引文格式：刘满义,牛坤. 基于文本纠错模型的时空数据标准化文档生成研究[J]. 地理空间信息, 2025, 23(11): 7-11. doi:10. 3969/j.issn. 1672-4623.2025.11.002

基于文本纠错模型的时空数据标准化文档生成研究

刘满义[1]，牛　坤[2*]
（1. 贵州省地质矿产勘查开发局测绘院，贵州 贵阳 550025；2. 贵州大学 计算机科学与技术学院，贵州 贵阳 550025）

摘　要：为提高时空数据文档成果的生成效率和质量，降低工作成本，针对特定的调查工作，提出了Textual-Feature BERT文本纠错模型；设计了时空数据标准化文档生成算法，通过自定义的数据抽取规则对属性数据进行智能化抽取，对空间数据成果图件进行自动编制，实现快速、便捷生成文档成果的目标。该算法已应用于地理国情监测、自然灾害综合风险普查、“两违”遥感监测等项目中，错别字纠错正确率达95%以上、文本缺失填充正确率可达93%。结果表明，该算法具有一定的应用价值，且可应用于类似项目中。

关键词：文本纠错模型；GIS动态匹配制图；时空数据；标准化工具
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2025）11-0007-05

Research on Spatio-temporal Data Standardization Document Generation Based on Text Error Correction Model

LIU Manyi[1], NIU Kun[2]
(1. Institute of Surveying and Mapping, Guizhou Bureau of Geology and Mineral Resources Exploration and Development, Guiyang 550025, China;
2. School of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: In order to improve the generation efficiency and quality of spatio-temporal data document achievement, reduce the work's cost, we proposed a Textual-Feature BERT text error correction model for a specific investigation work. We designed a spatio-temporal data standardized document generation algorithm, which intelligently extracted the attribute data through custom-defined data extraction rules, and automatically compiled the spatial data achievement map, achieving the goal of generating document achievements quickly and conveniently. The algorithm proposed in this paper has been applied in geographical conditions monitoring, comprehensive risk census of natural disasters, and remote sensing monitoring of illegal construction, with the error correction accuracy of incorrect characters reaching 95% or above and the accuracy of text filling reaching 93%, which show that this algorithm has certain application value and can be used in more similar projects.

Key words: text error correction model, GIS dynamic matching mapping, spatio-temporal data, standardization tool

随着社会经济的快速发展，我国开展了全国土地调查、全国土壤普查、自然灾害综合风险普查、地理国情监测、自然资源确权登记等一系列与空间数据相关的普查或调查，而空间数据成果文档是空间数据相关普查中不可或缺的成果资料之一[1-2]。通常空间数据成果文档的数量巨大，且包含的内容繁杂，涵盖了空间数据的属性信息、遥感影像数据统筹、遥感监测等图件成果和统计信息。通过人工方式完成空间数据成果文档的制作往往需要花费大量的人力、时间和经济成本；且在制作过程中难免会出现属性信息缺失、统计信息不全、图件成果错位等问题，导致生产的空间数据成果文档质量不高，难以满足实际生产项目的要求[3-4]。人工采集或填报时常出现音错、形错等文本错误，这些数据处于整项工作的上游，错误会蔓延从而严重影响自然资源普查整体质量，因此时空数据文本智能纠错是标准化文档生成迫切需要解决的问题。自2018年Google提出基于Transformer架构的预训练模型BERT[5]以来，陆续诞生了许多在错别字纠正方面表现优异的改进模型，如Soft-Masked BERT[6]、FASPell[7]、PLOME[8-9]等。本文根据不同调查工作领域文本内容的特性，提出了一种基于文本特性的Textual-Feature BERT中文文本纠错模型，在实际项目中取得了较好的应用效果。

结合自然资源普查文档标准化特定应用项目需求，本文对成果文档生成过程中的关键技术进行了研

项目来源：贵州大学引进人才科研资助项目（贵大人基合字2022-21）；贵州省科技计划资助项目（黔科合支撑[2023]一般169）；贵州省地质矿产勘查开发局地质科研资助项目（黔地矿科合[2024]12号）。
第一作者简介：刘满义（1997—），硕士，工程师，从事GIS时空数据建模、地理信息系统等工作，E-mail：1558761438@qq.com。
通信作者：牛坤（1985—），博士，讲师，研究方向为群智感知数据挖掘、人工智能与安全，E-mail：275553858@qq.com。

究。首先设计了基于文本特性的Textual-Feature BERT中文文本纠错模型，并在特定专用训练集进行模型训练和优化；然后设计了空间数据与成果文档之间的数据匹配规则；最后结合GIS、GDAL、Python等技术设计研发了空间数据成果文档标准化生成工具的架构。本文采用自顶向下的方法对系统进行整体设计，包括文本纠错智能算法模块、数据匹配规则设计模块、基于GIS的动态制图模块、空间数据成果预览模块和成果文档标准化生成模块等核心功能模块。

1 Textual-Feature BERT中文文本纠错模型

基于文本特性的Textual-Feature BERT中文文本纠错模型（图1）包括4个部分：输入层通过分层、序列化等操作将数据集转换为BERT模型可识别的输入数据[10-11]；模型层根据设定的优化器、学习率、分块大小、迭代次数、收敛条件等参数进行训练，经过多轮参数调整，保存较理想的模型；微调层根据保存的模型以及应用场景下的特定数据集，通过分块大小、学习率、验证集正确率等参数对模型进行微调[12]，并保存较理想的优化模型，同时可通过融合多个理想模型来提升模型的泛化能力；应用层将优化模型部署到应用场景的服务器上，通过RESTful API规范提供对外的接口，使用者可通过接口完成单个词语、单个句子或文档的文本纠错任务。

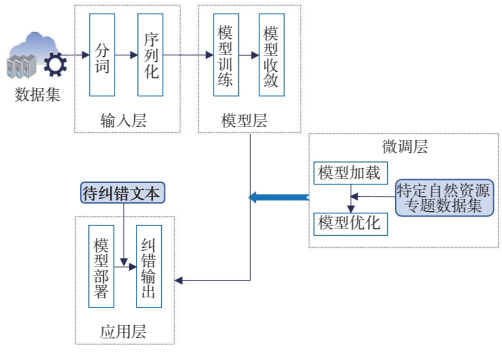


图1 Textual-Feature BERT模型框架

1.1 专题数据集构建

为了训练具有文本特性的Textual-Feature BERT中文文本纠错模型，根据实际应用场景需求，从地理国情监测、土壤普查、自然灾害综合风险普查、自然资源确权登记、遥感影像数据统筹、遥感监测等项目中收集错误文本实例和正确文本示例构成自定义数据集，共计5 000个样本。模型层数据的训练集来源于自定义数据集训练集、SIGHAN14和SIGHAN15训练集、CGED2018和CGED2020训练集、CGED2021训练集；测试集来源于自定义数据测试集、SIGHAN15测试集、CGED2016和CGED2017测试集部分数据。微调层中的特定数据集来源于自定义数据集中地理国情监测、自然灾害综合风险普查等特定场景的样本训练集和测试集[13-14]。验证集分为通用验证集和专用验证集，通用验证集随机从SIGHAN数据集抽取300个样本、CGED数据集抽取1 000个样本、自定义数据集抽取200个样本，共计1 500个样本。从自定义数据集抽取地理国情监测验证集记为专用验证集1，共计300个样本；从自定义数据集抽取自然灾害综合风险普查验证集记为专用验证集2，共计500个样本。

地理国情监测和自然灾害综合风险普查数据集均为根据实际项目情况制定的数据集。本文以自然灾害综合风险普查数据集为例，对自定义数据集内容进行介绍。该数据集主要包括行政区划名称样本和自然灾害综合风险普查常用名词样本两个部分（表1）。采用这两个部分样本的原因在于：①审核调查阶段数据时，经常出现县、乡镇、村（居委会）名称存在错别字的情况；②审核评估与区划阶段提交成果报告时，经常出现与评估、区划规范中专用名词不一致的错别字。

表1 自然灾害综合风险普查数据集样本示例

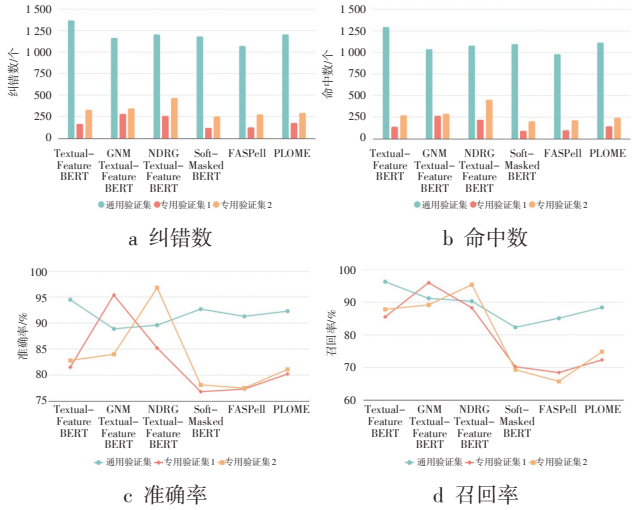
样本类别	正确样本值	错别字样本值
行政区划名称类	水口寺街道	水口是街道
	宗岭镇	鬃岭镇
	乌尧村	乌要村
自然灾害综合风险普查常用名词	承灾体	承载体
	风险评估	奉献评估
	防治图	防止图
	综合减灾能力	总和减灾能力

1.2 模型训练与实验分析

本文利用深度学习框架PyTorch对Textual-Feature BERT模型进行训练[15-16]。设置10组优化器、学习率、分块大小、最大迭代次数等实验参数，对训练时间、召回率、测试集正确率等进行评估和参数优化，最终选定模型的训练参数为优化器AdamW、学习率0.000 1、分块大小32、最大迭代次数10、隐藏层维度768、Transformer层数16、预热学习率调整策略ExponentialLR。

在选定优化模型参数的基础上，利用专题数据集再次进行参数微调生成地理国情监测（GNM Textual-Feature BERT）、自然灾害综合风险普查（NDRG Textual-Feature BERT）两个特定场景的专用模型。其中，对GNM Textual-Feature BERT模型的学习率、最大迭代次数进行微调，能提升模型的召回率和测试集正确率；对NDRG Textual-Feature BERT模型的分块大小、最大迭代次数进行微调，能提升模型的召回率、测

试集正确率。利用通用验证集、专用验证集1、专用验证集2对Textual-Feature BERT、GNM Textual-Feature BERT、NDRG Textual-Feature BERT模型进行验证，并与Soft-Masked BERT、FAS Pell和PLOME的结果进行对比分析。验证集上的纠错数、命中数、准确率、召回率结果见图2，结果表明：①Textual-Feature BERT模型在训练时加入自定义数据集，融合了特定数据的文本特性，因此在通用验证集上表现出较好的性能，说明该模型对特定文本特性具有较好的应用价值，也说明微调后Textual-Feature BERT模型的泛化能力降低；②专用模型在其对应的专用验证集上的表现优于其他模型，正确率均能达到95%，说明Textual-Feature BERT模型经过特定数据集微调后，在特定场景下具有较高的文本纠错准确率，可在实际项目中进行推广使用；③Textual-Feature BERT在特定场景下的召回率优于其他模型，均在95%以上；④NDRG Textual-Feature BERT模型在专用验证集1以及GNM Textual-Feature BERT在专用验证集2的表现较好，给人一种专用模型在专用验证集上的表现优于通用模型的错觉，通过分析两个对应的数据集发现，其原因是两个数据集中存在许多相似的空间数据描述，若采用两个几乎无相似之处的专用验证集，则不同专用模型的表现通常不及Textual-Feature BERT模型。综上所述，Textual-Feature BERT及其微调后的模型在特定应用场景下具有较好的文本纠错能力，微调后的模型通过牺牲其泛化能力进一步提高在特定应用场景下的文本纠错准确率，在后续还需在平衡纠错准确率与泛化能力方面进行深入研究。



a 纠错数　　b 命中数

c 准确率　　d 召回率

图2　不同模型在不同验证集的结果对比

2　时空数据文档生成关键算法

2.1　数据抽取规则

时空数据文档内容是从大量的基础数据中抽取而来，不同的时空数据文档成果对应不同的基础数据，本文采用配置数据抽取规则的方式搭建了基础数据与成果文档之间的桥梁。为使数据抽取规则可视化、便于编辑、通用性强[17]，本文采用JSON格式文本配置数据抽取规则，包括文本信息规则、图片信息规则和统计信息规则。所有规则信息的数据类型均为数组。

1）文本信息规则包括包括数据来源（DSource）、文档位置（DLocation）、文档位置值（DLValue）、文档值单位（DLVUnit）和字体样式（FSSetting）等元素。DSource项数据类型为数组，长度为2，第一个元素为字符串，可取值常量和动态；第二个元素为字段名称，若第一个元素为常量，则该值为固定字符串，若第一个元素为动态，则该值为基础数据字段名称，通过“,”连接可配置多个字段名称。DLocation项数据类型为字符串，可取值段落和表格。DLValue项数据类型为字符串，值根据DLocation项确定，若DLocation项为段落，则该值为段落顺序值，从0开始计数；若DLocation项为表格，则该值为单元格行列值，从0开始计数。DLVUnit项数据类型为字符串。FSSetting项数据类型为数组，长度为5，依次为字体名称、字体大小、字体居中设置（1为居中、0为默认）、字体是否加粗（1为加粗、0为不加粗）、字体颜色（取值为RGB值数组）。

2）图片信息规则包括图片来源（ISource）、图片元数据（IMetadata）、DLocation、DLValue和图片大小（ISize）等元素。ISource项数据类型为字符串，可取值静态和动态。IMetadata项数据类型为数组，长度为7，依次为文件路径、字段名称、图片名称前缀、图片名称后缀、图片格式名称（取值为PNG、JPEG、TIFF）、动态图片大小（仅当图片来源为动态时有效）、制图模板路径（仅当图片来源为动态时有效）。DLocation、DLValue项与文本信息规则配置一致。ISize项数据类型为数组，长度为2，依次为图片宽度、高度。

3）统计信息规则包括DSource、DLocation、DLValue、DLVUnit、FSSetting和数据统计类型（DSStatistical Type）等元素。DSource项数据类型为字符串，取值为基础数据字段名称，且字段数据类型必须为数值。DLocation、DLValue、DLVUnit、FSSetting与文本信息规则配置一致。DSStatistical Type项数据类型为字符串，可取值SUM、MEAN、MIN、MAX、STD、COUNT。

根据上述规则，将同一图斑两个时相的属性信息和影像图片信息写入时空数据成果的Word文档中，写入的内容需求是一个可变标题和一个动态内容表

格。对表格中的14项文本信息和两项图片信息配置数据匹配模板文件，其中行政区名称需配置乡镇名称和村名称两个字段。文本信息配置示例为：

{"DSuorce":["动态","TBZLZ,TBZLC"],"DLocation":"表格","DLValue":"0,0,1","DLVUnit":"","FSSetting":["宋体",12,1,1,[255,0,0]]}

图片信息配置示例为：

{"ISuorce":"静态","IMetadata":["E:\\img\\前","TBBH","","_1","PNG",[0,0]],"DLocation":"表格","DLValue":"0,8,0","ISize":"6.95,6.95"}

2.2 基于GIS的空间数据解析与动态制图

基于GIS和GDAL提供的Python语言包，根据空间数据的格式调用不同的空间数据解析模块对空间数据进行解析，并采用键值存储结构存储解析后的数据，以属性字段名称为键，属性字段值为值，geom为空间位置信息的键，采用OGC标准中的WKT格式表示空间位置信息。支持的数据类型包括ESRI Shapefile、ESRI FileGDB、ESRI Personal Geodatabase、GeoPackage、Geographic Javascript Object Notation、SQLite、Comma-Separated Values。若输入的空间数据类型为Comma-Separated Values，则表示空间位置信息的键值必须为geom且空间位置必须是标准的WKT格式[18]。

在空间数据解析的基础上，根据当前空间要素的位置进行动态制图。将输入的空间数据要素动态添加至GIS制图模板中，根据生产项目实际需求设置空间数据样式；并将制图模板空间显示范围设置为当前空间要素外接矩形范围，调用Python中相应的制图接口，将当前制图模板视图按照配置模板的要求导出为图片，支持ArcMap（.mxd）、ArcGIS Pro（.aprx）、QGIS（.qgz）3种格式的GIS制图模板。基于ArcGIS提供的Arcpy包完成空间数据解析和动态制图过程。

3 文档标准化生成方案设计与实现

文档生成功能的输入包括空间数据、成果文档模板、数据匹配模板、文档输出路径、文档命名规范；输出为文档成果。文档生成功能的总体流程包括：①数据检查，根据数据匹配模板检查所有配置的数据字段是否均已在输入的空间数据中，若不是则输出缺少数据字段，根据数据匹配模板检查所有配置的文档位置值是否存在，若不存在则输出该文档位置值；②数据提取，根据数据匹配模板配置的文本信息、图片信息、统计信息的数据来源，从输入的空间数据中提取信息，若图片信息规则中存在动态图片，则根据配置信息中的制图模板生成满足要求的图片文件；③文档生成，根据数据匹配模板中空间数据值与文档对应的位置关系，将提取空间数据值写入文档对应位置，按照字体样式或图片大小要求对文档相应内容进行调整，将文档按照规范命名后输出至指定的文件路径下。

3.1 文档数据解析

本文利用Python中相关的文件解析包对JSON格式的数据匹配模板、Word格式的成果文档模板进行解析。解析完成后，数据匹配模板采用键值存储结构进行存储，文本信息规则、图片信息规则、统计信息规则采用3个键值对存储变量进行分开存储；成果文档模板采用Document数据结构进行存储，Document包括当前文档中所有的文本段落、表格、图片信息。

3.2 文档数据写入

基于Python中Word文件相应的操作包，根据数据匹配模板信息在成果文档模板的对应位置写入相应信息，并依据不同的信息类型对写入信息进行样式定制。文档数据写入的关键在于将文本信息、图片信息、统计信息与文档模板中的位置一一对应。若写入信息为文本，则根据文档位置和文档位置值获取Document的Paragraphs或对应Cell的Paragraphs数组下的Paragraph元素，通过Paragraph的text属性写入文本信息，通过Paragraph的bold、font、element、alignment、color等属性设置字体样式。若写入信息为图片，则根据文档位置和文档位置值获取Document的Paragraphs或对应Cell的Paragraphs数组下的Paragraph元素，通过Paragraph的add_picture方法写入图片信息，通过add_picture方法返回值的width、height属性设置图片大小，如果图片在表格中，则默认设置图片样式为居中。若写入信息为统计数据，其写入方式与文本一致，不同点在于统计数据需先根据配置的字段信息、数据统计类型进行统计分析。

4 应用实例

地理国情监测项目中首先采用Python语言的GDAL库对空间数据进行解析，读取三调土地类型、地理国情土地类型、变化原因等重要属性值；然后加载GNM Textual- Feature BERT模型对属性值进行文本纠错，将正确属性值重新写回空间数据并对数据进行标记；最后采用人工抽检的方式对纠错属性值进行核查，并将核查结果加入地理国情监测数据集中，不断地对GNM Textual- Feature BERT模型进行改进。统计纠错文本发现，纠错正确率可达95%，错误类型主要为错别字。

自然灾害综合风险普查项目中首先采用Python语言的GDAL库对调查阶段的空间数据成果进行解析，读取行政区划相关字段值；然后加载NDRG Textual-Feature BERT模型对字段值进行文本纠错，将正确字段值重新写回空间数据相应字段并对数据进行标记；最后采用人工抽检的方式对纠错字段值进行核查，并将核查结果加入自然灾害综合风险普查数据集中。采用Python语言的docx库对评估与区划阶段的成果报告Word文档进行分段读取；然后采用段中分句输入NDRG Textual-Feature BERT模型的方式进行文本纠错，并将纠错前后的文本一并写入报告成果中；最后通过人工确定是否进行替换，并将错误的文本纠错内容加入自然灾害综合风险普查数据集中，不断地对NDRG Textual-Feature BERT模型进行改进。统计纠错文本发现，调查成果数据的纠错正确率可达96%，错误类型主要为错别字；成果报告的纠错正确率可达93%，错误类型主要为错别字和文本缺失。

利用开发的空间数据成果文档标准化生成工具将某“两违”遥感监测生产项目中涉及的成果数据（ESRI Shapefile文件）、成果文档模板（docx文件）、数据匹配模板文件（JSON文件）、成果文档输出文件夹生成成果文档，输出成果示例见图3，行政区名称、图斑编号设置为红色，行政区名称在配置文件中设置为乡镇名称和村名称两个字段，再动态写入两个字段的组合值。

XXX"两违"遥感监测图斑之记

行政区名称	XX镇XX村	图斑编号	NKZ2024020049
前时相	20231225	后时相	20240212
中心点经度	118.47196302	中心点纬度	24.80353315
前时相地类	建筑物	后时相地类	建筑物
图斑变化类别	建筑加建	图斑面积（㎡）	207.42
图斑占永久基本农田面积（㎡）	0.0	图斑占耕地面积（㎡）	6.04
历史图斑			
说明	1. 前后时相对比，红色线框区域为地类变化区域。 2. 变化区域面积采用软件计算。		
			
前时相卫片		后时相卫片	
填表人：XXX　　检查人：XXX　　填表时间：2023年4月1日			
统计值	66121.47		

图3　成果文档输出示例图

5　结　语

本文利用GIS、GDAL、Python等技术研究了空间数据成果文档生成过程的关键技术，并构建了基于数据匹配模板的动态空间数据成果文档标准化生产工具。该工具已在多个实际生产项目中得到应用，结果表明其可代替人工进行包含繁杂信息的成果文档生成，提升了工作效率和成果文档质量。工具运行所需的参数可按实际生产项目需求进行自定义配置，可应用于类似的空间数据成果文档生成项目中，具有较强的可复制性、可重用性、可维护性。

从GNM Textual-Feature BERT和NDRG Textual-Feature BERT模型在实际项目应用中的表现来看，在短文本纠错方面的表现较好，对于长句子的纠错能力较弱，特别是在文本缺失的纠错能力方面表现最差。其原因在于，地理国情监测数据集和自然灾害综合风险普查数据集样本的代表性有待提升，模型微调过程和最优模型选择条件有待改进，后续将针对这两个方面进行更加深入的研究，不断提升模型能力。

参考文献

[1] 张寿雨. 基于ArcEngine的三调和变更调查文档批量制作工具[J]. 测绘与空间地理信息,2023,46(6):219-221

[2] 方圆,宋阳,任立欣. 地理国情监测数据处理工具的设计与实现[J]. 测绘与空间地理信息,2023,46(增刊1):98-100

[3] 王冠,肖昶,代琦. 基于时空数据的武汉市绿色交通出行模型研究[J]. 地理空间信息,2023,21(1):20-25

[4] 王震,王珍,丁乐乐. 自然资源违法实时在线监测平台设计与实现[J]. 地理空间信息,2023,21(6):35-38

[5] Devlin J,Chang M W,Lee K,et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4 171-4 186

[6] Zhang S,Huang H,Liu J,et al. Spelling Error Correction with Soft-masked BERT[C]//Proceedings of the Association for Computational Linguistics,2020:882-890

[7] Hong Y,Yu X,He N,et al. FASPell: A Fast,Adaptable,Simple,Powerful Chinese Spell Checker Based on DAE-decoder Paradigm[C]. Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2019

[8] Liu S L,Yang T,Yue T C,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,2021:2 991-3 000

[9] Guo J X,Sainath T N,Weiss R J. A Spelling Correction Model for End-to-end Speech Recognition[C]// IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2019:5 651-5 655

（下转第69页）