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

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.