文章编号：1003-0077(2022)09-0001-18

中文文本自动校对综述

李云汉[1,2]，施运梅[1,2]，李 宁[1,2]，田英爱[1,2]

(1.北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 100101；
2.北京信息科技大学 计算机学院，北京 100101)

摘 要：文本校对在新闻发布、书刊出版、语音输入、汉字识别等领域有着极其重要的应用价值，是自然语言处理领域中的一个重要研究方向。该文对中文文本自动校对技术进行了系统性的梳理，将中文文本的错误类型分为拼写错误、语法错误和语义错误，并对这三类错误的校对方法进行了梳理，对中文文本自动校对的数据集和评价方法进行了总结，最后展望了中文文本自动校对技术的未来发展。

关键词：自动校对；拼写错误；语法错误；语义错误；数据集；评估指标

中图分类号：TP391 **文献标识码**：A

A Survey of Automatic Error Correction of Chinese Text

Li Yunhan[1,2], Shi Yunmei[1,2], Li Ning[1,2], Tian Ying'ai[1,2]

(1. Beijing Information Science and Technology University, Beijing Key Laboratory of Internet Culture Digital Dissemination, Beijing 100101, China;
2. School of Computer, Beijing University of Information Technology, Beijing 100101, China)

Abstract: Text correction, an important research field in Natural Language Processing (NLP), is of great application value in fields such as news, publication, and text input . This paper provides a systematic overview of automatic error correction technology for Chinese texts. Errors in Chinese texts are divided into spelling errors, grammatic errors and semantic errors, and the methods of error correction for these three types are reviewed. Moreover, datasets and evaluation methods of automatic error correction for Chinese texts are summarized. In the end, prospects for the automatic error correction for Chinese texts are raised.

Keywords: automatic correction; spelling errors; grammatical errors; semantic errors; datasets; evaluation indicators

0 引言

中文文本自动校对是自然语言处理技术的一个重要应用方面。随着互联网与信息技术的高速发展，中文文本数量呈爆炸式增长，这对传统的手工校对方式提出了严峻挑战。为了降低手工校对工作量，中文文本自动校对相关的研究工作得到了人们的重点关注。中文文本自动校对研究始于 20 世纪 90 年代，相对于英文文本自动校对研究开始较晚，但其发展速度快且取得了丰硕的研究成果，目前也出现了已经商业化的产品，如黑马校对软件、哈工大讯飞实验室发布的飞鹰智能文本校对系统等。

早期中文自动校对方法主要基于统计和规则相结合的方法[1-2]，采用了分词、统计语言模型、统计机器翻译(Statistical Machine Translation，SMT)和混淆字符集等技术。随着深度学习的发展，一系列端到端的方法在自然语言处理(Natural Language Processing，NLP)领域逐渐得到应用，如循环神经网络(Recurrent Neural Network，RNN)、序列到序列模型(Sequence-to-sequence，Seq2seq)[3-4]、注意力机制[5-6]、卷积序列到序列模型(Convolutional Sequence to Sequence，ConvS2S)[7]和基于自注意力的 Transformer 模型[8]，中文文本自动校对研究逐渐从基于规则和统

收稿日期：2021-10-07 **定稿日期**：2021-11-25

基金项目：国家重点研发计划项目(2018YFB1004100)

计语言模型相结合的方法转向基于深度模型的方法，并且使用序列标注模型、神经机器翻译模型(Neural Machine Translation，NMT)和预训练语言模型进行端到端的校对。

本文概述了中文文本中的常见错误类型，分析了中文文本校对技术的研究发展现状，对中文文本校对共享任务数据集以及校对系统的评估指标进行了归纳总结，最后探讨了中文文本自动校对技术未来发展的方向。

1 中文文本的错误类型

中文文本产生的错误可大体分为拼写错误、语法错误和语义错误三类。

拼写错误 张仰森等人[9-10]和Liu等人[11]指出音似、形似字错误是中文文本中常见的拼写错误。形似字错误主要发生在五笔输入和字符识别(Optical Character Recognition，OCR)过程中，音似错误则主要发生在拼音输入和语音识别(Automated Speech Recognition，ASR)过程中。其中，音似错误又可以进一步细分为同音同调、同音异调和相似音错误[12-13]。虽然大部分拼写错误是由音似、形似字误用导致，但也有些错误是由于缺少常识性知识或语言学知识所导致的，如表1所示。

表1 常见拼写错误举例

错误类型		错误	正确
形似字错误		诞续	延续
音似字错误	同音同调	火势向四周漫(man4)延	火势向四周蔓(man4)延
	同音异调	但是不行(xing2)还是发生了	但是不幸(xing4)还是发生了
	相似音	词青(qing1)标注	词性(xing2)标注
知识型错误		埃及有金子塔	埃及有金字塔
推断型错误		他的求胜欲很强，为了越狱在挖洞	他的求生欲很强，为了越狱在挖洞

语法错误 NLPTEA等[14-20]语法错误校对竞赛将中文文本常见语法错误归纳为字词冗余错误(Redundant words，R)、字词缺失错误(Missing words，M)、搭配不当错误(Selection errors，S)和字词乱序错误(Word ordering errors，W)，如表2所示。

表2 常见语法错误举例

错误类型	错误	正确
字词冗余	我根本不能理解这妇女辞职回家的现象。	我根本不能理解妇女辞职回家的现象。
字词缺失	我河边散步的时候。	我在河边散步的时候。
搭配不当	还有其他的人也受被害。	还有其他的人也受伤害。
字词乱序	世界上每天由于饥饿很多人死亡。	世界上每天很多人由于饥饿死亡。

语义错误 语义错误是指一些语言错误在字词层面和语法搭配上不存在问题，而是在语义层面上的搭配有误[21]，如表3所示。由于语义错误的处理需要模型理解上下文的语义信息，因而对模型提出了较高的要求，其校对难度要高于拼写错误校对和语法错误校对。

表3 常见语义错误举例

错误类型	错误	正确
知识错误	中国的首都是南京	中国的首都是北京
搭配错误	他戴着帽子和皮靴就出门了	他戴着帽子穿着皮靴就出门了

下文中将分别对拼写错误、语法错误和语义错误的自动校对方法进行总结与分析。

2 中文文本自动校对方法

2.1 拼写错误校对方法

中文文本拼写校对流程大致可以分为以下三步：①错误识别：判断文本是否存在拼写错误，并标记出错误位置；②生成纠正候选：利用混淆字符或通过模型生成字符等方法构建错误字符的纠正候选；③评估纠正候选：利用某种评分函数或分类器等，结合局部乃至全局特征对纠正候选排序，排序最高的纠正候选作为最终校对结果。事实上，大部分校对方法的流程都可以划分为上述三步，不过也有部分方法，如基于深度模型端到端的校对方法，将错误识别阶段省略，但本质上也属于此流程。

2.1.1 基于规则和统计语言模型结合的校对方法

中文拼写错误校对早期采用的主要是规则和统计语言模型(Statistical Language Model，SLM)相结合的校对方法，该类方法使用规则和统计语言模型进行检错，在生成候选阶段利用混淆字符或通过

模型生成字符的方式得到纠正候选字符，最后通过统计语言模型进行纠正候选的评估，其中，校对规则主要使用了混淆字符集、基于分词的查错规则和校对词典等，统计语言模型主要使用了 N 元语法（N-gram）、条件随机场（Conditional Random Fields，CRF）等，如表 4 所示。

表 4　基于规则和统计的拼写校对方法

引用	语言	规则	统计模型
[22]，1995	繁体	混淆字符集	Bi-gram
[23]，1998	简体	最长匹配分词	Tri-gram
[24]，2001	简体	—	互信息
[25]，2002		混淆字符集，最小编辑距离	Tri-gram，贝叶斯分类器
[26]，2006	简体	非多字词错误查错规则	互信息
[27]，2012	繁体	形似字符集	Bi-gram，线性回归
[28]，2013	繁体	混淆字符集	Bi-gram，线性回归
[29]，2013	繁体	混淆字符集，E-HowNet	N-gram
[30]，2013	繁体	混淆字符集，混淆字符替换规则	N-gram
[31]，2013	繁体	混淆字符集，校对词典	Tri-gram，CRF
[32]，2013	繁体	混淆字符集	N-gram
[33]，2013	繁体	—	最大熵
[34]，2013	繁体	混淆字符集，词典	SMT，N-gram，SVM
[35]，2013	繁体	校对词典，检错规则	SMT，N-gram
[36]，2014	繁体	混淆字符集	Tri-gram
[37]，2014	繁体	混淆字符集	噪声信道模型，N-gram
[38]，2014	繁体	校对规则	图模型，CRF
[39]，2014	繁体	校对词典，编辑距离，最长匹配分词	HMM，N-gram，SVM
[40]，2015	繁体	混淆字符集	CRF，N-gram
[41]，2015	繁体	—	N-gram
[42]，2016	简体	模式匹配，中文串相似度计算	N-gram
[43]，2017	繁体	模式匹配，E-HowNet，混淆字符集	N-gram

对于规则和统计相结合的校对方法的研究，通常是改进校对流程的不同阶段的方法，可大致分为三类：

错误识别阶段　基于统计语言模型的检错。基于统计语言模型的检错方法通常都需要先对原句进行分词，然后通过统计语言模型与词性标注序列等相结合的方式进行检错，其中统计语言模型主要用到 N-gram 等，如于勐等人[23]提出一种混合校对系统 HMCTC（Hybrid Method for Chinese Text Collation），采用最长匹配分词结合词典的方式将原句分词，然后以 Tri-gram 为基础结合语法属性标注进行检错，将相邻词共现频率低于阈值和语法序列标注不合理的地方标记为错误；张仰森等人[24]提出了一种基于互信息的字词接续判断模型，通过判断相邻字和相邻词的接续性进行检错。早期基于统计语言模型的检错方法通常都需要构建庞大的字字、词词同现频率库，这带来了严重的数据稀疏问题，造成这个问题的原因除了统计模型本身的缺陷外，还因为早期的检错方法没有深度地分析中文分词的特点。张仰森等人[26]通过分析中文文本的特点指出，中文文本大多由二字以上的词构成，分词后出现的连续单字词一般不超过 5 个，且出现的单字词多是助词、介词等，而含有拼写错误的文本分词后会出现连续的不合理的单字散串，并由此提出了“非多字词错误”，在检错时主要针对分词后出现的连续单字词进行判断，字字同现库通过正确文本中的连续单字词同现频率进行构建，减小了同现频率库的规模，缓解了数据稀疏问题；Xie 等人[41]对文本中长度等于 2 和大于 2 的连续单字词分别使用 Bi-gram 和

Tri-gram模型检错。“非多字词错误”概念的提出在中文文本校对方法中有着重要的意义，很多基于统计的检错方法都是针对这种错误特点提出的校对方法[32]，甚至影响了基于深度模型的校对思路[44]。

还有一些文献采用图模型、最大熵模型或CRF等模型检错。Han等人[33]将检错视为二分类问题，他们基于大量原始语料为每个字符训练了一个最大熵模型进行检错；Xin等人[38]在分词的同时构建一个有向无环图，将音似字或校对词典中的词加入到图中构建新的有向无环图模型，在图中寻找最短路径完成校对，如图1所示；刘亮亮等人[42]提出了一种模糊分词的“非多字词错误”自动校对方法，首先利用精确匹配算法与模糊相似度算法对中文文本进行精确切分和模糊全切分，建立词图，然后利用改进的语言模型对词图进行最短路径求解，将查错与纠错融于一体，其思路和Xin等人[38]的大致相同；Wang等人[40]将词向量、词长和词性等作为特征输入CRF进行检错。

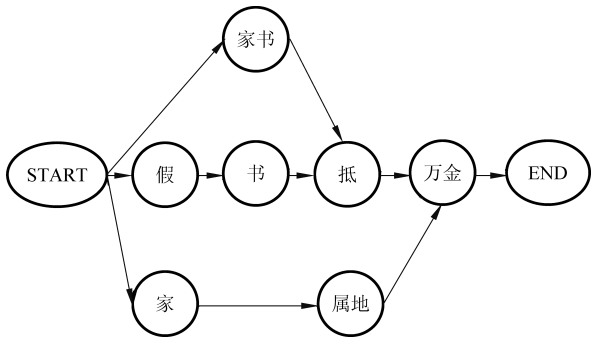

图1 相似词语有向无环图

生成纠正候选阶段 基于混淆字符集和模型生成的候选字符生成。混淆字符集是中文文本拼写错误校对中较为关键的数据，用于存储每个字词可能被混淆的错别字词，如表5所示。基于混淆字符集的候选字符生成方法假定原句中的错别字词是由于字符的视觉相似性或读音相似性导致的，因此，一些方法直接使用混淆字符作为候选字符对句中的错别字符进行替换，如Huang[36]和Chang等人[22]直接使用混淆字符作为候选字符，逐个替换原句字符得到多种修改句，再通过Bi-gram模型计算修改句流畅度，若修改句流畅度高于原句就保留替换结果。这种直接使用混淆字符替换的策略，虽然可以取得较好的召回率，但会造成大量的修改句，导致校对性能较差。为了缓解上述问题，一些研究使用基于规则的方法对混淆字符进行筛选后再得到候选字符，如Lin等人[30]先使用混淆字符替换文本中的连续单字词，形成多种修改句，然后将修改句再次分词，如果原先的连续单字词重新形成长词，就将该混淆字符加入到候选字符中；Yeh等人[43]同Lin的的思路相似，若再次分词后，原先的连续单字词重新形成了新词并出现在E-HowNet中，就将该混淆字符加入到候选字符中；张道行等人[27]考虑了字符在水平结构和垂直结构上的相似度，通过计算错误字符与混淆字符在字形上的相似度进行候选字符筛选；Chang等人[28]在张道行的基础上进一步引入了偏旁特征和读音特征；Yeh等人[29]则是根据混淆字符出现的频率进行倒排序索引，在替换时选择前N个字符作为候选字符。

表5 混淆字符集示例

类型	混淆字符
音似	右，幼鼬誘宥柚祐有侑...
形似	可，何呵坷奇河柯苛阿倚寄崎...

除了混淆字符集，还有一些方法则是通过模型生成候选字符，比较常用的模型有隐马尔科夫模型(Hidden Markov Model，HMM)和统计机器翻译模型，如Xiong等人[39]先使用HMM生成候选字符并形成多种修改句，再通过支持向量机(Support Vector Machine，SVM)的置信度评分对修改句排序，选出两个最佳的候选字符；Liu[34]和Chiu等人[35]则是将文本校对视为一项翻译任务，通过SMT模型直接生成校对结果。

评估纠正候选阶段 基于统计语言模型的候选评估。评估候选阶段主要使用统计语言模型评估校对后语句的流畅度，若流畅度高于原句则保留校对结果。其中N-gram、互信息和困惑度(Perplexity，PPL)等是常用的统计语言模型，如Huang[36]和Chang等人[22]使用Bi-gram评估校对句的流畅度；Lin等人[30]和He等人[32]使用N-gram评估校对结果的流畅度；Zhao等人[45]使用PPL评估校对句的整体流畅度，使用互信息评估校对句的局部流畅度。

从上面的介绍可以看到，在基于规则和统计语言模型相结合的校对方法中，统计语言模型，如N-gram、互信息、SVM、HMM等，在检错和候选字符评估阶段起着重要作用，统计语言模型有着参数易训练、可解释性强等优点，但同也样存在着缺乏长距离依赖、参数空间较大、数据稀疏严重和泛化能力较差等缺点，这也导致基于统计模型的校对方法逐

渐被基于深度模型的校对方法所替代。

除了统计语言模型外，混淆字符集同样在校对过程中起着重要作用，由于中文文本拼写错误通常是由音似、形似字错用导致，所以在纠错阶段通过混淆字符集生成候选字符的方法是不可或缺的，混淆字符集通常包括音似字符集和形似字符集。在基于统计和规则的校对方法中，混淆字符集的质量决定了校对模型的上限，构建一个高质量的混淆字符集对于纠错效果至关重要。

2.1.2　基于深度模型的校对方法

近年来随着深度学习在各领域取得的显著效果，基于深度模型的校对方法受到了更多关注。该类方法主要使用序列标注模型端到端的方式生成校对结果，但这样的校对方法有两个不足：①中文文本常见的拼写错误类型是音似字、形似字错误，虽然一些研究工作通过加入拼音特征和字形特征为模型提供更多的字符信息，但是并没有直接用到混淆字符信息；②序列标注模型大部分是基于字符级建模的，忽略掉了词信息。针对第一个问题，一些研究工作使用混淆字符过滤模型生成的候选字符或者使用深度模型提取混淆字符特征，将混淆字符融入深度模型中；针对第二个问题，一些研究工作通过联合词特征的方法利用词信息，提升校对效果。具体方法如表 6 所示。

表 6　基于深度模型的拼写校对方法

引用	语言	深度模型	特征
[46]，2019	简体	BiLSTM	拼音特征，字形特征
[47]，2019	简体	BiLSTM-CRF	—
[48]，2019	简体	Pointer-generator network	混淆字符矩阵
[49]，2019	简体	Attention LSTM	混淆字符集
[50]，2019	简体/繁体	DAE-CSD	拼音特征，字形特征
[51]，2020	繁体	BSST	—
[52]，2020	简体	BiGRU，Soft-Masked BERT	—
[53]，2020	简体	BERT，GCN	混淆字符集
[54]，2020	繁体	BERT	音似、形似、语义混淆集
[44]，2021	简体	FL-LSTM-CRF	拼音特征，词典
[55]，2021	简体	BERT，GRU	拼音特征，字形特征
[56]，2021	简体	BERT，GRU，ResNet[57]	拼音特征，字形特征
[58]，2021	简体	BERT，CNN	拼音特征，字形特征
[59]，2021	简体	BERT，Tacotron2[60]，VGG[61]	拼音特征，字形特征
[62]，2021	简体	BERT	混淆字符
[63]，2021	简体	BERT	混淆字符集
[64]，2021	简体	RoBERTa	拼音特征

通过深度模型生成候选字符　该类方法主要通过序列标注模型端到端地生成校对结果，如 Duan 等人[47]使用 BiLSTM 提取字特征，通过 CRF 生成候选字符；Han 等人[51]则是将拼音特征、字形特征和字特征拼接后输入 BiLSTM 生成候选字符。早期基于 LSTM 的校对方法很难解决表 1 中提到的知识型错误和推断型错误，这是由于 LSTM 的语义提取能力有限，无法通过上下文环境做出推断。随着 BERT 等预训练语言模型的提出，许多研究工作开始使用 BERT 处理拼写校对任务。相比 LSTM 模型，BERT 有着更强的语义特征提取能力，如沈峻毅等人[51]提出的校对模型 BSST(BERT for Single Sentence Tagging)，其本质是在 BERT 上加入一个线性分类器，先使用 BERT 提取语义信息，再由线性分类器生成候选字符；Zhang 等人[52]为了解决表 1 中提到的知识型错误和推断型错误，提出一种三阶段

校对模型 Soft-Masked BERT，第一阶段是检错模型，使用 BiGRU（Bi-directional Gated Recurrent Unit）得出每个字符存在拼写错误的概率，第二阶段是软遮掩机制，根据检错模型得到的概率“盖住”可能存在错误的字向量，第三阶段将软遮掩后的字向量输入到 BERT 中，通过 BERT 强大的上下文理解能力对遮掩字符进行推断。

为了更好地解决音似、形似字错误，通过加入拼音特征和字形特征提升效果，也是深度模型校对方法中常见的思路。如 Han 等人[46]通过加入拼音和字形统计特征的方法提升模型校对能力。除了统计特征外，还有一些研究工作利用深度模型提取拼音和字形特征，以表 1 中的相似音错误为例，“词青标注”中的“青”字为相似音错误，应被校对为“词性标注”，利用深度模型提取拼音和字形特征的校对流程如图 2 所示。

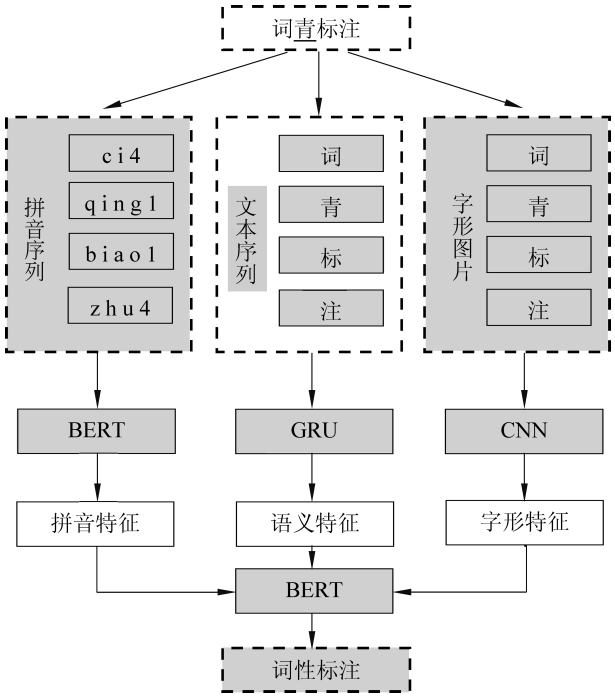

图 2 基于深度模型的拼写校对示例

Xu 等人[56]使用 BERT 提取字特征，GRU＋BERT 提取拼音特征，ResNet[57]提取字形特征，通过门控机制将三种特征融合后输入 BERT；Liu 等人[55]使用 GRU 分别提取拼音特征和字形特征，BERT 提取字特征，将三种特征拼接后输入 BERT；Sun 等人[58]使用 CNN 分别提取拼音特征和字形特征，BERT 提取字特征，将三种特征拼接后输入混淆层进行融合，再将融合后的特征输入 BERT；Huang 等人[59]使用 BERT 提取字特征，Tacotron2[60]提取拼音特征，VGG[61]提取字形特征，将三种特征通过门控机制融合后输入全连接层生成候选；Zhang 等人[62]虽然没有直接加入拼音或字形特征，但在预训练时他们使用字符的音似字对其进行“遮掩”，通过这种方式加强模型对音似字错误的识别能力；Wang 等人[64]提出了一种基于拼音特征增强的候选字符生成方法，使用卷积层提取拼音特征，RoBERTa 提取字特征，将两者拼接后进行标准化操作，得到候选字符评分，随后通过一个动态连接网络选择最佳的候选字符。

使用混淆字符信息筛选候选字符 该类研究主要是通过结合混淆字符特征的方法提升校对效果，主要有两种方法，一种是在输出端通过混淆字符限制候选字符范围，如 Wang 等人[48]将混淆字符矩阵与指针网络（Pointer Network，PN）[65]相结合，在解码端（Decoder）生成候选字符时不再是从整个字表预测，而是从错误字符的混淆字符表中进行预测；Li 等人[63]提出了校对模型 DCSpell，其校对思路和 soft-masked BERT[52]相似，不同的是在生成候选字符时要求模型从混淆字符表中预测。另一种是对混淆字符建模，如 Cheng 等人[53]提出了校对模型 SpellGCN，整个模型由三部分组成，第一部分使用 BERT 提取字特征，第二部分使用一个图模型提取混淆字符特征，第三部分将混淆字符特征与字特征融合后生成候选字符。

联合词信息 该类方法主要是通过在输入端或解码端联合词信息的方法加强校对效果，如 Bao 等人[54]提出了基于组块解码的检错方法，利用全局优化修改单字和多字词错误，其中单字、多字词、短语和成语同等地被视为一个块，基于块的解码器不断地将句子分块，模型假定块中字或词都可能存在错误并通过音似、形似和语义混淆集进行替换生成多种候选句，在候选句排序阶段引入了多种特征参与最终决策，如块数、字符编辑距离、语言模型的困惑值等，并最终用 MERT（Minimum Error Rate Training）[66]算法学习多种特征权重并做出最优选择；Wang 等人[44]则是提出了融入词特征和拼音特征的改进 LSTM 模型 FL-LSTM-CRF，输入时将拼音特征和字向量加权相加后输入 LSTM 单元，改进的 LSTM 删除了输出门，下一时刻的单元值通过当前时刻单元值和词向量加权得到。

2.2 语法错误校对方法

本节主要梳理了基于深度模型的中文文本语法

错误校对方法，主要有基于序列标注模型的方法、基于 NMT 的方法，具体方法见表 7。

表 7　语法错误校对方法

引用	方法
[69]，2016	BiLSTM-CRF
[70]，2017	BiLSTM-CRF
[67]，2017	BiLSTM-CRF
[68]，2018	BiLSTM-CRF
[72]，2018	LSTM-CRF，CRF
[71]，2018	BiLSTM-CRF
[73]，2018	BiLSTM-CRF，校对规则
[88]，2018	Transformer-NMT，*N*-gram，校对规则
[89]，2018	LSTM Seq2seq，SMT，*N*-gram，校对规则
[90]，2018	BiLSTM-CRF，NMT，SMT，校对规则
[81]，2018	ConvS2S
[82]，2019	ConvS2S
[84]，2019	Transformer，*N*-gram，校对规则
[83]，2020	Transformer，LaserTagger[94]，*N*-gram
[74]，2020	BERT-BiLSTM-CRF，Seq2Seq，*N*-gram
[75]，2020	RoBERTa-BiLSTM-CRF
[76]，2020	BERT-BiLSTM-CRF
[77]，2020	BERT
[78]，2020	BERT-ResNet
[79]，2020	BERT
[80]，2020	BERT-GCN-CRF
[87]，2020	Transformer，数据增强
[85]，2020	BERT-fused NMT[86]
[92]，2020	Dynamic-Masked Transformer
[91]，2021	Dropout-Transformer

2.2.1　基于序列标注模型的方法

该类方法将语法检错视为序列标注任务，早期的序列标注模型主要使用 LSTM 和 CRF 来实现，使用 LSTM 模型进行语法检错时，特征的选择十分重要，如 Liao 等人[67]将词向量特征输入到 BiLSTM 模型，通过 BiLSTM 提取句法特征并由 CRF 标记句中的语法错误；Liu 等人[68]将字向量特征输入 BiLSTM-CRF 模型进行语法检错；Soni 等人[69]将字向量特征、字二元向量特征和词性 POS 特征输入 BiLSTM-CRF 进行检错。

除了上述常用的特征，很多研究工作提出了许多新的统计特征，如 Yang 等人[70]将字向量特征、向量字共现特征、词性 POS 特征、词性 POS 概率、点间互信息特征等输入 BiLSTM-CRF；Fu 等人[71]将高斯互信息（ePMI）、POS＋PMI 等联合特征输入 BiLSTM-CRF 模型检错；Zhou 等人[72]将依存句法特征、词特征、POS *N*-gram 特征和词性 POS *N*-gram 特征等联合输入CRF 和 LSTM-CRF 模型检错；Zhang 等人[73]将字特征、词性 POS 特征、字二元特征和字三元特征等输入 BiLSTM 模型检错。这些研究工作的重点在于学习中文语法规律，基于无标注语料统计词语规律和词语用法，并利用相应的特征来提高检错效果，然而其特点是统计特征不能捕获深层的语法和语义信息，难以发现一些隐晦的语法错误。

随着预训练语言模型的提出，许多研究者开始基于 BERT 等预训练语言模型构建 CGEC 模型，如 Zan 等人[74]使用 BERT 提取句子深层语法、语义信息，将提取到的特征输入 BiLSTM-Attention-CRF 模型检错；Han 等人[75]则是使用 RoBERTa 提取语义特征，然后通过 BiLSTM-CRF 检错；Cao 等人[76]提出一种基于门控机制的特征融合方法，将 BERT 特征和经过门控机制筛选的词性 POS 特征等拼接后输入 BiLSTM-CRF 进行检错；Cheng 等人[77]则是直接使用 BERT 代替 LSTM 模型进行检错；Wang 等人[78]提出了 ResBERT 模型，借鉴 ResNet[56]的思想将 BERT 的输入和输出融合后进行语法检错；谢海华等人[79]通过语言学特征的多任务学习，如词性标注和依存句法分析等对 BERT 进行优化，提升语法错误检错效果；Luo 等人[80]在引入 BERT 的基础上还加入了 GCN 进一步提取文本依存结构信息，将 BERT 提取的语义信息与 GCN 提取的结构信息结合后输入 CRF 检错。

2.2.2　基于 NMT 的语法错误校对方法

基于 NMT 的方法将语法校对任务看成由错误句子到正确句子映射的翻译任务，在翻译任务的框架下，NMT 天然地可以使用一套框架同时校对多种类型的错误，但早期的 NMT 方法主要使用 LSTM 模型，与基于序列标注的校对方法一样，也面临特征提取能力有限的问题。同时，基于 NMT 的方法还一直存在训练语料不足的问题。为此，研究者们分别通过加强文本特征的方法提高文本特征提取能力，通过多模型混合校对和优化模型结构的方法缓解训练语料不足的问题。

处理文本特征提取问题　早期的 NMT 模型主

要使用 LSTM 模型进行编码(Encoder),受限于 LSTM 的特征提取能力,同基于序列标注的校对方法中提到的一样,研究者们使用加强文本特征的方式,通过引入字向量特征、词向量特征等传统特征和高斯互信息(ePMI)、词向量共现特征等统计特征来提高 LSTM 的提取效果。近些年随着 Transformer 和 ConvS2S 等模型的提出,模型的特征提取能力有了极大的加强,许多学者逐渐放弃基于 LSTM 的 Seq2Seq 结构,转用 Transformer 和 ConvS2S 模型处理语法错误校对任务,如 Ren 等人[81]和 Li 等人[82]使用 ConvS2S 模型进行语法错误校对;Hinson 等人[83]使用 Transformer 模型生成校对结果;Qiu 等人[84]提出一种两阶段校对方法,先用统计语言模型结合混淆字符集解决拼写错误,再通过 Transfomer 校对语法错误;Liang 等人[85]则使用 BERT-fused[86]处理语法错误校对任务,将 BERT 提取的文本特征融入 Transformer 结构中提升校对效果;王辰成等人[87]基于 Transformer 结构提出一种动态残差结构,动态结合不同的神经模块输出来增强模型的语义捕获能力。

缓解训练语料不足问题 早期基于 LSTM 模型的 NMT 校对方法中,许多研究者提出用多模型混合的方法缓解训练数据不足的问题,如 Fu 等人[88]先通过语言模型和混淆字符集解决句子表面的拼写错误,然后使用平型组合结构(图 3),将待校对句和不同的特征输入到多个 NMT 模型,最后通过 5-gram 语言模型对所有校对结果进行排序,选出

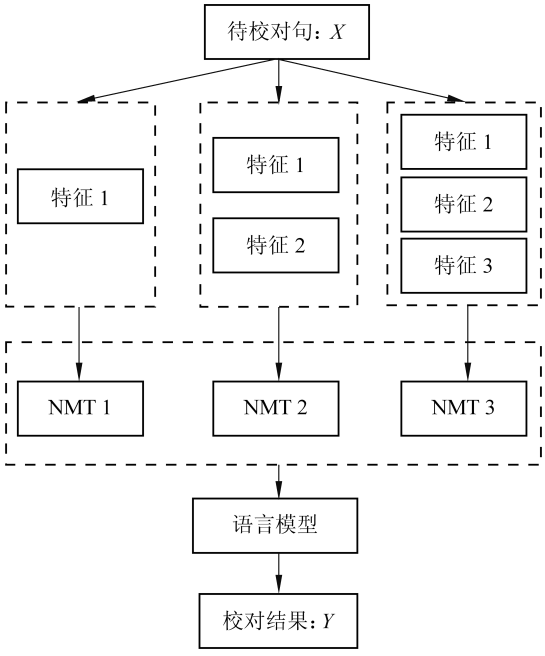

图 3 平型组合结构

最优的校对结果合并到一起输出;Zhou 等人[89]和 Li 等人[90]使用分层组合结构(图 4),将校对流程分为校对模块和结果合并模块,校对模块中将多个 NMT、SMT 和 Rule-based 模型平行组合处理待校对句,并将校对结果送入结果合并模块。结果合并模块分为低级合并与高级合并两个步骤,低级合并是将 NMT、SMT 和 Rule-based 模型的校对结果进行同类模型合并,得到校对候选集;高级合并则是进一步将校对候选集进行合并得到最终结果。近年来,研究者们逐渐采用 Transformer 等模型替换 LSTM,虽然 Transformer 等 NMT 模型有较好的校对效果,但复杂的模型结构和庞大的参数量需要更多的训练数据。面对缺少训练数据的问题,研究者们通过优化模型结构的方法进行改善,如汤泽成等人[91]提出一种基于源端词 Dropout 的 Transformer 模型,通过在输入端引入 Dropout 操作减少编码器对源端错误信息的依赖,增强模型的泛化能力,减少对数据的依赖;Zhao 等人[92]在训练阶段通过多轮动态遮掩输入实例的方法,在不依赖更多数据的情况下增强模型的校对能力;Qiu 等人[84]和 Hinson 等人[83]则是受 Ge 等人[93]的启发,使用循环生成的校对策略将 Transformer 生成的校对结果再次输入 Transformer,如此多轮迭代并记录每次的校对结果,最后在多轮校对结果中,取流畅度最高的作为最终结果。

不论是基于序列标注模型的方法还是基于神经机器翻译的方法,目前的语法校对主要是基于对句子的浅层分析,并不涉及对语义级别的深层理解,这就使得在涉及语义理解的语法错误校对时,机器的能力要远弱于人类专家的水平。但对于大多数如表 2 所示的语法错误,机器自动校对语法错误的能力已经达到人类的水平。

2.3 语义错误校对方法

中文文本语义错误校对一直是中文文本自动校对的难点。早期的中文文本语义错误校对方法主要是基于规则和统计模型相结合的方法,如骆卫华等人[95]在 2003 年提出一种基于实例、基于统计和基于规则的语义搭配关系相结合的方法进行语义检错,通过对待校对文本进行句法分析得到句子成分,再在语料库中寻找与句子结构相似的实例,计算两者的相似度,检查长距离的语义搭配错误,同时利用词义邻接矩阵等检查局部语义搭配错误。

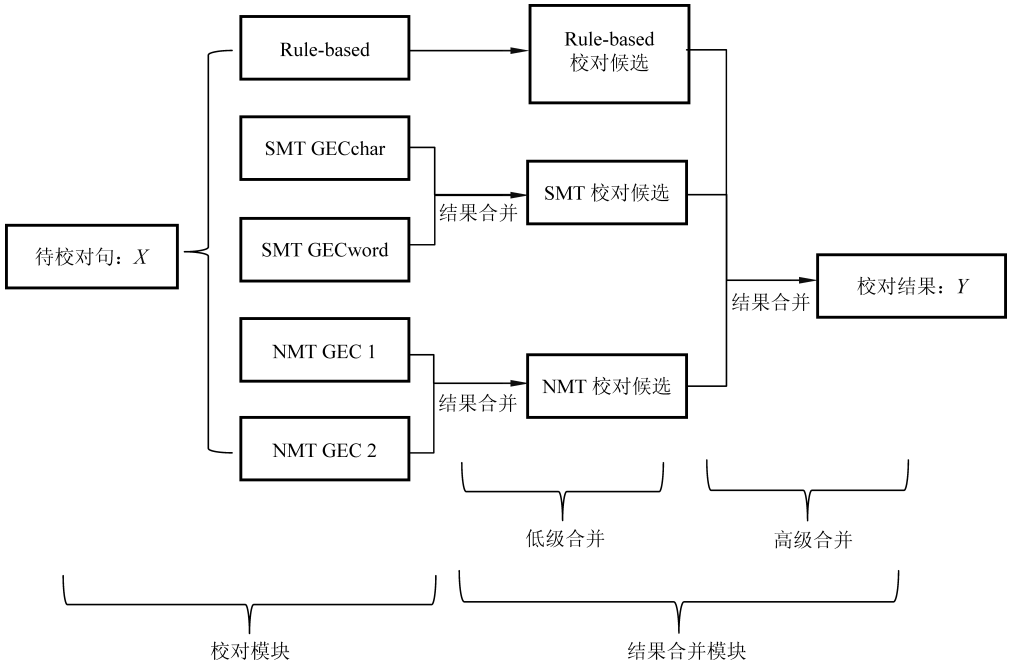

图 4　分层组合结构

从 2009 年开始，语义检错方法主要基于语义知识库，如程显毅等人[96]基于 HNC（概念层次网络）构建了一个校对系统模型，该系统利用句类知识库和概念关联知识库对待校对文本进行句类检查并构成语义块，借助概念关联知识库对语义块的语义成分进行分析，检查语义错误，但由于 HNC 本身没有语义推理机制，需要穷举 57 种句式语义来判断语义错误，并不适合大规模语义校对；2010 年张仰森等人[97]提出一种基于《知网》义原搭配的检错方法，构建了动词-名词搭配关系知识库，并通过知识库检测句中的动词-名词语义搭配错误。构建知识库时，统计语料中动词-名词的二元搭配组合，利用《知网》对词语义项和义原的描述，将动词与名词的二元搭配组合转变为义原间相互制约的多元组合，得到语义知识库，并通过语义知识库检测是否存在动词-名词语义搭配错误；2012 年张仰森等人[98]又提出基于知识库的多层级中文文本查错推理模型，通过关联词关系库和推理规则检查语法错误，通过语义搭配知识库和义原搭配知识库检测句中的语义搭配错误；2017 年张仰森等人[99]进一步提出基于语义搭配知识库和证据理论的检错模型，在语义搭配知识库的基础上利用证据理论的不确定性推理方法确定语义搭配的关联强度，以此判断是否存在语义搭配错误；同年姜赢等人[21]面对政治新闻领域提出一种基于描述逻辑本体推理的语义校对方法，利用本体技术提取文本中的语义内容并将其转化为结构化本体，再与领域背景知识库融合，通过描述逻辑推理机来判断语义内容是否存在逻辑一致性错误。

从上述介绍文字可以看到，语义知识库在中文文本语义错误校对中起着重要作用，从早期简单的动词-名词搭配语义知识库逐渐丰富到多种句子成分搭配知识库，《知网》和《HowNet》等词语义原抽取知识库将词语信息转换成相应的语义表示，在语义知识库的构建中又起着重要的作用，除了语义知识库，描述逻辑推理机等推理方法进一步提高语义检错的效果，如姜赢等人[21]和张仰森等人[99]都在语义知识库的基础上通过推理模型进一步判断句中是否存在语义搭配错误。

3　中文文本校对数据集

无论是统计模型还是深度模型，想要得到较好的结果，都需要高质量的数据集，高质量的数据集往往能够提高模型训练的质量和预测的准确率。在缺乏数据的情况，可以优先考虑公开数据集，特别是得到公认的被普遍使用的数据集。中文文本校对公开数据集主要来自 SIGHAN[100-103]和 NLPTEA[14-20]等竞赛提供的数据集，该类数据集由手工标注，质量较高。

3.1　拼写错误校对数据集

拼写错误校对共享任务公开数据集如表 8 所示。

表 8 拼写错误校对共享任务公开数据集

数据集	组成	句/篇	平均句长	错误	主题
SIGHAN-2013[100]	Train	700 句	—	350	台湾中学生作文
	Test	1 000 句	68.711	376	
	Valid	1 000 句	74.328	1 265	
SIGHAN-2014[101]	Train	1 301 篇	—	5 284	TOCFL
	Test	1 062 句	50.11	792	
SIGHAN-2015[102]	Train	970 篇	—	3 143	TOCFL
	Test	1 100 句	—	715	
NLPTEA-2017[103]	Train	1 000 句	—	—	台湾中学生作文
	Test	1 000 句	—	—	

SIGHAN-2013 是第一个举办的中文文本拼写错误校对竞赛，竞赛数据来自台湾经过手工标注的中学生的作文。除了数据集外，SIGHAN-2013 还整理并公开了一个由音似字和形似字组成的混淆字符集，这对之后的中文文本拼写错误校对发展起着重要的作用。

SIGHAN-2014 和 SIGHAN-2015 数据来自台湾华语文能力测验作文（Test of Chinese as a Foreign Language，TOCFL），每篇作文含多个拼写错误并全部经过手工标注。

NLPTEA-2017 数据来自香港小学生作文，由香港大学完成标注和校对。主办方从校对好的数据中选出长度合理、易于理解、不含歧义的 6 890 句与另外 3 110 条不含拼写错误的句子构成基础数据集，最后从基础数据集中随机选取各 1 000 句组成训练集和测试集。

中文拼写错误校对竞赛举办次数较少，提供的数据集规模较小，当使用深度模型解决拼写错误校对任务时，可以使用多个竞赛的数据集扩充数据量。可以看到，全部竞赛提供的数据集都由繁体中文构成，繁体中文与简体中文在文本组织结构、表述方式和用词习惯等方面有着较大的差异，如“幼稚园”在简体中文构成的文本中通常被写作“幼儿园”，在面向简体中文建模时，直接使用工具将繁体中文转换为简体中文则不合适。

3.2 语法错误校对数据集

语法错误校对共享任务数据集共包含字词缺失、字词冗余、搭配不当和字词乱序四种类型的错误，常用数据集如表 9 所示。

表 9 语法错误校对共享任务公开数据集属性和统计

数据集	组成	句/篇	错误	M	R	S	W	语言	主题
NLPTEA-2014[14]	Train	5 670 句	5 670 句	—	—	—	—	繁体	TOCFL
	Test	1 750 句	875 句	350	279	126	120		
NLPTEA-2015[15]	Train	2 205 句	2 205 句	620	279	126	120	繁体	TOCFL
	Test	1 000 句	1 000 句	126	132	110	132		
NLPTEA-2016[16]	Train	10 693 句	24 492 个	8 739	4 472	9 897	1 384	繁体	TOCFL
	Test	3 528 句	4 103 个	1 482	782	1 613	226		
	Train	10 071 句	24 797 个	6 623	5 538	10 949	1 687	简体	HSK
	Test	3 011 句	3 695	991	802	1 620	282		

续表

数据集	组成	句/篇	错误	M	R	S	W	语言	主题
IJCNLP-2017[17]	Train	10 449 篇	26 448 个	7 010	5 852	11 591	1995	简体	HSK
	Test	3 154 篇	4 876 个	1 247	1 062	2 155	385		
NLPCC-2018[18]	Train	717 241 句	300 004 句	—	—	—	—	简体	TOCFL
	Test	2 000 句	—	—	—	—	—		
NLPTEA-2018[19]	Train	402 篇	1 067 个	298	208	474	87	简体	HSK
	Test	3 549 句	5 040 句	1 381	1 119	2 167	373		
NLPTEA-2020[20]	Train	1 129 篇	2 909 个	801	678	1 228	201	简体	HSK
	Test	1 457 篇	3 654 个	864	769	1 694	327		

NLPTEA-2014 和 NLPTEA-2015 的数据来自 TOCFL，每篇文章含多个平行句对，每个平行句对由含有语法错误的病句和对应的校对句组成。

NLPTEA-2016 是第一个提供简体中文数据集的语法错误校对竞赛，繁体中文数据集来自 TOCFL，简体中文数据集来自汉语水平考试（Hanyu Shuiping Kaoshi，HSK）。从 NLPTEA-2016 开始，IJCNLP-2017、NLPCC-2018、NLPTEA-2018 和 NLPTEA-2020 等语法错误校对竞赛开始提供简体中文数据集。可以看到，中文语法错误校对竞赛提供的数据集逐渐从繁体中文转向简体中文，竞赛举办的频次也越来越高，人们对中文文本自动校对的需求在日益增长，中文文本自动校对研究逐渐受到更多的关注。

3.3　自动生成校对数据集

现阶段的中文文本校对主要基于深度模型的方法，由于深度模型的训练依赖大量高质量的标注数据，仅靠共享任务提供的数据集并不能满足需求，因此许多学者提出了自动生成中文校对数据集的方法。

拼写校对数据集　拼写校对数据主要生成音似、形似字错误数据，因为音似、形似字错误是中文文本中最常见的错误类型[11]。数据生成的主要思路是通过 OCR 工具、ASR 工具或规则等方法得到含有音似、形似字错误的句子，再进一步通过规则进行筛选，保留符合真实错误分布的句子，如 Wang 等人[12]使用 OCR 生成形似字错误数据，ASR 生成音似字错误数据。在生成形似字错误数据时，从正确句子中随机选取 1～2 个字符转换成图像并对部分区域进行高斯模糊[104]，然后使用 OCR 工具识别，若出现字符识别错误，则进一步计算错误字符与原字符的笔划相似度，保留笔划相似度大于阈值的句子作为形似字错误数据。在生成音似字错误数据时，通过 ASR 工具识别句子语音信息，若出现字符识别错误，则进一步通过规则进行数据筛选，保留与原句长度相同、错误字符个数小于 3 且错误字符与原字符拼音相似度大于阈值的句子作为音似字错误数据；Duan 等人[13]将音似字错误分为同音字错误和相似音字错误。在生成同音字错误数据时，先将句子分词，选择短词进行拼音提取并通过拼音得到多组同音字词，然后用读音相似度最高的字词替换原字词，得到同音字错误数据。生成相似音字错误数据的思路与 Wang 等人[12]相似，但 Duan 等人[13]选择保留拼音编辑距离在 1～2 的作为错误字符替换原字符得到相似音字错误数据。

语法校对数据集　中文语法校对数据主要采用数据增强的方法生成错误数据，其主要思路是通过对原句进行添加、删除、替换和乱序操作得到含有语法错误的句子，如汤泽成等人[91]提出了融合字词粒度噪声的数据增强方法，在原本词粒度的基础上加入了字粒度的添加、删除、替换和乱序操作；王辰成等人[87]提出一种基于腐化语料的单语数据增强方法，按照训练语料的错误率，通过随机增加、删除、替换字词和打乱字词顺序的方法生成错误样本。上述方法采用随机的方式构造语法错误样本，往往显得不够真实，甚至会破坏句子的语义信息和句子结构，因此一些研究工作设计了规则以保证改造的句子不会发生语义改变，如 Zhang 等人[73]统计已有训练语料的错误分布，通过近似词替换等规则构造相应的错误样本；谢海华等人[79]基于句法分析与预训练语言模型构造语料，通过随机添加、删除或打乱语句的句子成分的方法构造成分冗余、成分缺失和语序不当错误，用词不当错误则是随机遮掩一个词语，然后

用 BERT 预测出的候选字符替换原字符。同时，为了保证改造句的基本语义信息和结构不变，在改造时不对命名实体进行修改，根据句子长度控制错误个数。构造成分缺失和语序不当时，规避修改远距离依赖的结构成分，用词不当错误不对语气词等无意义词语进行修改。

4 中文文本校对系统评估指标

中文文本拼写校对和语法校对任务常用的评估指标混淆矩阵相同，但混淆矩阵在各自任务中的含义不同，详见下两节。

4.1 拼写错误校对评估指标

拼写错误校对评估混淆矩阵如表 10 所示。其中 TP(True Positive)表示被正确识别为错误的字符数；FP(False Positive)为被错误识别为错误的字符数；TN(True Negative)则代表被正确识别为无错误字符的字符数；FN(False Negative)表示未识别到的错误字符数。

拼写错误校对任务通常分为检错和纠错两个子任务，检错任务需要标记所有错误字符的位置，纠错则需要给出所有错误字符的标准校对结果。即拼写校对任务通常分为以下两个级别：

(1) 检错：标记出句中所有错误字符的位置；

(2) 纠错：给出所有错误字符的标准纠正答案。

拼写检错通常只需要对系统的检错性能做出评估，而拼写纠错则需要评估系统的全部两个指标。在混淆矩阵的辅助下，两个级别的系统性能评估指标如式(1)～式(5)所示，即误报率、精确率、准确率、召回率和 F_1 值。

表 10 评估指标混淆矩阵

混淆矩阵		系统预测	
		Positive	Negative
标准答案	Positive	TP	FN
	Negative	FP	TN

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{1}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

$$F_1 = \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}} \tag{5}$$

4.2 语法错误校对评估指标

语法错误校对评估混淆矩阵同拼写错误相同，如表 10 所示。其中 TP(True Positive)是系统正确识别有语法错误的句子数；FP(False Positive)是被系统错误识别为有语法错误的句子数；TN(True Negative)是正确识别无语法错误的句子数；FN(False Negative)是错误的识别为正确的具有语法错误的句子数。

语法错误校对通常也分为检错和纠错两个子任务，检错又可以进一步细分为检错、识别错误类型和定位错误三个级别，而纠错任务则要求模型不仅能定位到错误位置，还要能给出正确的校对结果。即语法错误校对任务通常分为以下四个级别：

(1) 检错：检测语句是否含有语法错误；

(2) 识别错误：语句含有的语法错误的类型；

(3) 定位错误：语句含有的语法错误的发生位置；

(4) 纠错：对于选择(S)和缺失(M)的错误类型需要纠错，每个 S 或 M 类型错误一般允许生成 3 个纠错结果。

语法错误检测任务通常只对系统前三个级别的成绩进行评估，而语法纠错任务需要评估全部四个级别。在混淆矩阵的辅助下，评估系统校对性能如式(2)～式(5)所示，即精确率、准确率、召回率和 F_1 值。

5 中文文本自动校对展望与总结

近年来中文文本自动校对研究越来越受到研究者的关注，基于深度模型的校对方法逐渐成为主流。回顾过去的工作，本文认为以下几个方面会是今后一段时间内研究者们持续探索和关注的问题：

(1) 自从基于深度模型的方法引入文本校对领域之后，已取得了相当大的成功，成为文本自动校对研究的主流方法，对基于深度模型的文本校对研究尤其是基于 BERT 预训练语言模型的文本校对研究会持续受到关注。

(2) 目前中文文本自动校对研究主要围绕 BERT 和 Seq2Seq 结构的 Transformer 模型展开探

索，这些模型参数多、规模大、计算速度慢，难以满足搜索引擎搜索和语音交互等实时场景。如何在效果损失较小的情况下缩小模型规模、缩短迭代周期、加快预测速度是一个重要的研究方向，如 Sanh 等人[105]提出了 LTD-BERT 模型(Learning to Distill BERT)对 BERT 进行了模型压缩，在效果损失很小的基础上，降低了存储和运算开销。

(3) 现有的文本自动校对研究主要面向通用领域，随着无纸化办公的普及，针对不同领域具体场景下的文本校对需求迫在眉睫，将受到越来越多研究人员的关注。具体应用场景下的文本校对通常需要在传统校对的基础上进行更加有针对性的建模，以公文领域为例，张仰森等人[106]指出政治新闻领域存在的文本错误除常见的拼写、语法错误以外，还有领导人顺序错误和领导人姓名-职务对应错误等，针对政治新闻等领域的文本校对，需要分析领域错误特点，单独构建领域词典。

(4) 现阶段中文文本语法错误校对方法主要还是基于 Seq2Seq 的 NMT 方法，通常生成模型需要大规模的平行语料进行训练，而语法纠错相关的语料则比较匮乏，因此如何自动构建大量中文语法校对训练语料将受到更多学者的关注。目前针对语法校对训练数据不足的问题，部分英文语法校对的研究者提出通过构造伪数据的方法来增加训练数据，如 Ge 等人[107]将正确语句输入 Seq2Seq，将错误语句作为输出，训练得到一个错误语句生成模型；Lichtarge 等人[108]使用翻译系统将英文翻译成一种中间语言，如日语、法语等，再将中间语言翻译回英文，生成的英语语义和原始英语语句基本保持不变，但是往往会存在一些语法错误。中文语法校对也可以参考上述办法构造大规模平行语料。

(5)语义问题的研究一直是 NLP 研究中的薄弱环节，也是中文文本校对的难点[95]，已有的语义错误校对方法主要是基于规则、知识库和语义推理的方法[21,96,98]。基于规则、知识库等的校对方法需要人工建立规则，整理领域词典，不适用于大规模的语义错误校对，随着深度学习的不断发展，如何通过深度学习的方法解决语义错误会持续受到学者们的关注。

中文文本自动校对作为自然语言处理领域一个重要研究方向，一直以来受到相当广泛的关注。本文主要阐述了中文文本拼写错误和语法错误的校对方法，整理了相关共享任务数据集，并对未来的研究方向进行了分析和展望。

参考文献

[1] 徐连诚，石磊. 自动文字校对动态规划算法的设计与实现[J]. 计算机科学，2002，29(9)：149-150.

[2] 龚小谨，罗振声，骆卫华. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用，2003，39(8)：98-100.

[3] Cho K，Van Merrienboer B，Gulcehre C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg，PA，USA：Association for Computational Linguistics，2014：1724-1734.

[4] Sutskever I，Vinyals O，Le Q V. Sequence to sequence learning with neural networks[C]//Proceddings of the 27th International Conference on Neural Information Processing Systems. Cambridge，USA：MIT Press，2014：3104-3112.

[5] Bahdanau D，Cho K，Bengio Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of 3rd International Conference on Learning Representations. San Diego，United States：International Conference on Learning Representations，2015：940-1000.

[6] Luong T，Pham H，Manning C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg，PA，USA：Association for Computational Linguistics，2015：1412-1421.

[7] Gehring J，Auli M，Grangier D，et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning. United States：JMLR，2017：2029-2042.

[8] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook，NY，USA：Curran Associates Inc，2017：6000-6010.

[9] 张仰森、丁冰青. 中文文本自动校对技术现状及展望[J]. 中文信息学报，1998(301)：51-57.

[10] 张仰森，俞士汶. 文本自动校对技术研究综述[J]. 计算机应用研究，2006，23(6)：8-12.

[11] Liu C L，Lai M H，Tien K W，et al. Visually and phonologically similar characters in incorrect Chinese words[J]. ACM Transactions on Asian Language Information Processing，2011，10(2)：1-39.

[12] Wang D, Song Y, Li J, et al. A hybrid approach to automatic corpus generation for Chinese spelling check[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 2517-2527.

[13] Duan J, Pan L, Wang H, et al. Automatically build corpora for Chinese spelling check based on the input method[C]//Proceedings of the 8th Natural Language Proceedings and Chinese Computing. Cham, Swizerland: Springer, 2019: 471-485.

[14] Yu L C, Lee L H, Chang L P. Overview of grammatical error diagnosis for learning Chinese as a foreign language[C]//Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications. Nara, Japan: Asia Pacific Society for Computers in Education, 2014: 42-47.

[15] Lee L H, Yu L C, Chang L P. Overview of the NLP-TEA shared task for Chinese grammatical error diagnosis [C]//Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 1-6.

[16] Lee L, Rao G, Yu L, 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. Osaka, Japan: Natural Language Processing Techniques for Educational Applications, 2016: 40-48.

[17] Rao G, Zhang B, Xun E, et al. IJCNLP-2017 task 1: Chinese grammatical error diagnosis[C]//Proceedings of the IJCNLP, Shared Tasks. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017: 1-8.

[18] Zhao Y, Jiang N, Sun W, et al. Overview of the NLPCC shared task: Grammatical error correction [C]//Proceedings of the 7th CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, PEOPLES R CHINA: Springer, Cham, 2018: 439-445.

[19] Rao G, Gong Q, Zhang B, 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. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 42-51.

[20] Rao G, Gong Q, Zhang B, et al. Overview of NLPTEA shared task for Chinese grammatical error diagnosis[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 25-35.

[21] 姜赢, 庄润钹, 吴烨凡, 等. 基于描述逻辑本体推理的语义级中文校对方法[J]. 计算机系统应用, 2017, 26(4): 224-229.

[22] Chang C H. A new approach for automatic Chinese spelling correction[C]//Proceedings of Natural Language Processing Pacific Rim Symposium. Japan: Information Processing Society of Japan, 1995: 278-283.

[23] 于勐, 姚天顺. 一种混合的中文文本校对方法[J]. 中文信息学报, 1998, 12(2): 32-37.

[24] 张仰森, 丁冰青. 基于二元接续关系检查的字词级自动查错方法[J]. 中文信息学报, 2001, 15(3): 37-44.

[25] Li J, Wang X. Combining trigram and automatic weight distribution in Chinese spelling error correction[J]. Journal of Computer Science and Technology, 2002, 17(6): 915-923.

[26] 张仰森, 曹元大, 俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. 中文信息学报, 2006, 20(4): 3-9.

[27] 张道行, 苏守彦. 字形相似别字之自动校正方法[C]//Proceedings of the 24th Conference on Computational Linguistics and Speech Processing. Taiwan: The Association for Computational Linguistics and Chinese Language Processing, 2012: 125-139.

[28] Chang T, Chen H, Tseng Y H, et al. Automatic detection and correction for Chinese misspelled words using phonological and orthographic similarities[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 97-101.

[29] Yeh J F, Li S F, Wu M R, et al. Chinese word spelling correction based on N-Gram ranked inverted index list[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 43-48.

[30] Lin C J, Chu W C. NTOU Chinese spelling check system in SIGHAN bake-off [C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 102-107.

[31] Wang Y. Conditional random field-based parser and language model for traditional Chinese spelling checker[C]//Proceedings of the 7th SIGHAN Workshop

on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 69-73.

[32] He Y, Fu G. Description of HLJU Chinese spelling checker for SIGHAN bakeoff[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 84-87.

[33] Han D, Chang B. A maximum entropy approach to Chinese spelling check[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 74-78.

[34] Liu X, Cheng F, Luo Y, et al. A hybrid Chinese spelling correction using language model and statistical machine translation with reranking[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 54-58.

[35] Chiu H W, Wu J C, Chang J S. Chinese spelling checker based on statistical machine translation[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 49-53.

[36] Huang Q, HuanG P, Zhang X, et al. Chinese spelling check system based on tri-gram model[C]//Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Wuhan, China: Association for Computational Linguistics, 2014: 173-178.

[37] Chiu H, Wu J C, Chang J S. Chinese spell checking based on noisy channel model[C]//Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Wuhan, China: Association for Computational Linguistics, 2014: 202-209.

[38] Xin Y, Zhao H, Wang Y, et al. An improved graph model for Chinese spell checking[C]//Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 157-166.

[39] Xiong J, Zhang Q, Hou J, et al. Extended HMM and ranking models for Chinese spelling correction[C]//Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Wuhan, China: Association for Computational Linguistics, 2014: 133-138.

[40] Wang Y R, Liao Y F. Word vector/conditional random field-based Chinese spelling error detection for SIGHAN-2015 evaluation[C]//Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing. Beijing, China: Association for Computational Linguistics, 2015: 46-49.

[41] Xie W, Huang P, Zhang X, et al. Chinese spelling check system based on *N*-Gram model[C]//Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing. Beijing, China: Association for Computational Linguistics and Asian Federation of Natural Language Processing, 2015: 128-136.

[42] 刘亮亮，曹存根. 中文“非多字词错误”自动校对方法研究[J]. 计算机科学，2016，43(10)：200-205.

[43] Yeh J F, Chang L T, Liu C Y, et al. Chinese spelling check based on N-Gram and string matching algorithm[C]//Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017: 35-38.

[44] Wang H, Wang B, Duan J, et al. Chinese spelling error detection using a fusion lattice LSTM[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(2): 1-11.

[45] Zhao H, Cai D, Xin Y, et al. A hybrid model for Chinese spelling check[J]. ACM Transactions on Asian and Low Resource Language Information Processing, 2017, 16(3): 1-22.

[46] Han Z, Lv C, Wang Q, et al. Chinese spelling check based on sequence labeling[C]//Proceedings of International Conference on Asian Language Processing. Shanghai, China: IEEE, 2019: 373-378.

[47] Duan J, Wang B, Tan Z, et al. Chinese spelling check via bidirectional LSTM-CRF[C]//Proceedings of IEEE 8th Joint International Information Technology and Artificial Intelligence Conference. Chongqing, China: IEEE, 2019: 1333-1336.

[48] Wang D, Tay Y, Zhong L. Confusionset-guided pointer networks for Chinese spelling check[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 5780-5785.

[49] Wang Q, Liu M, Zhang W, et al. Automatic proofreading in Chinese: detect and correct spelling errors in character-level with deep neural networks[M]. Lecture Notes in Computer Science. Springer International Publishing, 2019: 349-359.

[50] Hong Y, Yu X, He N, et al. FASPell: A fast, adaptable, simple, powerful Chinese spell checker

based on DAE-Decoder paradigm[C]//Proceedings of the 5th Workshop on Noisy User-Generated Text. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 160-169.

[51] 沈峻毅, 张道行. 基于 BERT 任务模型之低误报率中文别字侦测模型[C]//Proceddings of the 32nd Conference on Computational Linguistics and Speech Processing. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing, 2020: 319-330.

[52] Zhang S, Huang H, Liu J, et al. Spelling error correction with soft-masked bert[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 882-890.

[53] Cheng X, Xu W, Chen K, et al. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 871-881.

[54] Bao Z, Li C, Wang R. Chunk-based Chinese spelling check with global optimization[C]//Proceddings of the Association for Computational Linguistics: EMNLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 2031-2040.

[55] Liu S, Yang T, Yue T, et al. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2991-3000.

[56] Xu H D, Li Z, Zhou Q, et al. Read, Listen, and See: Leveraging multimodal information helps Chinese spell checking[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 716-728.

[57] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceddings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vgas, NV, USA: IEEE, 2016: 770-778.

[58] Sun Z, Li X, Sun X, et al. Chinese BERT: Chinese pretraining enhanced by Glyph and Pinyin information [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2065-2075.

[59] Huang L, Li J, Jiang W, et al. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 5958-5967.

[60] Shen J, Pang R, Weiss R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2018: 4779-4783.

[61] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. arXiv preprint arXiv: 1409.1556, 2014.

[62] Zhang R, Pang C, Zhang C, et al. Correcting Chinese spelling errors with phonetic Pre-Training[C]//Proceddings of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2250-2261.

[63] Li J, Yin D, Wang H, et al. DCSpell: A detector-corrector framework for Chinese spelling error correction[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, 2021: 1870-1874.

[64] Wang B, Che W, Wu D, et al. Dynamic connected networks for Chinese spelling check[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2437-2446.

[65] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1073-1083.

[66] Och F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003: 160-167.

[67] Liao Q, Wang J, Yang J, et al. YNU-HPCC at IJCNLP task 1: Chinese grammatical error diagnosis using a bi-directional LSTM-CRF model[C]//Proceed-

ings of the 8th International Joint Conference on Natural Language Processing. Taipei, Tiwan: Asian Federation of Natural Language Processing, 2017: 73-77.

[68] Liu Y, Zan H, Zhong M, et al. Detecting simultaneously Chinese grammar errors based on a BiLSTM-CRF model[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 188-193.

[69] Soni M, Thakur J S. Chinese grammatical error diagnosis with long short-term memory networks[C]//Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Application. Osaka, Japan: The COLING Organizing Committee, 2016: 49-56.

[70] Yang Y, Xie P, Tao J, et al. Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into LSTMs for Chinese grammatical error diagnosis task [C]//Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017: 41-46.

[71] Fu R, Pei Z, Gong J, et al. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 52-59.

[72] Zhou Y, Shao Y. Chinese grammatical error diagnosis based on CRF and LSTM-CRF model[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 165-171.

[73] Zhang Y, Hu Q, Liu F, et al. CMMC-BDRC solution to the NLP-TEA Chinese grammatical error diagnosis task[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 180-187.

[74] Zan H, Han Y, Huang H, et al. Chinese grammatical errors diagnosis system based on BERT at NLPTEA CGED shared task[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 102-107.

[75] Han Y, Yan Y, Han Y, et al. Chinese grammatical error diagnosis based on RoBERTa-BiLSTM-CRF model[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 97-101.

[76] Cao Y, He L, Ridley R, et al. Integrating BERT and score-based feature gates for Chinese grammatical error diagnosis[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 49-56.

[77] Cheng Y, Duan M. Chinese grammatical error detection based on bert model[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 108-113.

[78] Wang S, Wang B, Gong J, et al. Combining resnet and transformer for Chinese grammatical error diagnosis[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 36-43.

[79] 谢海华，陈志优，程静，等. 基于数据增强和多任务特征学习的中文语法错误检测方法[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics. Haikou, China: Chinese Information Processing Society of China, 2020: 761-770.

[80] Luo Y, Bao Z, Li C, et al. Chinese grammatical error diagnosis with graph convolution network and multi-task learning[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020: 44-48.

[81] Ren H, Yang L, Xun E. A sequence to sequence learning for Chinese grammatical error correction [C]//Proceddings of the CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2018: 401-410.

[82] Li S, Zhao J, Shi G, et al. Chinese grammatical error correction based on convolutional sequence to sequence model[J]. IEEE Access, 2019, 7: 72905-72913.

[83] Hinson C, Huang H H, CHEN H H. Heterogeneous recycle generation for Chinese grammatical error correction[C]//Proceedings of the 28th International Conference on Computational Linguistics. Strouds-

burg, PA, USA: International Committee on Computational Linguistics, 2020: 2191-2201.

[84] Qiu Z, Qu Y. A two-stage model for Chinese grammatical error correction[J]. IEEE Access, 2019, 7: 146772-146777.

[85] Liang D, Zheng C, Guo L, et al. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistic, 2020: 57-66.

[86] Zhu J, Xia Y, Wu L, et al. Incorporating BERT into neural machine translation[C]//Proceddings of the 8th International Conference on Learning Representations. New Orleans, America: Neural Information Processing Systems, 2020: 1-18.

[87] 王辰成，杨麟儿，王莹莹，等. 基于 Transformer 增强架构的中文语法纠错方法[J]. 中文信息学报，2020，34(6)：106-114.

[88] Fu K, Huang J, Duan Y. Youdao's winning solution to the NLPCC task 2 challenge: A neural machine translation approach to Chinese grammatical error correction[C]//Proceddings of the Natural Language Processing and Chinese Computing. Cham, Switzerland: Springer, 2018: 341-350.

[89] Zhou J, Li C, Liu H, et al. Chinese grammatical error correction using statistical and neural models [C]//Proceddings of the Natural Language Processing and Chinese Computing. Cham, Switzerland: Springer, 2018: 117-128.

[90] Li C, Zhou J, Bao Z, et al. A hybrid system for Chinese grammatical error diagnosis and correction[C]// Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 60-69.

[91] 汤泽成，纪一心，赵怡博，等. 基于字词粒度噪声数据增强的中文语法纠错[C]//Proceedings of the 20th Chinese National Conference on Computational Linguistics. Huhhot, China: Technical Committee on Computational Linguistics, 2021: 813-824.

[92] Zhao Z, Wang H. MaskGEC: Improving neural grammatical error correction via dynamic masking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(01): 1226-1233.

[93] Ge T, Wei F, Zhou M. Reaching human-level performance in automatic grammatical error correction: an empirical study[J/OL]. arXiv preprint arXiv: 1807.01270, 2018.

[94] Malmi E, Krause S, Rothe S, et al. Encode, tag, realize: High-precision text editing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 5053-5064.

[95] 骆卫华. 中文文本自动校对的语义级查错研究[J]. 计算机工程与应用，2003，12：1-4.

[96] 程显毅，孙萍，朱倩. 基于 HNC 的中文文本校对系统模型的研究[J]. 微电子学与计算机，2009，26(10)：49-52.

[97] 郭充，张仰森. 基于《知网》义原搭配的中文文本语义级自动查错研究[J]. 计算机工程与设计，2010，31(17)：3924-3928.

[98] 吴林，张仰森. 基于知识库的多层级中文文本查错推理模型[J]. 计算机工程，2012，38(20)：21-25.

[99] 张仰森，郑佳. 中文文本语义错误侦测方法研究[J]. 计算机学报，2017，40(4)：911-924.

[100] Wu S H, Liu C L, Lee L H. Chinese spelling check evaluation at SIGHAN bake-off[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 35-42.

[101] Yu L C, Lee L H, Tseng Y H, et al. Overview of SIGHAN bake-off for Chinese spelling check[C]// Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 126-132.

[102] Tseng Y H, Lee L H, Chang L P, et al. Introduction to SIGHAN bake-off for Chinese spelling check [C]//Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 32-37.

[103] Fung G, Debosschere M, Wang D, et al. NLPTEA shared task - Chinese spelling check[C]//Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications. Taipei, Tiwan: Asian Federation of Natural Language Processing, 2017: 29-34.

[104] Bradski G. The OpenCV library[J]. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 2000, 25(11): 120-123.

[105] Sanh V, Debut L, Chaumond J, et al. DistilBERT, A distilled version of BERT: Smaller, faster, cheaper and lighter[J/OL]. arXiv preprint arXiv: 1910.01108, 2019.

[106] 张仰森，唐安杰，张泽伟. 面向政治新闻领域的中文文本校对方法研究[J]. 中文信息学报，2014，28(6)：79-84. （下转第 27 页）

杨进才(1967—)，通信作者，博士，教授，博士生导师，主要研究领域为现代数据库与信息系统、中文信息处理、人工智能与自然语言处理。
E-mail：jcyang@mail.ccnu.edu.cn

胡泉(1980—)，博士，讲师，主要研究领域为中文信息处理，现代教育技术。
E-mail：123750955@qq.com

曹元(1996—)，硕士，主要研究领域为中文信息处理。
E-mail：cyuan@mails.ccnu.edu.cn

(上接第 18 页)

[107] Ge T, Wei F, Zhou M. Fluency boost learning and inference for neural grammatical error correction [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 1055-1065.

[108] 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. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 3291-3301.

李云汉(1993—)，硕士研究生，主要研究领域为文档信息处理。
E-mail：18618243871@163.com

李宁(1964—)，博士，教授，主要研究领域为文档信息处理。
E-mail：ningli.ok@163.com

施运梅(1969—)，通信作者，硕士，副教授，主要研究领域为文档信息处理。
E-mail：sym@bistu.edu.cn