＋MHA	61.39	75.20	67.60
mBERT＋CharCNN＋MHA	62.04	74.44	67.68
mBERT＋feature（CharLSTM）＋MHA	62.08	73.47	67.30
本文方法	62.50	**75.46**	**68.37**

如图 3 显示，在 mBERT 序列标注模型基础上单独加入多头注意力机制或词性与字符特征，会对模型错误检测性能造成负面影响，主要体现在词缺失错误（两个模型分别损失 2.23%和 0.78%的 F_1 值）和语序不当错误（两个模型分别损失 2.29%和 1.17%的F_1值）上，检测其他错误类型的性能与

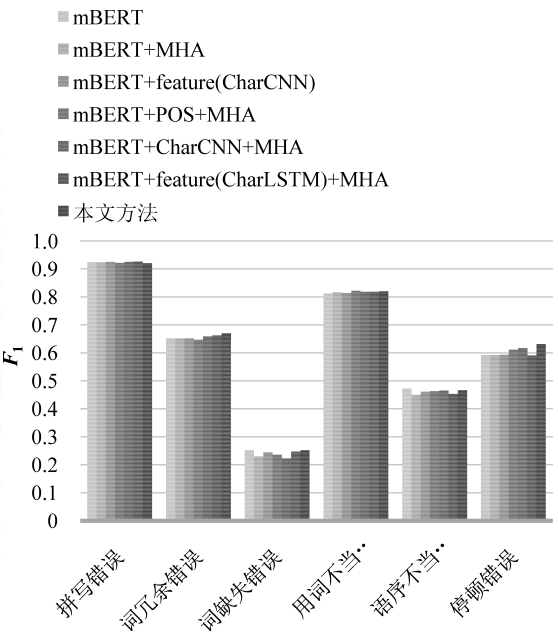

图 3　本文方法中不同组件对模型检测不同错误类型的性能影响

mBERT 序列标注模型基本持平。而单独融合字符特征或词性特征并加入多头注意力机制仅在个别错误类型的检测上会使模型损失一些性能（在词缺失错误的检测上两个模型分别损失 1.6%和 2.93%的 F_1 值，在语序不当错误的检测上两个模型分别损失 1.00%和 0.74%的 F_1 值）。总体而言，对模型是起到正面作用的，在拼写错误的检测上两个模型分别提升－0.22%和 0.12%的 F_1 值，在词冗余错误的检测上两个模型分别提升－0.49%和 0.74%的 F_1 值，在用词不当错误的检测上两个模型分别提升 0.96%和 0.63%的 F_1 值，在停顿错误的检测上两个模型分别提升 1.96%和 2.52%的 F_1 值。当在 mBERT 模型上同时融合词性和字符特征并加入注意力机制时，相比仅依赖于 mBERT 的模型，在六类语法错误的检测上分别提升了－0.32%、1.82%、0.01%、0.82%、－0.62%和 3.99%的 F_1 值，由此可进一步证明本文模型中各组件存在的必要性和有效性。

为验证本文所提语料库构建方法的有效性，在验证集和测试集规模不变的前提下，本文分别使用训练集的 25%、50%、75%和 100%训练表 5 中的基线模型与本文方法，结果如图 4 所示。从图中可以看出，随着训练集规模的扩大，本文方法和基线方法的性能均会逐步提升。此外，本文在四个训练数据规模下相比 mBERT 模型分别提升了0.7%、1.72%、1.14%和 1.37%的 F_1 值，由此可证明本文语料库构建方法的有效性，并可进一步证明本文提出的泰语

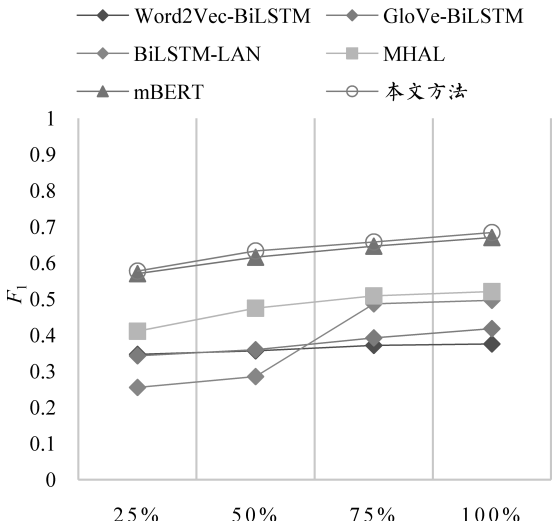

图 4 不同方法在不同数据规模下的性能对比

文本语法错误检测方法的有效性。

3.4 实例与分析

本文对检测失败的示例进行分析，如表 7 所示，其中，“×”表示含有语法错误的文本，“√”表示对应的正确文本，发现当句子中出现以下情况时通常会使模型检测失败：

(1) 拼写错误的单词存在于词典中。例如，模型可以识别出示例 1 成功示例中拼写错误的单词“เวิา”，该单词不存在于词典中，属于 Non-word 拼写错误。而模型无法识别出失败示例中拼写错误的单词“เฉียง”，该词存在于词典中，为“倾斜，对角”的意思，模型将其误判成用词不当错误。

表 7 本文方法检测泰语文本语法错误成功与失败的示例

示例 1	成功	×	วัน/ที่/7/_/ตุลาคม/ตาม/[เวิา]/ท้องถิ่น O B-PM O O O O B-C O
		√	วัน/ที่/[_]/7/_/ตุลาคม/ตาม/[เวลา]/ท้องถิ่น 当地时间 10 月 7 日。
	失败	×	ใคร/ทำ/[อยู่/อะไร]/ใน/ครัว/[เฉียง]/ดัง/แก๊ก O O B-W I-W O O B-C O O(正确标签) O O B-W I-W O O B-S O O(预测标签)
		√	ใคร/ทำ/[อะไร/อยู่]/ใน/ครัว/[เสียง]/ดัง/แก๊ก 谁在厨房里做什么？
示例 2	成功	×	ไม่ว่า/สถานการณ์/ระหว่าง/ประเทศ/[ไป/เปลี่ยน/จะ]/อย่าง/ไร O O O O B-W I-W I-W O O
		√	ไม่ว่า/สถานการณ์/ระหว่าง/ประเทศ/[จะ/เปลี่ยน/ไป]/อย่าง/ไร 无论国际形势如何变化。
	失败	×	นอกเหนือ/[อุปสงค์/ด้าน/ได้/อุปทาน/จาก]/รับ/การ/ปรับปรุง/ดี/[ขึ้น]/แล้ว O B-W I-W I-W I-W I-W O O O O B-S O(正确标签) O B-W I-W I-W O O O O O O B-S O(预测标签)
		√	นอกเหนือ/[จาก/ด้าน/อุปสงค์/อุปทาน/ได้]/รับ/การ/ปรับปรุง/ดี/[ขึ้น]/แล้ว 除了需求端，供给端也有所改善。
示例 3	成功	×	ใน/[การ]/สมัย/โบราณ/ชาว/[คหัฐ]/เชื่อถือ/ผี/สาง/เทวดา O B-R O O O B-S O O O O
		√	ใน/[การ]/สมัย/โบราณ/ชาว/[บ้าน]/เชื่อถือ/ผี/สาง/เทวดา 在古代，村民信奉若虫。
	失败	×	เรา/ต้อง/เธียนรู้/ศัพท์/แสง/ทาง/คอมพิวเตอร์/ไว้/[อะไร]/บ้าง O O O O O O O O B-R O(正确标签) O O O O O O O O O O(预测标签) 我们需要学习哪些计算机术语？
		√	เรา/ต้อง/เธียนรู้/ศัพท์/แสง/ทาง/คอมพิวเตอร์/ไว้/[อะไร]/บ้าง 我们需要学习一些计算机术语。

续表

示例 4	成功	×	ราว/ผ้า/ที่/ทำ/[ด้วญ]/ลวด/จะ/[_]/เป็น/สนิม/_/ทำ/ให้/เปื้อน/ผ้า/ได้ O O O O B-C O O B-PR O O O O O O O O O
		√	ราว/ผ้า/ที่/ทำ/[ด้วย]/ลวด/จะ/[_]/เป็น/สนิม/_/ทำ/ให้/เปื้อน/ผ้า/ได้ 金属丝挂衣杆会使织物生锈染色。
	失败	×	ผม/[นั่ซ่ง/เธอ/กับ/คุย]/อยู่/ริม/หน้าต่าง O B-C B-W I-W I-W O O O(正确标签) O B-C O O O O O O O(预测标签)
		√	ผม/[นั่ง/คุย/กับ/เธอ]/อยู่/ริม/หน้าต่าง 我坐在窗边和她说话。

（2）语序不当错误中乱序短语较长。例如，模型可以识别出示例 2 成功示例中较短短语的语序不当错误，对失败示例中较长短语的语序不当错误识别不完整。

（3）需要依靠更丰富的上下文信息（例如篇章级信息）进行判断的错误。如示例 3，模型可以识别出仅依赖于单个句子的词冗余错误，识别不出在单句层面没有错误，但可能在更长的上下文环境下或者特定语境中为冗余词的错误。

（4）当句子中有多处语法错误时，错误之间距离太近。如示例 4 的成功示例中，两处语法错误之间存在间隔，模型可以准确地识别出与之对应的错误类型，但对于失败示例中紧紧相邻的两处错误，模型仅识别出了一处。

4　结论

本文针对泰语文本语法错误检测与纠正任务语料匮乏问题，结合泰语语言学及文本语法错误类型特点，基于人工启发式规则构建了一定规模的泰语文本语法错误检测与纠正语料库。基于该语料库，提出融合语言学特征的泰语文本语法错误检测方法，在多语言 BERT 模型的基础上融合字符、词与词性的深层语义表达。实验结果表明，本文方法能在多语言 BERT 模型上进一步提升性能。下一步的工作，本文将考虑扩大语料库规模，尝试构建其他东南亚小语种的文本纠错语料库，并探索使用更有效的模型方法和更丰富的语言学特征来提升东南亚小语种文本语法错误检测的效果，同时开展文本语法错误校正相关的研究。

参考文献

[1] KANEKO M, MITA M, KIYONO S, et al. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4248-4254.

[2] ROTHE S, MALLINSON J, MALMI E, et al. A simple recipe for multilingual grammatical error correction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 702-707.

[3] SUN X, GE T, WEI F R, et al. Instantaneous grammatical error correction with shallow aggressive decoding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 5937-5947.

[4] MEKNAVIN S, KIJSIRIKUL B, CHOTIMONGKOL A, et al. Combining trigram and winnow in thai OCR error correction[C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998:836-842.

[5] RODPHON M, SIRIBOON K, KRUATRACHUE B. Thai OCR error correction using token passing algorithm[C]//Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. Piscataway, NJ: IEEE, 2001: 599-602.

[6] WATCHARABUTSARAKHAM S. Spell checker for Thai document [C]//Proceedings of the IEEE Region Conference, 2005: 1-4.

[7] NG H T, WU S M, BRISCOE T, et al. The CoNLL-2014 shared task on grammatical error correction[C]//Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task, 2014: 1-14.

[8] LEE L H, RAO G Q, YU L C, et al. Overview of NLP-TEA shared task for chinese grammatical error diagnosis[C]//Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, 2016: 40-48.

[9] RAO G Q, ZHANG B L, XUN E D, et al. IJCNLP-2017 Task 1: Chinese grammatical error diagnosis [C]//Proceedings of the IJCNLP, Shared Tasks, 2017: 1-8.

[10] RAO G Q, GONG Q, ZHANG B L, et al. Overview of NLPTEA share task Chinese grammatical error diagnosis[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, 2018: 42-51.

[11] RAO G Q, YANG E, ZHANG B L. Overview of NLPTEA shared task for Chinese grammatical error diagnosis [C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, 2020b: 25-35.

[12] WANG L, ZHAO W, JIA R Y, et al. Denoising based sequence-to-sequence pre-training for text generation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 4003-4015.

[13] LICHTARGE J, ALBERTI C, KUMAR S, et al. Corpora generation for grammatical error correction [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3291-3301.

[14] ZHOU W C S, GE T, MU C, et al. Improving grammatical error correction with machine translation pairs[C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2020: 318-328.

[15] KANEKO M, KOMACHI M. Multi-head multi-layer attention to deep language representations for grammatical error detection[J]. Computaciónу Sistemas, 2019, 23(3): 883-891.

[16] 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.

[17] YUAN Z, TASLIMIPOOR S, DAVIS C, et al. Multi-class grammatical error detection for correction: A tale of two systems[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021:8722-8736.

[18] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators[C]//Proceedings of the International Conference on Learning Representations, 2020: 1-18.

[19] BRYANT C, FELICE M, BRISCOE T. Automatic annotation and evaluation of error types for grammatical error correction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 793-805.

[20] 裴晓睿. 泰语语法新编[M]. 北京: 北京大学出版社, 2001.

[21] COLLIER N, KIM J D. Introduction to the bio-entity recognition task at JNLPBA[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004: 73-78.

[22] SORNLERTLAMVANICH V, TAKAHASHI N, ISAHARA H. Building a thai part-of-speech tagged corpus (ORCHID)[J]. Journal of the Acoustical Society of Japan, 1999, 20(3):189-198.

[23] 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.

[24] CUI L Y, ZHANG Y. Hierarchically-refined label attention network for sequence labeling[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 4115-4128.

施灿镇(1998—),硕士研究生,主要研究领域为自然语言处理、文本纠错。
E-mail: crassphage1998@163.com

余正涛(1970—),教授,博士生导师,主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail: ztyu@hotmail.com

朱俊国(1982—),通信作者,博士,讲师,主要研究领域为自然语言处理、机器翻译。
E-mail: zhujunguohit@gmail.com