文章编号:1671-4598(2023)02-0210-07 DOI:10.16526/j.cnki.11—4762/tp.2023.02.033 中图分类号:TP317.1 文献标识码:A

基于深度学习的文本自动纠错系统设计与实现

杨 辉[1],张静静[2],熊 涛[2],蔡红维[3],刘皓挺[2],
才金山[1],杜晓平[1],高美萍[1]

(1.中国航天员科研训练中心,北京 100094;
2.北京科技大学 自动化学院 北京市工业波谱成像工程技术研究中心,北京 100083;
3.西昌卫星发射中心,四川 西昌 615000)

摘要:为解决办公人员在进行文档写作时存在各种文本格式和内容错误的问题,设计基于深度学习的文本自动纠错系统,用于辅助办公人员的写作和校对工作;分析办公人员的文本纠错需求,并进行文本格式与内容纠错方法研究;设计系统由写作模板生成、文本格式纠错和文本内容纠错三个功能组成;首先,设计文本要素识别与检查算法并基于VBA技术实现文本格式校对;然后基于Seq2Seq深度学习模型训练字词、语法和标点符号查错模型完成公文内容纠错,并根据办公人员工作需求建立纠错辅助字库提升系统纠错准确率;最终,通过系统测试实验结果表明,设计系统能够极大地提升办公人员写作效率并减轻文本校对工作负担。

关键词:Word文档;格式纠错;内容纠错;VBA技术;Seq2Seq模型

Design and Implementation of Official Document Automatic Error Correction System

YANG Hui[1], ZHANG Jingjing[2], XIONG Tao[2], CAI Hongwei[3], LIU Haoting[2],
CAI Jinshan[1], DU Xiaoping[1], GAO Meiping[1]

(1. China Astronaut Research and Training Center, Beijing 100094, China;
2. Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China;
3. Xichang Satellite Launch Center, Xichang 615000, China)

Abstract: In order to solve the problem of various text format and content errors in document writing by office staff, an automatic text error correction system based on deep learning was designed to assist office staff in writing and proofreading; The paper analyzed the text error correction requirements of office workers, and researched on text format and content error correction methods; The design system consisted of three functions of writing template generation, text format correction and text content correction; Firstly, the algorithm of text element recognition and check was designed, and the text format proofreading based on VBA technology was realized; Then, based on the Seq2Seq deep learning model, the error detection model of words, grammar and punctuation was trained to complete the error correction of official document content, and the error correction auxiliary word bank was established according to the work needs of office staff to promote the system error correction rate; Finally, the system test results showed that the design system can effectively improve the writing level of the office staff and greatly ease the burden of text proofreading.

Keywords: word document; format error correction; content error correction; VBA technology; Seq2Seq model

0 引言

电子文档是现代重要的信息交换媒介,对电子文档进行编辑排版、格式校对和内容检查是办公人员重要的日常工作。随着电子文档应用越来越广泛,且在特定的办公场合,固定版式文档的规范性和标准性标准越来越高[1]。文档格式排版繁琐且具有重复性,若办公人员对文档格式不熟悉,将会造成工作效率低下无法保证排版质量。当办公人员对语言掌握不足或工作疏忽大意时,电子文档中就不可避免地会存在字词、语法和标点符号错误,这些错误严重时会造成其他人对文本内容的理解偏差[2]。目前虽然存在文本内容纠错技术,然而这些技术的纠错率比较低,文

收稿日期:2022-11-05; **修回日期**:2022-11-16。
作者简介:杨 辉(1985-),男,北京人,硕士研究生,高级工程师,主要从事公文信息化系统、航天信号处理等方向的研究。
通讯作者:刘皓挺(1981-),男,甘肃兰州人,博士,教授,主要从事模式识别与智能系统、人—机—环境系统工程方向的研究。
引用格式:杨 辉,张静静,熊 涛,等.基于深度学习的文本自动纠错系统设计与实现[J].计算机测量与控制,2023,31(2):210-216.

本内容查错主要还是依赖办公人员细致的检查。为减轻办公人员进行文档编辑写作时繁重的文本校对工作，并提升文本内容纠错准确率，本文研究基于深度学习的文本自动纠错系统，用来辅助办公人员的文档写作和文本纠错工作，以提升办公人员的工作效率并确保文档内容的规范性和正确性。

目前已有相关研究和技术实现对文档格式校对和文本内容的纠错。文献［3］和［4］开展了对标准论文模板的分析与设定，提出了毕业论文格式自动检查系统。文献［5］实现了软件项目文档格式自动检查和修改，降低了文档编写人员在文档格式编写的出错率。文献［6］开发的Word文档格式自动排版系统，能够自定义文档的格式。随着自然语言处理的迅速发展，中文文本纠错技术也愈加成熟。文献［7］将中文文本纠错技术分为基于字词混淆集而形成候选字符串方法、基于概率统计分析上下文方法和基于规则和固定搭配关系的方法等。文献［8］结合二元语法模型和散串技术，在混淆集中选出最优纠错候选集，提出了一种基于窗口技术的校对方法。文献［9］对句子进行分词和词性标注解决字词错误，通过模式匹配解决搭配错误问题，利用成分分析解决成分错误问题。文献［10］采用规则与概率统计相结合的方法实现中文文本自动查错。文献［11］构建专业领域词语搭配知识库，并设计基于语法和词语搭配的双重中文文本校对算法。

随着信息化时代的迅速发展，现有文本纠错技术难以适应多样性的纠错任务，近几年许多研究者纷纷将深度学习方法引入文本纠错任务。深度学习算法自动编码器的自主学习特征可以通过对语言模型的不断训练实现纠错[12]。目前常用的基于深度学习纠错方法主要包括基于序列模型的纠错方法、基于注意力机制的纠错方法和基于预训练模型的纠错方法。文献［13］通过基于规则、统计和深度学习网络结合的方式提升中文文本纠错率。文献［14］将神经网络模型引入于中文文本纠错，并设计两个纠错子模块检查中文语法错误和拼写错误。文献［15］提出基于注意力机制的深度学习纠错方法，在非标注中文语料库上用降噪编码器训练纠错模型实现字级别和句子级别纠错任务。为实现文本的格式和内容纠错任务，本文提出设计一种基于深度学习的文本自动纠错系统。系统主要由文本格式纠错和文本内容纠错功能组成，能够检查文本格式和内容错误同时生成检查报告并通过一键校对实现文本自动纠错任务。

1　文本自动纠错系统总体方案设计

1.1　系统相关原理介绍

分析总结常见的文本格式错误主要包括以下几种：1）文档结构错误：表现为固定版式文档缺乏一些文本要素，如缺乏文档标题；2）正文段落格式不符合规范，表现为文本段落缩进和行间距不正确等；3）文本格式不符合规范：表现为文本的字体和字号不正确等；4）页码不符合规范：表现为文档页码的对齐方式或字体不正确等。系统的文本格式纠错功能主要基于VBA（visual basic for applications）技术实现的。VBA是微软用来拓展Office功能设计的开发语言[16]。Office中元素都以对象形式表示出来，而VBA具有特有的对象属性和方法，可用于表示Office对象并对其进行查询和调整[17]。对于固定版式的文档，文本要素组成及对应格式是固定的。因此系统格式校对功能实现主要在于识别出文档中的文本要素，根据格式要求检查并校对文本的格式。此外VBA宏代码保存在“所有文档Normal.dotm”中，可利用Python编程语言设计程序调取VBA宏代码对Word文档进行格式校对。

分析总结常见文本内容错误主要包括以下几种：1）字词错误：包括音似、形似和易错字词的错误；2）语法错误：分为搭配错误和成分相关错误；3）标点符号错误：表现标点符号冗余和半角全角标点符号误用；4）用语不符合规范：主要是固定地名、人名、单位名称和专业用语使用错误[18]。基于深度学习的序列模型是典型的自然语言处理模型，它采用自循环的计算方式，从序列起始端开始计算循环迭代一直计算到序列最后一个节点，以获取更多的特征信息，从而提高模型分类和预测精度。序列模型常用于机器翻译、语音识别、文本自动摘要和自动问答的任务处理[19]，将其引入文本纠错任务中是将错误句子作为源文本，正确句子作为目标文本，将源文本和目标文本一起作为训练数据来训练文本纠错模型[20]。本文采用序列到序列模型（Seq2Seq，sequence to sequence）来训练查错模型实现字词和语法查错，标点符号纠错采用代码逻辑去判断。Seq2Seq模型的思路表示从一个序列到另一个序列，输入和输出均为序列，它有自由度高、方便灵活的特点，是一种比较常见的处理文本序列数据的模型[21]。

1.2　系统总体设计方案

办公人员在进行固定版式文档写作时，会因为不清楚文档标准格式而造成写作效率低下且校对负担较大。即使文档标准格式模板是固定的，办公人员在写作时的复制粘贴、调整格式等操作也会使得文本格式不符合规范要求。此外文档中会存在的一些字词、语法和标点符号的错误，文本错误将严重影响到相关人员对内容的理解。因此本文对文本格式和内容纠错方法进行研究，以实现对文档中字体、段落格式和页码格式错误，以及文本中字词、语法、标点符号和规范用语错误进行全方位的错误识别与自动校对。根据办公人员文档写作和文本纠错需求，设计文本自动纠错系统由三个功能组成，即文档模板生成、文本格式校对和文本内容纠错功能。系统的总体设计方案如图1所示。

办公人员在起草文档时，利用文档模板生成功能选择生成标准格式Word模板，在该格式模板的基础上进行文档写作。当文档编辑完成时，利用文本格式校对功能对成稿

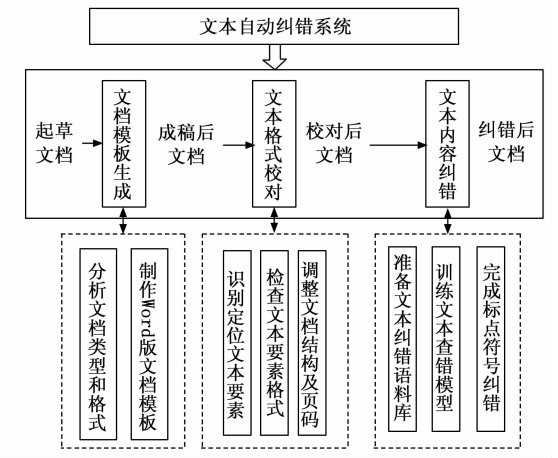


图 1 系统设计框图

文档进行格式校对，最后利用文本纠错功能对文本内容进行检查和纠错。本文设计的文本自动纠错系统能够极大地提升办公人员的工作效率，并在一定程度上保证文本的规范性和正确性。

2 文本自动纠错系统功能设计与实现

2.1 基于 VBA 技术的文本格式校对

本文格式校对功能是基于 VBA 技术对 Word 文档解析完成。使用 VBA 技术可以调用 Word 对象，获取 Word 文档的文本信息和格式信息，并对文档页面、文本段落和字体等格式进行调整。Word 对象结构层次如图 2 所示，Application 是 Word 应用中最大的对象，它共包含四种对象，它们分别为 Document 是文件类，Selection 是文字内容类，Bookmark 是书签类，还有 Range 是区域类。由于可通过不同方式访问同一个对象，所以类之间存在各样的重叠。除了这些属于顶层类型还有设置文本字体、段落等的格式的类。使用 VBA 技术可以获取 Word 对象，分析对象格式信息并对 Word 文档进行读写及格式调整。

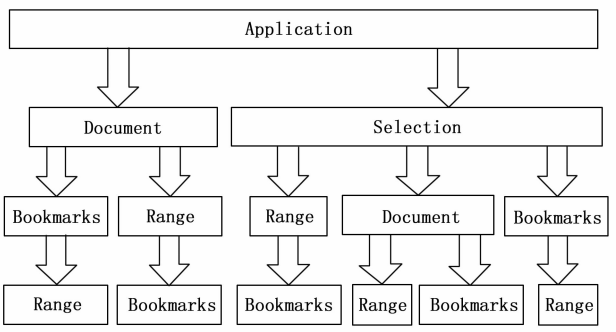


图 2 Word 对象结构层次

本文研究固定版式文档的格式校对功能，即文档中文本要素组成和要素格式是固定的。文本格式校对流程如图 3 所示。首先要对文档进行初始化设置，主要是删除文档中的超链接、调整页面大小和页边距以及进行字体初始化设置。接着要根据文本要素特点识别出文本要素所在区域并插入书签，然后检查文本要素的字体和段落格式是否满足标准格式要求，当不符合标准格式就修改文本要素的字体和段落格式。对于不同文档类型，其文档结构是不同的，主要表现为存在不同的文本要素，所以要根据文档类型对文档结构做出调整。最后由于不同类型文档页码格式不同且办公人员制作的文档页码常常不符合规范要求，因此删除原来的页码重新插入符合标准格式的页码。

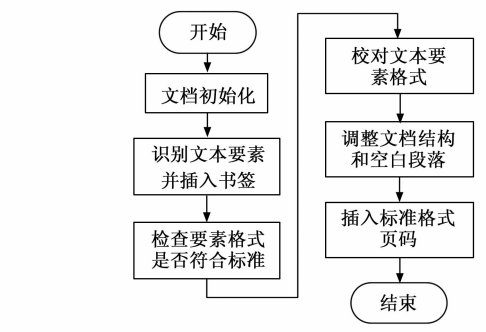


图 3 格式校对流程图

对于不同固定版式的文档如公司招投标文件、高校毕业论文、期刊杂志论文和机关公文等，文档的文本要素组成和格式标准相差很大。对这些文档格式进行校对，必须通过分析文本要素特点，明确一篇文档都由哪些必要文本要素组成，并根据文本要素特点设计相应的文本要素识别方法。文本要素特点包括文本要素内容特点，如论文标题一般含有“基于”、“方法”和“研究”等词语；文本要素格式特点，如论文标题字体字号和正文标题字体、字号不同；本文要素位置特点，如论文标题一般位于首页的第一行等等。结合文本要素的内容、格式和位置特点，设计方法识别出文档中各文本要素的区域并插入对应名称的书签。以标题为例，文本要素的定位流程如图 4 所示。即使文本内容有了新的插入或删除操作，书签定位的位置和内容也不会发生改变。后续文本要素的格式检查和校对都是对这些书签的内容分析与格式调整，极大地简化程序，确保格式校对的准确率。

识别出文本要素并在文本区域插入书签之后，就可以对文本要素进行格式检查。以文档标题为例各文本要素的检查流程包括如图 5 所示的部分。主要检查文本要素的字体、字号以及段落的间距和对齐方式等，当文本要素格式不符合要求，系统会以批注的形式展示在当前文档。

为便于对文本要素进行格式校对，系统利用 Word 文档样式设置功能，将文本要素的样式保存为格式模板文档 style.dotm 的内置样式。程序开始将文本要素的样式导入当前文档，当检查到文本要素格式不符合要求就将文本要素的格式设置为对应的样式，并对段落格式进行调整。对文档结构的调整主要是对文档中必要文本要素的调整，对文档的结构调整只需根据文档类型寻找必要文本要素的书签并进行增加或删减操作。一篇文档各文本要素之间会存在

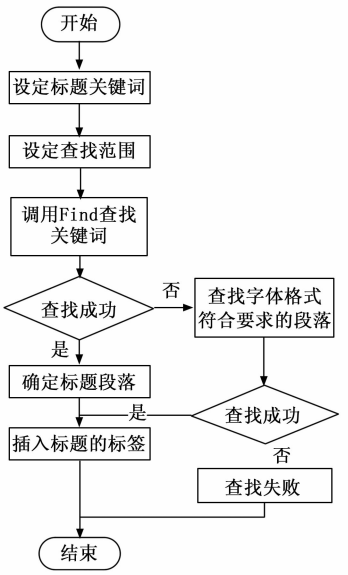


图 4　标题定位流程图

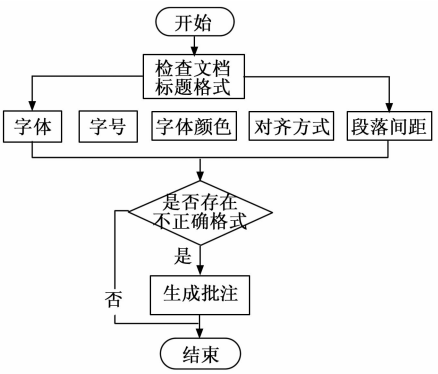


图 5　文本要素检查流程图

空行的要求，对这些空行的调整主要是定位两文本要素区域并进行删除和增添空行的操作。文档页码格式复杂且会存在奇偶页码格式不一致或首页码不同的情况，而办公人员在文档写作时一般不检查页码格式，因此系统必须对文档页码进行校对。首先删除原来文档的页码，接着根据文档要求设置是否奇偶页不同或首页不同，然后分别插入不同区域的页码并进行格式设置。通过这种方式插入页码，当文档新增或删减页时页码格式依旧符合规范要求。

2.2　基于 Seq2Seq 模型的文本内容纠错

为实现文本内容纠错功能，系统采用深度学习模型来训练中文文本纠错模型，并总结文档写作时的规范用语与固定搭配，创建辅助词库以提升纠错准确率。文本内容纠错流程如图 6 所示。首先对文本内容进行预处理，主要是提取文档正文内容部分并把正文的每一段分别存放于列表中。接着把预处理后的正文内容输入到预先训练好的算法模型。然后算法模型对正文内容进行检测，若检测结果与原文一致，则原文基本不存在内容错误，若不一致，则原文可能存在疑似错误。此时模型会输出正文中可能出现的字词错误、语法错误和标点错误的检测结果。然后系统查找可能发生错误的文字与标点，并对其以批注的形式展示在当前文档。最后一键纠错将会按照批注的纠错建议直接替换错误文本。

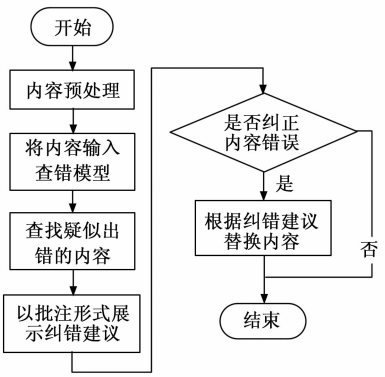


图 6　内容纠错流程图

本文采用 Seq2Seq 模型来训练字词和语法纠错模型。文本纠错任务可以看作不同序列的转化过程，原来的句子是源语句，正确的句子是目标语句，因此可以把 Seq2Seq 模型作为序列转化模型引入文本纠错。并且 Seq2Seq 深度学习模型相比于传统的规则和统计的模型能够更好地拟合错误语句到正确语句的转化过程。Seq2Seq 模型的底层结构是一个 Encoder-Decoder 的网络模型，其中 Encoder 是编码器，它会对输入的文本序列进行编码，使其变成长度一定的向量表达。Decoder 解码器对 Encoder 编码器获得的长度一定的向量表达进行解码，并将其转化为输出序列，解码和编码器模型一般用循环神经网络（recurrent neural network，RNN）模型，Encoder-Decoder 的设计决定了 Seq2Seq 模型的核心功能[21]。

从统计概率学方面看，Seq2Seq 模型是在给定输入文本 x 的条件下，找出使条件概率为最大值的目标文本 y，即使条件概率 $p(y_1,y_2,\cdots,y_{N'}\mid x_1,x_2,\cdots,x_N)$ 最大，其中 $x_1,x_2,\cdots,x_N$ 是给定的输入文本序列，$y_1,y_2,\cdots,y_{N'}$ 是该给定文本对应的输出序列，两个序列的长度 N 和 N' 不一定相等。我们先使用训练的数据集对算法模型进行拟合，使句子对的条件概率 $p(y_1,y_2,\cdots,y_{N'}\mid x_1,x_2,\cdots,x_N)$ 最大化。当训练的模型对参数拟合完成之后，给算法模型输入文本，训练好的 Seq2Seq 模型就会寻找使条件概率最大的文本，并将其作为模型的输出，计算公式如式（1）所示：

$$p(y_1,y_2,\cdots,y_{N'}\mid x_1,x_2,\cdots,x_N)=\prod_{n=1}^{N'}p(y_n\mid v,y_1,y_2,\cdots,y_{n-1}) \tag{1}$$

从算法模型结构角度看，当向 Seq2Seq 模型输入一个文本序列后，Seq2Seq 模型的 Decoder 会通过本身 RNN 模型对其进行编码，使文本序列成为长度固定向量，该向量包含文本语义。如图 7 所示，可以将 Encoder 编码器输出的隐状态直接定义为语义向量，也可以对其先进行变换处理，

然后把变换处理后的结果定义为语义向量。接下来 Seq2Seq 模型会把该语义向量输入到 Decoder 解码器，Decoder 解码器会以该语义向量作为依据并计算，并得出一个长度不固定的文本序列。在图 7 的常规 Seq2Seq 模型结构中，语义向量只作为一个 Decoder 解码器的输入数据，它并不参加 Decoder 解码器内部的后续计算。

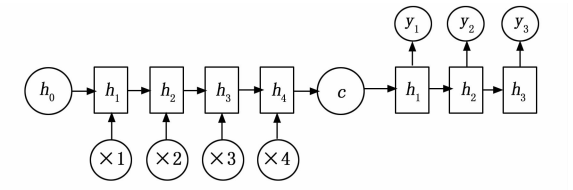


图 7 常规 Seq2Seq 模型结构

由于存在错误文本的句子在自动分词时会受到错误文本影响导致分词错误。本文引入注意力机制（Attention Mechanism）来解决分词时长序列到定长序列转化过程部分信息丢失的问题[22]。Attention 机制的本质是对给定目标通过生成权重系数对输入加权求和，来识别输入中哪些对于目标是重要的特征。将 Attention 机制引入 Seq2Seq 纠错模型，可以加强编码端和解码端的对应关系。Attention 机制把原数据看作键值对形式，根据给定任务目标的查询值计算键值和查询值的相似系数，即得到向量值得权重系数，之后利用权重系数对向量值加权求和得到注意力机制输出。在 Seq2Seq 纠错模型中加入 Attention 机制来学习句子之间的长距离依赖，计算公式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = soft\max(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \tag{2}$$

式中，$\boldsymbol{K}$ 代表词的键向量，$\boldsymbol{Q}$ 代表编码词的查询向量，$\boldsymbol{V}$ 代表值向量。通过计算权重得到注意力机制分布情况，从而得到对于当前输出位置相对重要的输入位置权重，在预测输出时相应地也会占较大的比重。即解码端自动选择与正在生成目标词相关源语句的词汇，进而提升了模型纠错的准确率。

基于 Seq2Seq 文本纠错模型的纠错准确率会受到训练数据集影响，因此要对训练语料预处理。先去除语料中非文字符号，接着对句子进行分词编码；然后分析数据获取训练数据的统计性信息，最后对句子重新排列以优化训练过程。本文基于 Seq2Seq 的纠错模型由编码器和解码器构成，在解码端加入注意力机制分散对输入语句各个词的关注度以掌握输入序列的细节信息，降低了错误文本对最终生成结果的影响。由于解码器前一时刻输出影响当前时刻输入，因此在模型训练时要清楚每一刻正确输入形式。那么对于所有训练样本，训练结果应使得全部训练样本预测概率之和最大，通过最大化似然函数获取最佳纠错模型。

本文内容纠错算法的流程如图 8 所示，首先，系统加载并初始化算法模型，然后使用正则匹配方法，对正文部分以中文标点符号的分句规则进行分句，待检测句子逐句输入文本纠错模型，模型读入该检测句子后会计算并返回一个算法认为正确的句子并与原句子进行对比，若两句相同，则待检测语句没有错误；若两句不同，则待检测语句可能存在疑似错误，算法将返回结果作为纠错建议。

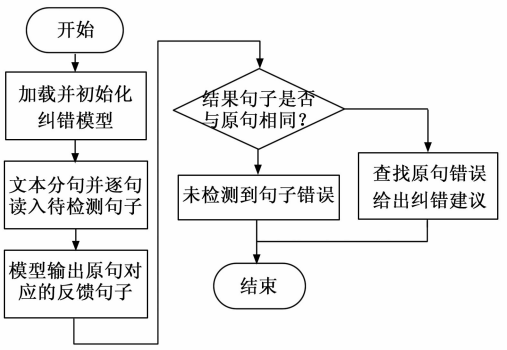


图 8 内容纠错算法流程图

对于标点符号的错误，由于办公人员疏忽而多打了一个标点符号，造成标点符号的冗余。本系统采用遍历循环的算法遍历全文，定位每一个标点符号的位置，判断该位置标点符号的数量，若出现两个以及以上的标点符号，判断标点符号冗余，返回标点符号错误的位置并以批注的形式显示出来。此外文档的标点大部分是全角符号，而办公人员在写作时可能会误用半角的标点符号。本系统在遍历全文标点的基础上，判断每个标点是否为半角，若判断结果为半角标点，则返回标点错误的位置并以批注的形式标注出来。

3 文本自动纠错系统测试结果分析

3.1 系统开发软硬件条件

本文采用 Python 编程语言，基于 VBA 技术和 Seq2Seq 深度学习模型，在硬件设备为 Windows10 的 64 位操作系统，处理器为 Intel Core i5-7200U，内存为 8G 的计算机上开发了一套文本自动纠错系统。系统界面如图 9 所示，主要由模板生成、格式校对和内容纠错三个功能组成。办公人员利用模板生成功能选择对应类型的文档模板，在模板基础上进行内容写作，内容编辑完成分别利用格式校对和内容纠错功能对文本进行格式和内容纠错，保证最终成稿文档的规范性和正确性。

3.2 系统测试结果分析

将文本自动纠错系统应用到某机关的公文纠错工作中。分析机关常用公文类型与格式要求，系统制作常见公文类型 Word 版的格式模板供办公人员起草公文使用。以请示公文为例，Word 版公文模板如图 10 所示。公文模板一般由版头、主体和版记三部分组成，其中文本要素的段落与字体格式是固定不变的，办公人员只需要在相应文本要素的位置进行内容编辑即可。

办公人员在使用格式校对功能时先选择需校对的文档及文档类型，再选择是否检查文档格式并对文档进行格式校对。首先点击“选择文件”按钮打开需要格式校对的文

图 9　文本自动纠错系统界面

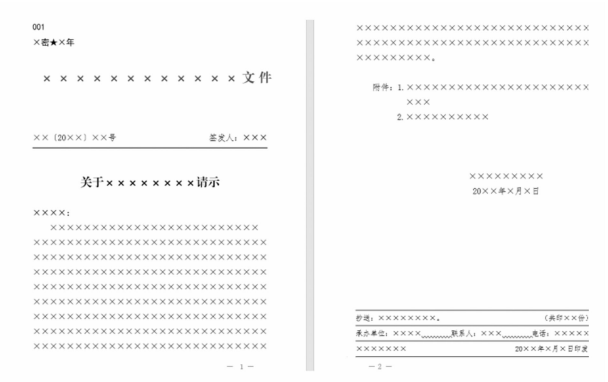
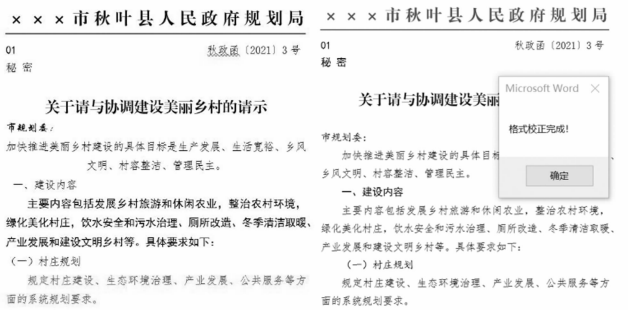


图 10　请示公文模板

档，接着点击“选择公文类型”下拉框选择需要校对的文档类型。若点击“开始检查”按钮，系统会打开文档检查文本格式错误并生成批注，同时如图 11 所示，会在“检测信息”文本框显示格式检查的结果供用户参考。若点击“保存结果”按钮，就会将“检测信息”框的内容保存为 .txt 文件。若点击“格式校对”按钮，系统就会对当前文档进行格式校对，并弹出“格式校正完成”提示信息框。该功能减轻了办公人员格式校对的工作负担，并且通过分析格式检查结果有利于了解文档常见的格式错误以加强办公人员的写作规范性。

图 11　格式检查的结果

以信函公文为例，图 12 展示了文档格式校对结果的部分截图。该截图展示了格式校对功能对文本要素的字体段落格式以及空白段落的校对。在字体格式校对方面：调整主送机关为楷体，一级标题为黑体等；在段落格式校对方面：调整主体内容第一段为两端对齐、首行缩进两字符的格式，调整一级和二级的层次标题段落缩进两字符等；在结构调整方面：在标题和主送机关之间插入一行空白段落。系统的格式校对功能可以实现对文档写作中常见格式错误的校对。



(a) 格式校对前　　(b) 格式校对后

图 12　格式校对结果图

文本内容检查功能无需选择文档类型。首先点击“选择文件”按钮打开需要内容校对的文档，若点击“开始检查”按钮，系统会打开公文文档检查公文内容错误并生成批注。以通知公文为例，图 13 展示了文本内容纠错的部分截图。

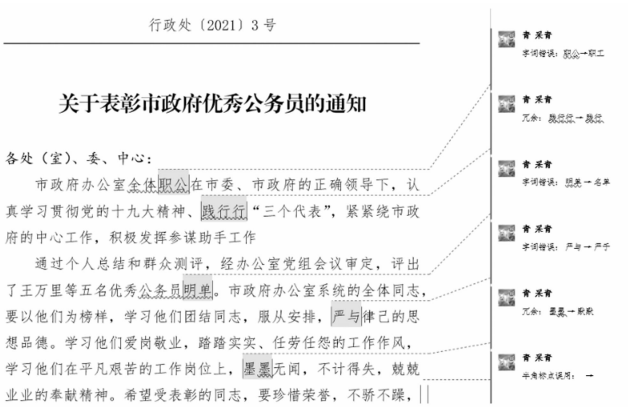


图 13　内容检查的结果

此时在“检测信息”文本框也会显示内容检查的结果供用户参考。若点击“保存结果”按钮，就会将“检测信息”框的内容保存为. txt 文件。若点击“一键纠正”按钮，系统就会根据内容查错结果对当前文档进行内容纠错。如图 14 所示，内容检查文档以文本文档保存，文档内包含了字词错误、冗余错误、标点错误的查错结果及纠错建议。其中“侦查”是基于文档常出现文本错误建立的自定义纠

错库而检查出来的文本错误。内容纠错功能保证了文档内容正确性，通过将内容检查的纠错建议加入到用户自定义纠错库中有助于提升文本纠错准确率。

图 14 内容检查文档的效果图

本文将基于深度学习的文本自动纠错系统应用到机关公文纠错工作中，通过对 30 篇公文文档的纠错测试发现，对文档常见格式纠错准确率在 60%以上，对文本内容纠错准确率在 20%以上。该系统能够满足办公人员基本的文本纠错需求，极大地提升办公效率并降低校对工作的负担。

4 结束语

随着信息化时代不断发展，电子文档应用越来越广泛，面对电子文档复杂的格式和内容错误，需要文本自动纠错系统作为辅助以减轻校对工作的负担。结合当前办公人员的纠错需求并参考当前文本纠错技术，本文提出基于深度学习的文本自动纠错系统，设计系统由文档模板生成、文本格式纠错和文本内容纠错三个功能组成，实现生成不同类型的标准格式文档模板，对文档进行格式校对与内容纠错同时生成检查文档供用户参考纠错建议。本文首先设计文本要素识别与检查算法并基于 VBA 技术实现文本格式校对。然后基于 Seq2Seq 深度学习模型训练字词、语法和标点符号查错模型完成公文内容纠错。然而系统的格式校对功能无法校对一些特殊的文本格式，如调整两行标题居中对齐及分别设置不同行的缩进格式。此外可以通过降低模型训练过程的过拟合或优化神经网络结构参数的方式进一步提升文本纠错准确率。最终，通过系统测试实验结果表明，本文的文本自动纠错系统的格式和内容纠错率都基本了满足办公人员的纠错需求，极大地提升了办公人员的写作效率和文本纠错的准确性，推动了电子办公时代的快速发展。

参考文献：

[1] 李 宁，田英爱. 办公文档与固定版式文档格式关系探讨 [J]. 电子学报，2008，36 (11)：128－132.
[2] 段良涛，郭曙超. 中文文本校对技术研究 [J]. 电脑知识与技术，2014，10 (19)：4602－4604.
[3] 王帅群，夏 斌，孔 薇. 基于. NET 的论文格式自动检查系统 [C] //全国第 21 届计算机技术与应用学术会议暨全国第 2 届安全关键技术与应用学术会议论文集，2010：335－339.
[4] 袁 敏. 学术论文格式检查和内容校对的研究 [D]. 北京：北京交通大学，2019.
[5] 侯伟波. 软件项目文档格式审查系统的设计与实现 [D]. 西安：西安电子科技大学，2016.
[6] 李 响. VC 中用 Word 实现文档的自动生成和排版的研究 [D]. 北京：华北电力大学，2015.
[7] 白雪丽，李建义，王洪俊. 中文文本自动校对方法研究综述 [J]. 软件导刊，2022，21 (8)：228－234.
[8] 汪维家，陈笑蓉，秦 进. 一种基于窗口技术的中文文本自动校对方法 [J]. 贵州大学学报，2003 (2)：161－164.
[9] 骆卫华，罗振声，宫小瑾. 中文文本自动校对技术的研究 [J]. 计算机研究与发展，2004 (1)：244－249.
[10] 徐全生，陈 莹. 基于二元、三元统计模型与规则相结合的中文文本自动查错研究 [J]. 科技信息 (学术研究)，2008 (36)：677－678.
[11] 陶永才，刘亚培，马建红. 一种结合压缩激发块和 CNN 的文本分类模型 [J]. 小型微型计算机系统，2020，41 (9)：1925－1929.
[12] 李 威. 基于深度学习的文本纠错关键技术研究 [D]. 武汉：华中科技大学，2021.
[13] 叶俊民，罗达雄，陈 曙. 基于层次化修正框架的文本纠错模型 [J]. 电子学报，2021，49 (2)：401－407.
[14] 邱肇泉. 基于序列到序列模型的中文语法纠错研究 [D]. 北京：北京交通大学，2021.
[15] 王匆匆，张仰森，黄改娟. 基于注意力机制与端到端的中文文本纠错方法 [J]. 计算机应用与软件，2022，39 (6)：141－147.
[16] 张槐权. 基于 VBA 技术对 Word 文档的公文格式审核 [J]. 电脑知识与技术，2017，13 (25)：209－212.
[17] 李小遐. Office 自动化技术在办公中的应用 [J]. 无线互联科技，2015 (2)：94－96.
[18] 张 蕾. 中文文本的词语纠错方法研究 [D]. 南昌：江西财经大学，2020.
[19] 董 谱. 改进的 Seq2Seq 文本摘要生成方法 [D] 广州：广东工业大学，2021.
[20] 袁 阳. 基于深度学习的中文文本纠错方法研究 [D]. 北京：北方工业大学，2021.
[21] 龚永罡，吴 萌，廉小亲. 基于 Seq2Seq 与 Bi-LSTM 的中文文本自动校对模型 [J]. 电子技术应用，2020，46 (3)：42－46.
[22] 李丹丹. 基于 Transformer 的中文纠错系统设计与实现 [J]. 数字技术与应用，2021，39 (12)：213－215.