DOI:10.19651/j.cnki.emt.2210763

语音转录后文本的中文拼写纠错模型*

邢月晗 郑 岩
(北京邮电大学人工智能学院 北京 100876)

摘 要: 针对目前语音转录文本错误率较高的问题,本文提出一种基于 MacBERT 的文本先检错后纠错模型,对语音转录后文本进行校正。检错阶段使用 MacBERT-BiLSTM-CRF 模型检查文本是否有错及出错位置。纠错阶段从置信度和字音相似度两个维度出发,划定"置信度-字音相似度"曲线判断候选字是否进行纠错。候选字的置信度使用 MacBERT 语言模型计算,并提出一种基于拼音码的字音相似度计算方法。在语音公开数据集 Thchs-30 上通过调用百度语音识别 API 进行实验,相比现有方法,在检错阶段和纠错阶段的精确率、召回率、F1 值都得到了提高,其中纠错阶段精确率达到 83.32%,提高了转录文本的正确性。

关键词: 语音;文本纠错;MacBERT;拼音码;Thchs-30
中图分类号: TP3 **文献标识码:** A **国家标准学科分类代码:** 520.2020

Chinese spelling error correction model for transcribed text

Xing Yuehan Zheng Yan
(Beijing University of Posts and Telecommunications, School of Artificial Intelligence,, Beijing 100876, China)

Abstract: Aiming at the high error rate of speech transcription text, proposes a text error detection and correction model based on MacBERT to correct the text after speech transcription. In the error detection stage, the MacBERT-BiLSTM-CRF model is used to check whether the text is wrong and where it is. In the error correction stage, starting from the two dimensions of confidence and phonetic similarity, a curve of "confidence-phonetic similarity" is delineated to determine whether candidate words are to be corrected for errors. The confidence of the candidate words is calculated using the MacBERT language model, and a phonetic similarity calculation method based on pinyin code is proposed. Experiments were conducted on the public speech dataset Thchs-30 by calling Baidu speech recognition API. Compared with the existing methods, the precision rate, recall rate and F1 value in the error detection stage and error correction stage have been improved. Among them, the error correction stage The accuracy rate reaches 83.32%, which improves the accuracy of the transcribed text.

Keywords: speech;text error correction;MacBERT;pinyin code;Thchs-30

0 引 言

随着智能制造中人机交互需求的增加,看剧听播客时 AI 字幕的兴起,社交软件语音转文字功能的广泛应用,如何保证语音转录文字的正确性[1]成为广泛研究的课题[2]。针对目前语音转录端错误率高的问题,本文提出一种基于 MacBERT(MLM as corrector BERT)[3]的文本检错纠错模型,针对语音转录后的中文文本进行拼写纠错。

目前针对中文拼写错误的纠错大致有以下 3 种方法:第一,基于 N-Gram[4]模型的中文自动纠错模型[5],利用 N-Gram 模型结合混淆字表、词表进行纠错,在专用领域能得到较好的结果,由于混淆字表词表的局限性,很难在通用领域得到很好的结果;第二,基于 Word2Vec[6]表征的中文纠错模型[7],利用 Word2Vec 模型表征中文词向量纠错,但 Word2Vec 模型本质是词袋模型(bag-of-words model)[8],无法表示一个词在不同语境下不同意义;第三,基于 BERT (bidirectional encoder representation from transformers)[9]等预训练语言模型掩码策略的中文纠错模型[10],预训练语言模型应用 Transformer[11]的 Encoder 结构学习双向语义解决了 Word2Vec 等语言模型无法结合上下文表示词义的问题,用海量自然数据进行预训练解决了检错纠错模型需要海量标注中文训练文本的问题。但以上 3 种方法均没有

收稿日期:2022-07-21
* 基金项目:教育部-中国移动科研基金(MCM20190701)项目资助

考虑到语音转录文本错误的特点，没有在纠错加入拼音特征，导致纠错的精确率不高。

由文献[12]统计 NIST02 和 NIST03 数据集下 2 806 处语音转录错误，获得忽略标点符号转录错误和数字转录错误的汉字转录错误比例，其中同音错误 48.58%，近音错误 24.20%，多字少字错误 17.99%，其他错误 9.23%，同音错误和近音错误占绝大部分比例。

先前 3 种方法均没有考虑到语音转录文本错误的特点，在纠错的时候只在混淆集中取字或取词，或者是只关注到候选字在语言模型中的可能概率，都没能在纠错时最大化利用到字音这一特征。本文针对语音转录错误的特点，提出基于 MacBERT 的检错纠错模型，结合基于拼音码的字音相似度计算方法，在纠错时加入字音相似度概念，用 Thchs-30 语音公开数据集实验，以期达到更高的纠错精确率和召回率。

1　基于 MacBERT 检错纠错模型

基于 MacBERT 的检错纠错模型，利用 MacBERT 预训练语言模型表征语音转录文本，与 BERT 模型相比，MacBERT 模型在预训练阶段发生两个部分变化：第一，在预训练阶段不再使用“[MASK]”的方式进行以字为单位的掩码，而是基于相似词语进行替换；第二，在预训练阶段将 NSP(next sentence predict)任务替换为 SOP(sentence order predict)任务。任务一将句子中 15%的词进行替换，其中 80%近义词替换，10%随机替换，10%不进行替换。采用 Synonyms 词向量工具进行相似词的选择。其中相似词语替换利用 LTP 分词工具对句子分词后，采用 N-Gram 的方式进行 Mask，对 1-Gram、2-Gram、3-Gram 和 4-Gram 分别按照 40%、30%、20%、10%的比列进行词语替换。用相似词替换代替随机掩码替换策略，使得模型分辨表达能力更强，更符合文本纠错任务的特点。本文基于 MacBERT 预训练语言模型设计检错纠错模型结构对语音转录文本出现错误进行纠正。

1.1　检错模型

LSTM(long short term memory)[13]一种循环神经网络[14]，相比于传统的循环神经网络结构，LSTM 在循环单元内部引入了门限结构，门限控制了内循环中的权重信息，权重的大小可以根据上下文动态调整，这种结构也能很好的解决“长依赖问题”，将循环开始的一些重要信息保留下来。LSTM 的核心思想在于保证了在较长序列长度的情况下梯度能正常地传播，从而减缓了梯度爆炸或者消失的问题。

BiLSTM 是双向的 LSTM 模型，功能是特征提取，需要在模型后添加下游模型完成具体的任务，而 CRF(conditional random field)就是进行序列标注的下游模型[15]。线性链条件随机场表示的是给定一组输入随机变量 X 的条件下另一组输出随机变量 Y 的马尔可夫随机场，是图概率模型[16]的一种。

在检错阶段采用 MacBERT-BiLSTM-CRF 模型用序列标注的方式返回出错位置。检错分为 3 层结构，第 1 层是 MacBERT 层，在 MacBERT 预训练模型基础上用语音转录后文本数据进行微调，语言表征更加符合语音转录后文本的特点，MacBERT 模型获取文本语义信息并输出融合上下文特征的词向量 $\boldsymbol{X}$；第 2 层是 BiLSTM 层，接受 MacBERT 层的词向量 $\boldsymbol{X}$ 信息，并输出双向序列信息 $\boldsymbol{H}$；第 3 层为 CRF 层，用规则来保证预测的合法性 $\boldsymbol{P}$，并输出最终预测结果。如图 1 所示，检错模型检查文本是否出错并返回出错位置，对“暴雨洪劳”用序列标注的形式来检错，对检测错误的字用 W 表示，对检测正确的字用 O 表示。

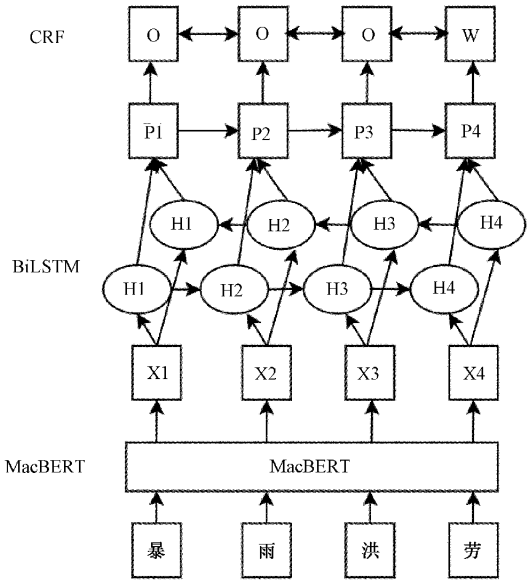

图 1　检错阶段的 3 层结构示例

1.2　纠错模型

纠错阶段用基于 MacBERT 的置信度[17]与字音相似度相结合的纠错模型，将检错模型输出的句子出错位置作为输入进行纠错。纠错时用 MacBERT 语言模型得出错字的置信度前三候选字，候选字按照置信度大小排序，分别获取 3 个候选字与原字的字音相似度，依次取候选字看其是否通过消融曲线，若通过消融曲线则完成纠错，未通过消融曲线则进行下一个候选字的判定，3 个候选字均未通过消融曲线则不进行纠错。

如图 2 所示，为整个纠错阶段的流程示例，“婴”的置信度第一候选字为“鹰”，置信度和字音相似度均通过消融曲线，进行纠错；“尉”的置信度第一候选字为“取”，因字音相似度较小不进行纠错，置信度第二候选字为“喂”，置信度和字音相似度均通过消融曲线，进行纠错。

1)置信度

由 MacBERT 语言模型产生的预测字概率值，值越大代表候选字在这的置信度越高。

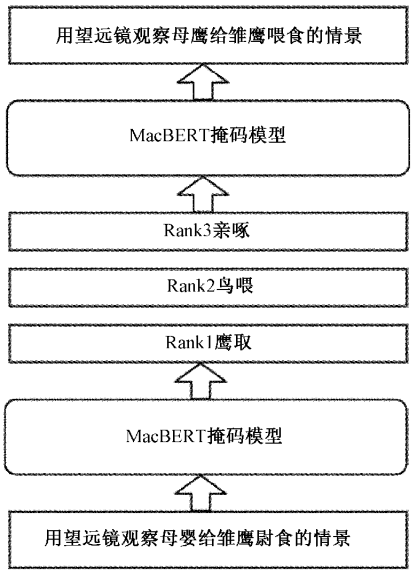

图 2 纠错流程示例

2)字音相似度

原字与候选字之间的字音相似度,采用生成拼音码,用拼音码计算字音相似度。

3)消融曲线

根据训练集的数据,在以置信度为横坐标,字音相似度为纵坐标的坐标系,生成“置信度-字音相似度图”。如图 3 所示,其中红色点表示该候选字是正确的,蓝色的点表示该候选字是错误的,由图可看出一般置信度越高,字音相似度越大的候选字是正确选择:

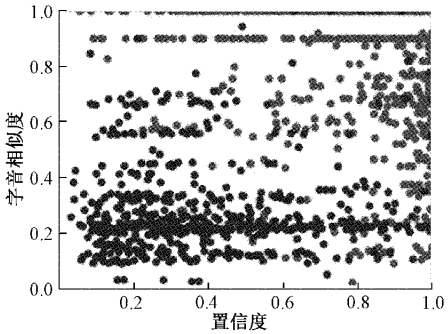

图 3 置信度-字音相似度图

根据训练集图像划出消融曲线,如图 4 消融曲线图所示,通过消融曲线的候选字进行纠正,未通过消融曲线的候选字(阴影部分)不纠正。

在训练集进行消融实验,划定更好消融曲线,进一步控制候选字“置信度-字音相似度”的通过范围;并在候选字未通过消融曲线时,对后续候选字继续进行消融实验,以期提高纠错的精确率和召回率。

2 字音相似度计算方法

字音相似度是指两个汉字发音相似的程度,汉语拼音

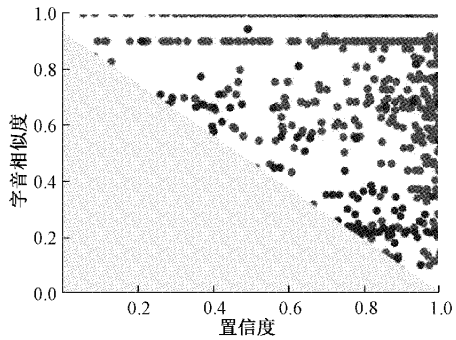

图 4 消融曲线图

是辅助汉字发音的有效工具,本文用 pypinyin 工具获取汉字拼音,使用汉语拼音构造拼音码,求得字音相似度。具体构造方式如下文所示。

2.1 拼音码

将汉语拼音分为声母、介音、韵母和声调 4 个码位。第 1 位韵母位,包含 24 个韵母如表 1 所示。

表 1 韵母表

zz	A	B	C	D	E	F
韵母	a	o	e	i	u	v
码值	G	H	I	J	K	L
韵母	ai	ei	ui	iou	ie	ve
码值	M	N	O	P	Q	R
韵母	er	an	en	in	un	vn
码值	S	T	U	V	W	X
韵母	ang	eng	ing	ong	ao	ou

第 2 位是声母位,包含如表 2 所示 23 个声母。

表 2 声母表

码值	A	B	C	D	E	F
声母	b	p	m	f	d	t
码值	G	H	I	J	K	L
声母	n	l	g	k	h	q
码值	M	N	O	P	Q	R
声母	x	zh	sh	ch	r	z
码值	S	T	U	V	W	
声母	c	s	y	w	j	

第 3 位是介音位,包含如表 3 所示 3 个介音。

表 3 介音表

码值	A	B	C
介音	u	v	i

第 4 位是音调位,包含 5 种音调,如表 4 所示。

表 4　音调表

码值	A	B	C	D	E
音调	-	ˊ	ˇ	ˋ	轻声

拼音码由以上 4 个码位组成，若码位为空则补“0”。如“纠”和“错”的拼音码，分别为“JX0A”和“BSED”。

2.2　字音相似度计算

由于不同声母或韵母之间读音区分的难易程度不同，如声母“n”与声母“l”前鼻音难以区分，韵母“in”与韵母“ing”后鼻音难以区分，声母“n”和声母“zh”容易区分。直观看声母“n”与声母“l”的相似度要比声母“n”与声母“zh”的相似度高。因此计算候选字与原字拼音码码值间相似度的大小，获得码间相似度矩阵，再将候选字与原字的 4 个码位的相似度加和，求得字音相似度，比简单的用编辑距离来求候选字与原字间字音相似度的方式会更加精准。

基于收集的字音易混字表，针对 4 992 个常用汉字的 39 463 个读音易混字，针对声母、介音、韵母和声调进行统计计算字音相似度。

$$Sheng_{ij}=\frac{a_{ij}}{a_i} \tag{1}$$

以声母为例，其中 $Sheng_{ij}$ 为声母 i 和声母 j 的相似度，其中 a_{ij} 为声母 i 在易混字表中错读成声母 j 的次数，a_i 为在易混字表中声母 i 发生错读的次数。

依次计算出声母位 $Sheng_{ij}$、韵母位 Yun_{ij}、介音位 Jie_{ij} 和声调位 $Diao_{ij}$ 的相似度，进而求得候选字与原字之间字音相似度 Sim_{ij}。

$$Sim_{ij}=\alpha_1\times Yun_{ij}+\alpha_2\times Sheng_{ij}+\alpha_3\times Jie_{ij}+\alpha_4\times Diao_{ij} \tag{2}$$

字音相似度的计算如式(2)所示，由 4 个码位的相似度加权求和得到，α_1、α_2、α_3 和 α_4 四个超参数的设置由实验效果决定。

3　实　　验

本模型实验运用 Thchs-30 语音公开数据集，Thchs-30 公开数据集是由清华大学语音与语言技术中心发布的开放式中文语音数据集，该数据集语音总时长超过 30 h，录音者大多是会说流利普通话的大学生。

实验将 Thchs-30 数据集通过调用百度智能云平台语音转文本 API，进行语音转录，将转录后的文本与正确文本对比，共 13 388 篇文本，得出转录完全正确文本 4 699 篇，转录拼写错误文本 7 141 篇，多字少字文本 1 548 篇。针对转录完全正确文本和转录拼写错误文本，将文本按 8∶1∶1 的比例划分训练集、验证集和测试集。在检错阶段和纠错阶段与 Pycorrector 纠错模型[18]、基于 BERT 掩码策略的纠错模型[19]，在精确率、召回率和 F1 值三个指标下进行对比。其中精确率(P)、召回率(R)和 F1 值三个指标的具体计算公式如下：

$$P=\frac{T_P}{T_P+F_P}\times 100\% \tag{3}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\% \tag{4}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{5}$$

其中，T_P 是错误字被检错(纠错)模型正确预测(纠正)的次数，F_P 是正确字被检错(纠错)模型错误的识别(纠正)的次数，F_N 是错误字没被检错(纠错)模型正确识别(纠正)的次数。

检错阶段将本文模型与方案 1 基于 Pycorrector 纠错模型和方案 2 基于 BERT 掩码策略的纠错模型相对比，得到结果如表 5 所示，在同量级训练数据下，基于 MacBERT 和 BERT 预训练模型的 BiLSTM-CRF 的错误检测模型在精确率、召回率和 F1 值都显著优于 Pycorrector 模型；其中 MacBERT 预训练语言模型结合 BiLSTM-CRF 错误检测模型在检错任务的精确率、召回率和 F1 值 3 类指标上都优于 BERT-BiLSTM-CRF 模型，证实了 MacBERT 预训练语言模型更适合文本检错任务。

表 5　检错阶段结果对比　　%

	P	R	F1
方案 1	86.24	69.31	69.37
方案 2	90.31	76.62	82.90
本文模型	90.37	76.87	83.07

纠错阶段将本文模型与方案 1 基于 Pycorrector 纠错模型和方案 2 基于 BERT 掩码策略的纠错模型相对比，得到结果如表 6 所示，MacBERT 结合字音相似度方法在精确率、召回率和 F1 值指标上都明显优于上述两种方案。其中方案 1 方法精确率较高，但召回率偏低，印证了按混淆字表召回的方式覆盖面不广的问题。本文模型在精确率和召回率上相对直接用 BERT 纠错的方案 2 都有大幅提高，特别在精确率指标上提升大于 20%。

表 6　纠错阶段结果对比　　%

	P	R	F1
方案 1	61.92	30.05	40.47
方案 2	58.97	37.48	45.83
本文模型	82.32	46.63	59.54

将本文模型与不使用字音相似度纠错的方案相对比，得到结果如表 7 所示。结合字音相似度的方式，在精确率和召回率上相对直接用置信度纠错的方式都有大幅提高，特别在精确率指标上提升大于 20%，印证了在语音转录文本结合字音相似度进行纠错思路的正确性。

将本文模型与不使用置信度纠错的方案相对比，得到结果如表 8 所示，纠错时结合置信度的方式，比单用字音相

表7 是否使用字音相似度结果对比 %

	P	R	F1
本文模型-字音相似度	61.73	41.80	49.85
本文模型	82.32	46.63	59.54

表8 是否使用置信度结果对比 %

	P	R	F1
本文模型-置信度	72.43	40.18	51.69
本文模型	82.32	46.63	59.54

似度的方式在精确率、召回率都有大幅提升。

表7、8两个实验,印证了在纠错时划定“置信度-字音相似度”曲线纠错的正确性。

将使用相同纠错策略,预训练语言模型不同的两方案相对比,得到结果如表9所示,基于MacBERT的纠错模型也在精确率、召回率和F1值上优于基于BERT的纠错模型,论证了MacBERT预训练语言模型更适合文本纠错任务。

表9 不同预训练语言模型结果对比 %

	P	R	F1
方案2+字音相似度	81.44	40.93	54.48
本文模型	82.32	46.63	59.54

综上所述,本文设计的基于MacBERT的检错纠错模型,相比现有方法在精准率和召回率上都有大幅提升,在语音文本校正领域有较大应用价值。

4 结 论

在人类对智能制造、智能生活期待越来越高的今天,语音转文本的正确性是其中值得被研究的课题。本文针对语音转录后的文本提出一种基于MacBERT的检错纠错模型,对语音转录文本进行校正,得到较先前方法更好的效果。主要贡献点:

使用MacBERT公开预训练语言模型,并论证其有效性;

设计检错网络、纠错网络两阶段纠错;

针对语音转录文本特点设计拼音码,创新字音相似度计算方式;

纠错阶段对候选字划定“置信度-字音相似度”消融曲线,对置信度前三的候选字依次判定纠错。

但针对语音转录文本的多字少字问题没有得到很好的解决,今后也会针对此类问题进行进一步的研究。

参考文献

[1] 张琳涵. 面向转录文本的语音识别错误检测和纠正方法研究[D]. 哈尔滨:哈尔滨工业大学,2020, DOI:10.27061/d.cnki.ghgdu.2020.004594.

[2] 关健. 基于深度神经网络和多元损失的说话人识别[J]. 电子测量技术,2019,42(5):39-43.

[3] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]. Proceedings of the 2020 Conference on Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 657-668.

[4] 尹陈. N-gram模型综述[J]. 计算机系统应用,2018,27(10):33-38.

[5] 王匆匆. 基于上下文的中文文本自动校对方法[D]. 北京:北京信息科技大学,2020.

[6] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]. International conference on machine learning. PMLR, 2014: 1188-1196.

[7] 刘峻松. 基于Word2Vec的编程领域词语拼写错误检测算法[J]. 计算机应用与软件,2022,39(3):277-284.

[8] 黄春梅. 基于词袋模型和TF-IDF的短文本分类研究[J]. 软件工程,2020,23(3):1-3.

[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. NAACL-HLT(1), 2019: 4171-4186.

[10] ZHANG S, HUANG H, LIU J, et al. Spelling error correction with soft-masked BERT[J]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020:882-890.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017,30: 6000-6010.

[12] 田新宇. 语音识别错误对翻译性能的影响分析[J]. 厦门大学学报(自然科学版), 2022,61(4):682-688.

[13] SHI X J, CHEN Z, WANG H, et al, 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 802-810.

[14] SCHUSTER M, PALIWALK K. Bidirectional recurrent neural networks[J]. IEEE Transactionson SignalProcessing, 1997,45(11):2673-2681.

[15] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. ArXiv Preprint,2015, ArXiv:1508.01991.

[16] 刘建伟. 概率图模型表示理论[J]. 计算机科学,2014,41(9):1-17.

[17] 崔铁军. 系统多功能状态表达式构建及其置信度研究[J]. 智能系统学报,2023, 18(1):7.

[18] XU M. Pycorrector: Text Error Correction Tool[J]. 2019.

[19] 汪苏琪. 面向规范性文件的基于BERT的文本纠错模型[J]. 山西大学学报(自然科学版),2022,45(2):257-263.

作者简介

邢月晗,硕士研究生,主要研究方向为自然语言处理。
E-mail: xyh@bupt.edu.cn

郑岩(通信作者),副教授,主要研究方向自然语言处理、数据挖掘等。
E-mail: yanzheng@bupt.edu.cn