DOI: 10.19783/j.cnki.pspc.240838

基于 Sentence-MacBERT 模型的同源录波数据匹配方法

戴志辉[1]，张富泽[1]，韩 笑[1]，王冠南[2]

(1. 河北省分布式储能与微网重点实验室(华北电力大学)，河北 保定 071003;
2. 国网江西省电力有限公司电力科学研究院，江西 南昌 330096)

摘要：由于不同时期的录波数据记录标准有所不同，以及各个生产厂家对标准的解读存在偏差，造成同源录波数据的通道名称存在个性化差异，且通道索引号不同，难以进行录波数据的同源匹配。针对上述问题，提出基于句向量掩码纠错双向编码器表征语言模型(sentence-masked language model as correction bidirectional encoder representations from transformers, Sentence-MacBERT)的同源录波数据匹配方法。首先，分析录波文件的记录格式特点，根据录波文件的格式特点完成核查信息表的构建。然后，通过构建的核查信息表进行录波文件自动校核。最后，在双向编码器表征(bidirectional encoder representations from transformers, BERT)模型的基础上构建 Sentence-MacBERT 同源通道匹配模型，完成同源录波数据匹配。算例分析表明，根据核查信息表能够完成录波文件的自动校核，并对解析失败的录波文件发出告警信息。利用 Sentence-MacBERT 模型进行通道名称匹配的效果良好，能够有效地完成录波数据的同源匹配，帮助运行人员进行故障分析。

关键词：录波数据；Sentence-MacBERT；自动校核；通道名称；同源匹配

Homologous recording data matching method based on the Sentence-MacBERT model

DAI Zhihui[1], ZHANG Fuze[1], HAN Xiao[1], WANG Guannan[2]

(1. Hebei Key Laboratory of Distributed Energy Storage and Microgrid (North China Electric Power University), Baoding 071003, China; 2. Electric Power Research Institute of State Grid Jiangxi Electric Power Co., Ltd., Nanchang 330096, China)

Abstract: Because of differences in recording standard over different periods and variations in manufacturers' interpretation of these standards, homologous recording data often exhibit personalized differences in channel names and channel index numbers, making it difficult to achieve accurate matching of homologous recording data. To solve this problem, a method for matching homologous recording data based on the Sentence-MacBERT model is proposed. First, the characteristics of the recording format are analyzed, and a verification information table is constructed based on these format characteristics. Then, the verification information table is used to automatically verify the recording files. Finally, a Sentence-MacBERT homologous channel matching model is constructed based on the BERT model, and the homologous recording data matching is completed. Case studies show that the verification information table can be used to automatically verify the recording files, and alerts are generated for the recording files that fail to parse. The Sentence-MacBERT model is excellent in channel name matching, effectively completing the homologous matching of recording data and helping operators in analyzing faults.

This work is supported by the National Natural Science Foundation of China (No. 51877084).

Key words: recording data; Sentence-MacBERT; automatic checking; channel name; homologous matching

0 引言

为保证智能变电站的安全稳定运行，继电保护系统往往遵循双重化配置原则[1]，并且规定按电压等级和网络配置故障录波装置，要求能够完成继电保护开关量和电气量的采集和记录[2-3]。这种冗余配置会在电网发生故障或扰动时产生大量的同源录波数据[4-5]，现场运行人员在进行故障诊断时通常需要根据两个或多个录波文件中的同源录波数据进行综合分析判断。然而，不同生产厂家制造的保护装置与

基金项目：国家自然科学基金项目资助(51877084)

故障录波装置的型号种类多样[6]。不同时期的录波数据记录标准也有所不同，导致录波文件的通道名称和索引号存在个性化差异，难以进行同源录波数据的准确匹配。因此，亟须研究同源录波数据匹配方法，增强对海量录波数据的整合利用，帮助现场运行人员进行事故分析。

目前，录波文件的记录格式主要采用 IEEE 制定的电力系统暂态数据交换通用格式(common format for transient data exchange for power systems, COMTRADE)标准[7]，用于规范电力数字记录设备在进行故障录波时的存储格式，便于第三方解析软件的分析和处理。COMTRADE 的标准版本共有 1991 版、1999 版和 2013 版。其中，1999 版在 1991 版的基础上增加了信息文件，并在配置文件中给出互感器变比、字段格式等扩展信息。2013 版又在 1999 版的基础上进一步修订和完善配置文件格式。另外，电力系统的建设年代长久，存在大量规格型号不一，性能不同的保护和录波装置分布在各变电站[8-10]。以录波装置为例，当前国内主要有 9 个制造厂商、25 个主流型号，且均是根据厂家自身对标准的解读去定义录波文件格式，存在多种版本和不同程度的差异[11]，导致在录波文件的解读过程中不可避免地会出现文件乱码、数据缺失等问题，加大了同源录波数据匹配任务的技术难度。

由于标准版本的不同以及各个厂家对标准的解读存在偏差，不同设备对录波文件相同通道的命名和通道排列顺序会有所不同，而提取录波数据需要预先知道其所在的通道名称与通道索引号，所以无法直接进行同源录波数据的匹配任务。当前工程上主要采用人工方式进行同源通道的匹配任务，但该方法工作量大，所需时间长，且易发生匹配错误。对此，已有部分专家学者做出了相应研究，文献[12]提出利用 Word2vec 模型实现通道名称的中文分词及中文词向量构建，并利用余弦相似度和逆文本频率方法实现了通道名称的识别匹配。文献[13]基于正则表达式规范化录波通道的命名形式，并利用 Jaccard 相似度系数实现智能变电站录波通道同源匹配。但是上述文献仅停留在配置文件中的通道名称识别层面，没有考虑到录波文件乱码、数据缺失等问题给同源匹配带来的影响，未能真正实现录波数据的整合利用，且均需要对通道名称进行分词、去无用信息等预处理，前期需要对各种格式的通道名称进行预分析，增加了通道名称识别的工作量和复杂度。文献[14]通过词向量技术计算语义相似度实现虚端子的链接匹配，但采用分词表示的方法难以考虑文本的全局语义信息。而随着深度学习技术的快速发展，词嵌入语言模型(embeddings from language modeling, ELMo)[15]、生成式预训练转换器(generative pre-trained transformer, GPT)[16]、BERT[17]等动态语义特征学习模型受到广泛应用，通过在预训练模型上实施迁移学习，可以获取更丰富的语义特征信息。文献[18]基于 BERT 模型计算配置信息点与描述文本之间的相似性，有效地实现了信息点与描述文本的映射匹配。文献[19]利用轻量级 BERT 模型作为非结构化文本的预训练模型，完成了电力变压器运维领域的命名实体识别任务。由于预训练模型具有强大语义特征提取能力和泛化能力，可以考虑跳过文本预处理步骤，直接获取通道名称的全局语义特征进行匹配，为通道同源匹配提供新的思路和可行途径。当前针对同源录波数据难以匹配问题的研究尚处于起步阶段，缺乏具体的实施方法与步骤来完成同源录波数据的智能、准确、高效匹配。

为此，本文提出一种基于 Sentence-MacBERT 模型的同源录波数据匹配方法。首先，分析录波文件的记录格式和文件解读时易发生的错误问题，构建核查信息表对录波文件进行自动校核，解决录波文件出现乱码、数据缺失等问题的影响。然后，利用 Sentence-MacBERT 模型获取通道名称的特征向量，通过计算余弦相似度实现同源通道匹配。最后，根据同源通道定位提取出同源录波数据，提供给运行人员进行综合分析判断，提升智能变电站故障分析水平和效率。

1 同源录波数据匹配技术框架

电网在实际运行过程中发生故障或扰动时，保护装置和故障录波装置会启动录波记录各通道电气量的变化情况，生成 COMTRADE 格式的录波文件。然而，由于部分装置导出的录波文件可能存在采样率标识缺失、文件乱码等问题，并且同源录波数据的通道名称和通道索引号可能存在差异，增加了同源录波数据匹配的难度。对此，本文基于 Sentence-MacBERT 模型提出同源录波数据匹配技术框架，如图 1 所示。

根据图 1 中的技术流程，利用 Python 语言编写程序，在线校核录波文件并自动匹配同源录波数据，具体实现步骤如下所述。

1) 读取双重化配置的保护装置或者同一间隔下的保护装置与故障录波装置产生的同源录波文件，提取出同源配置文件与同源数据文件。

2) 根据构建的核查信息表分别对配置文件与数据文件进行校核，判断两文件的校核结果。若有文件出现校核失败，则根据核查信息表中对应的错

误类型发出相应告警信息并退出程序。若两文件均校核成功则进行下一步流程。

3) 定位提取同源配置文件中的通道名称与通道索引号，利用已训练完成的 Sentence-MacBERT 模型对提取出的通道名称进行同源匹配，并保存对应的通道索引号。

4) 根据通道索引号定位提取出同源数据文件中的对应录波数据，完成同源录波数据匹配，帮助运行人员进行综合分析诊断。

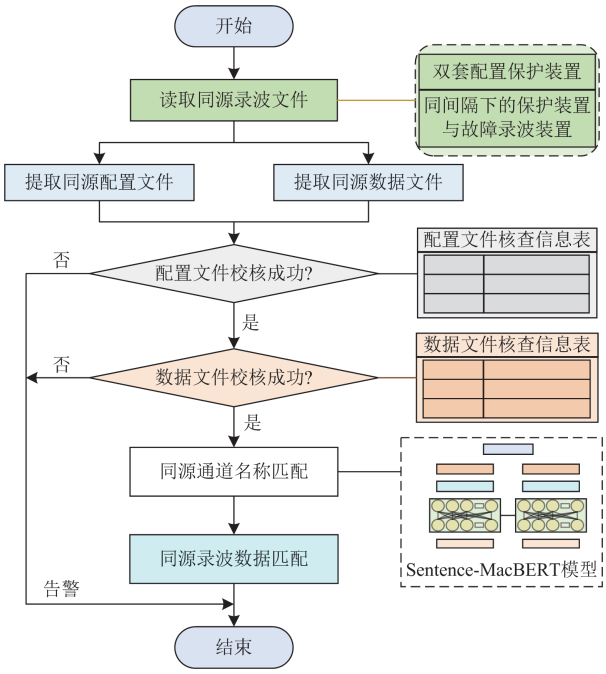


图 1　同源录波数据匹配技术框架

Fig. 1 Technical framework of homologous recording data matching

2　录波文件格式分析

录波文件主要包括 4 个子文件：头标文件、配置文件、数据文件和信息文件[20]。其中，配置文件与数据文件是录波文件当中的必选文件，也是提取录波数据的关键性文件，需对其格式特点进行分析。

2.1　配置文件标准格式

配置文件采用 ASCII 文本格式编写，用于供工作人员或计算机程序读取和分析数据文件中的录波数据，所包含的信息如下：1) 厂站名、装置标识和 COMTRADE 标准版本年号；2) 通道总数和类型；3) 通道名称、单位和转换因子；4) 标称电网频率；5) 采样率信息；6) 第一个数据点的日期和时间；7) 触发点的日期和时间；8) 数据文件的类型；9) 时标倍率因子；10) 时间编码和当地编码；11) 采样的时间品质。配置文件具有预定的标准化格式，如图 2 所示。

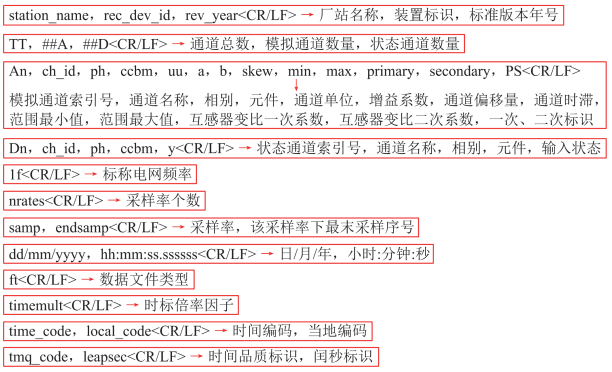


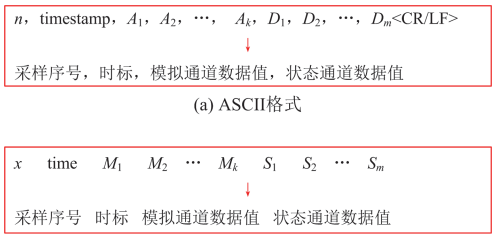
图 2　配置文件标准格式

Fig. 2 Standard format of the configuration file

由图 2 可知，配置文件的信息由若干行组成，行尾使用<CR/LF>作为每一行的结束符，每行的各个数据域以逗号分隔符“，”进行隔离。对于没有信息输入的数据域，也需保留逗号分隔符。

2.2　数据文件标准格式

数据文件用于记录采样得到的录波数据，包含每次采样的采样序号、时标和每个采样通道的数据值。数据文件格式类型可为 ASCII 格式或二进制格式，应与配置文件中定义的类型保持一致，其标准格式如图 3 所示。



(a) ASCII格式

(b) 二进制格式

图 3　数据文件标准格式

Fig. 3 Standard format of the data file

图 3(a)中的 A_k 为 ASCII 格式的第 k 个模拟通道数据值；D_m 为 ASCII 格式的第 m 个状态通道数据值。图 3(b)中 M_k 为二进制格式的第 k 个模拟通道数据值；S_m 为二进制格式的第 m 个状态通道数据值。对于 ASCII 格式的数据文件，每行的行尾仍使用<CR/LF>作为结束符，数据之间使用逗号分隔符隔离。而对于二进制格式的数据文件，行尾不再使用<CR/LF>标注，数据之间相互连续，无逗号分隔符，若任意元素缺失，变量的序列也将被破坏。

3　录波文件自动校核

由于在读取录波文件时会出现乱码、数据缺失

等情况，导致文件解析失败，无法进行后续的同源录波数据匹配。对此，本文根据配置文件、数据文件的格式特点以及时常发生的错误类型构建核查信息表，实现录波文件的自动校核，配置文件与数据文件核查信息表分别如表 1、表 2 所示。

表 1 配置文件核查信息表

Table 1 Verification information table for the configuration file

编号	错误类型	校核方案
A1	第一行信息缺失	版本年号设为 1991，解析继续
A2	第二行信息缺失或错误	解析失败，发出告警
A3	模拟通道信息行缺失	解析失败，发出告警
A4	状态通道信息行缺失	解析失败，发出告警
A5	采样率信息缺失	根据数据文件中的时标信息计算填充，若时标信息列缺失则认为解析失败，发出告警
A6	数据文件类型缺失或错误	根据数据文件有无逗号分隔符或结束符进行判断填充
A7	配置文件乱码	解析失败，发出告警

表 2 数据文件核查信息表

Table 2 Verification information table for the data file

编号	错误类型	校核方案
B1	ASCII 格式录波数据值缺失	缺失值填充为 0
B2	ASCII 格式采样序号列缺失	按顺序排号自动填充
B3	ASCII 格式时标列缺失	根据配置文件采样率信息计算填充，若采样率信息缺失则认为解析失败，发出告警
B4	ASCII 格式采样序号行缺失	根据相邻行数值计算填充
B5	ASCII 格式时标行缺失	根据相邻行数值计算填充
B6	二进制格式任意元素缺失	解析失败，发出告警
B7	数据文件乱码	解析失败，发出告警

表 1 和表 2 中列出了错误类型以及对应的校核方案。其中，编号为 A2、A3、A4、A7、B6、B7 的错误类型认为文件无法校核，解析失败并发出相应的错误信息告警；编号为 A1、B1、B2、B4、B5 的错误类型认为文件校核成功，可进行后续的同源录波数据匹配；编号为 A5、A6、B3 的错误类型需进行联合判断是否校核成功。

4 同源通道名称匹配

在进行同源录波数据匹配前首先需要匹配各个录波数据所处的通道名称和通道索引号[21]，然而智能变电站在不同建设时期配置的录波通道名称和排列顺序存在不一致情况。对此，本文根据通道名称的短文本特性，利用文本相似度匹配技术对同源通道名称进行自动匹配，并根据匹配得到的通道索引号提取同源数据文件中对应的录波数据。

4.1 BERT 预训练模型

进行文本相似度匹配首先应利用语言模型将通道名称转化为词向量的表示形式，以便计算机能够进行识别和处理[22-24]。常见的语言模型有 One-Hot、Word2Vec、ELMo、GPT 和 BERT。其中，文本经 One-Hot 编码得到的高维稀疏矩阵会浪费计算和存储资源，不同词的向量表示互相正交，无法衡量不同词之间的关系和重要程度；Word2Vec 模型能够学习到语义和语法的信息，但是训练出来的词向量属于静态 Word Embedding，无法解决多义词的问题；ELMo 采用双向 LSTM 语言模型来捕获句子的依赖关系，能够解决一词多义的问题，但是当数据数量较大时，该模型的训练速度较慢，并且模型精度没有 BERT 模型高；GPT 属于单向模型，虽然能够进行一词多义表示，但是该模型无法获取词的上下文信息。BERT 模型的普适性强，具有 Word2Vec、ELMo 以及 GPT 模型的优点，主要由双向 Transformer 的 encoder 结构组成，如图 4 所示。

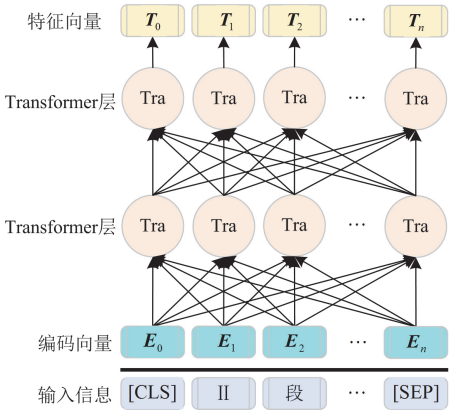


图 4 BERT 模型结构

Fig. 4 BERT model structure

图 4 中的[CLS]是用来作为输入文本开始的标志，[SEP]用来作为句子间分隔或文本结束的标志。输入信息经过双向 Transformer 编码器进行特征提取后，最终得到具有文本特定信息的动态特征向量。BERT 模型的预训练任务由掩码语言模型(masked language model, MLM)和下句预测(next sentence prediction, NSP)两个子任务构成。MLM 通过随机掩码词汇来学习上下文信息特性，从而进行词汇预测，同时赋予 BERT 模型一定的纠错能力；NSP 则通过学习句子间的特征关系来预测句子之间的位置是否相连。BERT 模型将 MLM 与 NSP 任务进行联合训练，使得 BERT 模型输出的特征向量能够表示输入文本的整体信息。

4.2 Sentence-MacBERT 模型构建

MacBERT 是在 BERT 模型基础上提出的改进预训练语言模型，其通过设计更巧妙的 MLM 任务来解决 BERT 模型在预训练任务和下游微调任务存在的不一致问题，能够提高模型的训练效果与计算速度。具体改进策略如下所述。

1) 提出 MLM 校正策略(MLM as correction, Mac)，利用相似词替代被掩码的字符，减轻了预训练和微调阶段之间误差，并随机替换没有近义词的字词。

2) 采用全词掩码策略(BERT whole-word mask, BERT-wwm)来代替随机掩码，同时利用 N-gram 掩码策略来决定需要掩码的字词。

3) 提出利用句子顺序预测(sentence order prediction, SOP)任务来代替 NSP 任务，让模型去预测两个句子的前后顺序，帮助模型获取更多的文本语义信息。

然而，仅使用 MacBERT 模型生成的特性向量进行文本相似度计算会造成巨大的计算开销，且其句子表征效果不理想。对此，本文构建 Sentence-MacBERT 模型，利用孪生网络结构对 MacBERT 模型进行微调，生成具有语义信息的句子嵌入向量，增强通道名称匹配模型的特征提取能力，提高其计算效率，模型结构如图 5 所示。

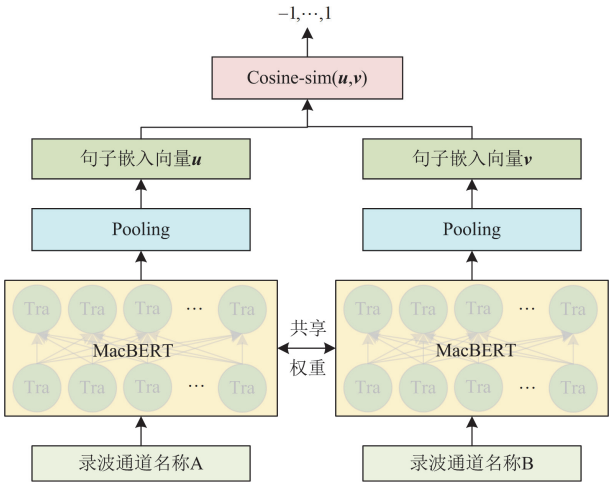


图 5 Sentence-MacBERT 模型结构

Fig. 5 Sentence-MacBERT model structure

由图 5 可看出，Sentence-MacBERT 使用孪生网络结构，将录波通道名称 A、B 传入两个相同的 MacBERT 中进行编码，MacBERT 共享相同的权重参数，并且在 MacBERT 层后加入 Pooling 层进行平均池化操作，本文采用均值池化策略计算每个字词输出向量的平均值，生成两个固定维度的句子嵌入向量 $\boldsymbol{u}$ 、 $\boldsymbol{v}$ 。通过计算两向量之间的余弦相似度 $\cos\theta$ 来度量录波通道名称的相似性，计算表达式为

$$\cos\theta=\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{|\boldsymbol{u}||\boldsymbol{v}|} \tag{1}$$

余弦相似度的绝对值越大表示通道名称越相似，本文选择相似度最高的匹配项作为最终结果，并从数据文件中提取对应的录波数据进行匹配。

5 算例分析

本文利用 Python 编程语言进行同源录波数据匹配实验，算例分析的实验环境如下：操作系统为 Windows 11，显卡为 NVIDIA GeForce RTX 4060，处理器为 Intel(R) Core(TM) i5-13500HX，内存大小为 16 GB，编程平台为 PyCharm，编程语言环境为 Python 3.7.1，建模环境为 Pytorch 1.10.0。

5.1 录波文件自动校核实验

为检验本文录波文件的自动校核效果，从智能变电站中抽取解析失败的录波文件，并选取部分正确录波文件人为制造其他错误类型。根据核查信息表利用 Python 编程语言搭建实验环境，进行录波文件自动校核实验，部分实验结果如表 3 所示。

表 3 录波文件自动校核部分实验结果

Table 3 Partial experimental results of automatic check of recording files

错误类型	编号	校核结果
标准版本年号缺失	A1	兼容至 1991 版，解析成功
缺失某行的模拟通道信息	A3	发出模拟通道信息缺失告警
配置文件采样率缺失但数据文件有时标信息	A5	根据时标信息成功计算出采样率进行填充，解析成功
数据文件类型标识缺失	A6	数据文件有逗号分隔符，填充为 ASCII 格式，解析成功
ASCII 格式数据文件缺失某个录波数据值	B1	根据逗号分隔符发现无数据项，自动填充为 0，解析成功
ASCII 格式数据文件采样序号行缺失	B4	定位出相邻行的采样序号，计算出本行采样序号，解析成功
二进制格式数据文件某录波数据值缺失	B6	发出二进制格式数据文件数据缺失告警
数据文件全乱码	B7	发出数据文件乱码告警

由表 3 可知，错误类型的校核结果与核查信息表的校核方案一致，且表 3 中未列出的其他错误类型实验结果也均一致，都可完成录波文件的自动校核，并对无法校核的录波文件发出相应的错误告警。

5.2 同源通道名称匹配实验

为验证基于 Sentence-MacBERT 模型的通道名称匹配效果，本文从智能变电站中抽取录波文件中的通道名称构建数据集进行模型训练。构建的样本

总量为 3440 条，其中正样本为同源录波通道名称文本，对应的标签为 1；负样本为非同源录波通道名称文本，对应的标签为 0，以 8:1:1 的比例划分训练集、测试集和验证集，部分数据集如表 4 所示。

表 4 部分数据集示例

Table 4 Examples of partial datasets

通道名称 A	通道名称 B	标签
218XXII 线电流 A 相 I_a	218XX2 线 a 相电流 I_a	1
线路 7 电流 B 相 I_b	线路 7 电流 $-I_B$	1
线路 4 零序电流 $3I_o$	线路 4 电流 I_o	1
218XXII 线电压 A 相 U_a	218XX2 线 b 相电压 U_b	0
218XXII 线零序电流 $3I_o$	218XX1 线零序电流	0
线路 7 电流 B 相 I_b	线路 4 电流 I_B	0

5.2.1 实验评估指标

根据混淆矩阵能够直观地看出识别结果的偏差值，只有在矩阵对角线上的数据才是被正确预测的样本比例，混淆矩阵如表 5 所示。

表 5 混淆矩阵

Table 5 Confusion matrix

实际值	预测值	
	1	0
1	T_P	F_N
0	F_P	T_N

表 5 中：T_P 代表实际值与预测值都为 1 的样本比例；F_N 代表实际值为 1 但预测值为 0 的样本比例；F_P 代表实际值为 0 但预测值为 1 的样本比例；T_N 代表实际值与预测值都为 0 的样本比例。

本文根据混淆矩阵计算出准确率 A 和 F_1 值作为通道名称匹配模型的评估指标。其中，准确率为预测结果正确的样本数据占总样本个数的百分比，计算如式(2)所示。

$$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{2}$$

然而，仅依靠准确率难以全面衡量通道名称匹配模型的识别性能，因此引入 F_1 值评估指标，F_1 值为精确率 P 和召回率 R 的调和平均值，计算如式(3)—式(5)所示。

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

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

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

F_1 值越高表示通道名称匹配模型的综合识别效果越优。

5.2.2 对比实验结果与分析

为验证本文提出的通道名称匹配模型的识别效果，本文利用构建的通道名称数据集分别训练 Sentence-BERT 模型和 Sentence-MacBERT 模型，对模型进行微调，保存最优模型进行通道名称匹配实验。其中，Sentence-MacBERT 模型的训练参数设置如表 6 所示。

表 6 Sentence-MacBERT 模型的训练参数

Table 6 Sentence-MacBERT model training parameter

训练参数	设置值	参数含义
Epoch	20	训练次数
Batch_size	16	批处理尺寸
evaluation_steps	100	评估步长
warmup_steps	150	预热步长

使用准确率 A 和 F_1 值作为 Sentence-BERT 模型和 Sentence-MacBERT 模型在训练过程中的监控指标，训练结果分别如图 6、图 7 所示。

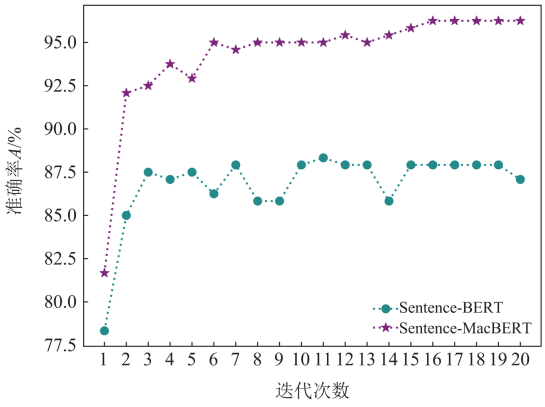


图 6 准确率的训练曲线

Fig. 6 Training curve for accuracy

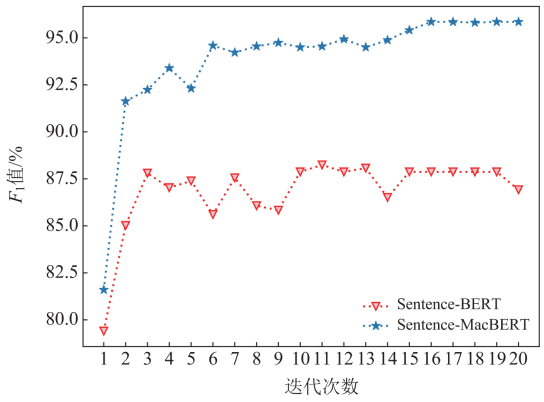


图 7 F_1 值的训练曲线

Fig. 7 Training curve for F_1

由图 6 和图 7 可以看出，Sentence-BERT 和 Sentence-MacBERT 模型在训练初期就取得较高的准确率和 F_1 值。此外，随着训练次数的增加，两种监控指标的变化趋势均为先增高然后趋于稳定。Sentence-BERT 模型在迭代 11 次时取得最优模型，准确率和 F_1 值均在 87.5%左右；Sentence-MacBERT 模型在迭代 16 次时模型取得最优模型，准确率和 F_1 值均在 95%以上。表明本文提出的 Sentence-MacBERT 模型的训练效果要优于 Sentence-BERT 模型。

为进一步验证本文所提 Sentence-MacBERT 模型的优越性，本文使用原始 BERT、MacBERT 模型针对测试集直接进行通道名称匹配实验，然后利用微调训练后保存的最优模型 Sentence-BERT 和 Sentence-MacBERT 模型，针对同一测试集进行比对实验，4 种模型的实验结果如表 7 所示。

表 7 不同模型的实验结果

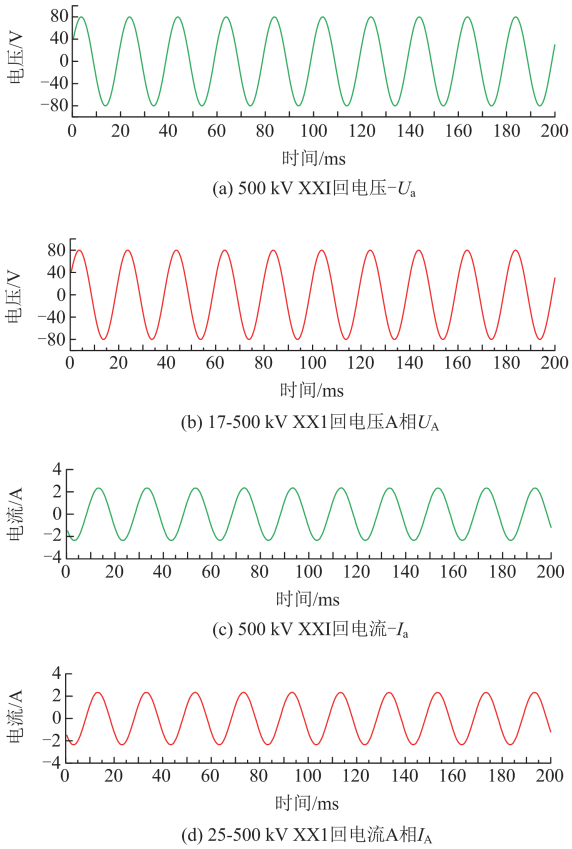
Table 7 Experimental results of different models

模型	准确率 A/%	F_1 值/%
BERT	67.08	72.47
MacBERT	74.17	75.17
Sentence-BERT	85.42	87.80
Sentence-MacBERT	95.83	96.09

由表 7 可知，经过微调之后的 Sentence-BERT 模型相比于 BERT 模型，其准确率提升了 18.34%，F_1 值提升了 15.33%；Sentence-MacBERT 模型相比于 MacBERT 模型，其准确率提升了 21.66%，F_1 值提升了 20.92%。Sentence-MacBERT 模型的准确率能够达到 95.83%，相比于 Sentence-BERT 模型提升了 10.41%，且其 F_1 值达到了 96.09%，相比于 Sentence-BERT 模型提升了 8.29%。表明了本文所提的 Sentence-MacBERT 模型在通道名称匹配方面具有更好的识别效果，能够有效地完成同源录波通道名称匹配任务。

5.3 同源录波数据匹配实验

本文利用训练好的 Sentence-BERT 模型匹配同源通道名称与通道索引号，然后根据录波通道索引号定位提取出同源数据文件中对应的录波数据，完成同源录波数据匹配。以智能变电站导出的某同源录波文件为例，进行同源录波数据匹配实验，实验结果如图 8、表 8 所示。



(a) 500 kV XXI回电压−U_a

(b) 17-500 kV XX1回电压A相U_A

(c) 500 kV XXI回电流−I_a

(d) 25-500 kV XX1回电流A相I_A

图 8 同源录波数据匹配结果

Fig. 8 Matching results of homologous recording data

表 8 同源通道索引号匹配结果

Table 8 Matching results of the index number of the homologous channels

通道名称	通道索引号	同源通道名称	同源通道索引号	匹配的正误
500 kV XXI 回电压 $-U_a$	21	17-500 kV XX1 回电压 A 相 U_A	17	正确
500 kV XXI 回电压 $-U_b$	22	18-500 kV XX1 回电压 B 相 U_B	18	正确
500 kV XXI 回电压 $-U_c$	23	19-500 kV XX1 回电压 C 相 U_C	19	正确
500 kV XXI 回电压 $-U_o$	24	20-500 kV XX1 回零序电压	20	正确
500 kV XXI 回电流 $-I_a$	25	25-500 kV XX1 回电流 A 相 I_A	25	正确
500 kV XXI 回电流 $-I_b$	26	26-500 kV XX1 回电流 B 相 I_B	26	正确
500 kV XXI 回电流 $-I_c$	27	27-500 kV XX1 回电流 C 相 I_C	27	正确
500 kV XXI 回电流 $-3I_o$	28	28-500 kV XX1 回零序电流	28	正确

由表 8 可看出，同源录波文件的同源通道均能被正确匹配。图 8 是根据通道索引号提取出的部分录波数据，实现了录波数据的同源匹配，验证了本文方法的有效性。

6 结论

提出了一种基于 Sentence-MacBERT 模型的同源录波数据匹配方法，实现了同源通道名称以及通道索引号存在差异情况下的同源录波数据匹配。

1) 完成了录波文件核查信息表的构建，能够对录波文件进行自动校核，增强了文件解析的容错能力，并对解析失败的文件发出错误告警信息。

2) 构建了 Sentence-MacBERT 同源通道匹配模型，经实验验证该模型的准确率能够达到 95.83%，F_1 值能够达到 96.09%，可以有效完成同源通道匹配任务。

3) 提出了同源录波数据匹配技术框架，帮助运行人员获取同源录波数据进行故障分析，增强了录波数据的整合利用水平，提高了智能变电站的运维水平。

参考文献

[1] 崔亚芹, 熊蕙, 祁忠, 等. 基于保信主站的双重化配置继电保护装置录波同源比对研究[J]. 电气技术, 2022, 23(7): 74-80.
CUI Yaqin, XIONG Hui, QI Zhong, et al. Research on recording homology comparison of relay protection devices with dual configuration based on protection information management master station[J]. Electrical Engineering, 2022, 23(7): 74-80.

[2] 叶艳军, 陈水耀, 潘武略, 等. 基于控保信号序列化和力引导算法的高压直流故障智能诊断可视化技术研究[J]. 电力系统保护与控制, 2023, 51(19): 155-163.
YE Yanjun, CHEN Shuiyao, PAN Wulüe, et al. Intelligent diagnosis and visualization technology for high voltage DC faults based on control and protection signal serialization and a force guidance algorithm[J]. Power System Protection and Control, 2023, 51(19): 155-163.

[3] 彭曙蓉, 郭丽娟, 陈慧霞, 等. 基于特征增强的变电站保护装置录波通道同源匹配研究[J]. 电力科学与技术学报, 2024, 39(6): 53-59.
PENG Shurong, GUO Lijuan, CHEN Huixia, et al. Matching of homologous recording channels of substation protection devices based on feature enhancement[J]. Journal of Electric Power Science and Technology, 2024, 39(6): 53-59.

[4] 戴志辉, 张富泽, 杨鑫, 等. 基于 FDTW 算法的故障录波数据智能比对方法[J]. 电力系统保护与控制, 2023, 51(23): 82-91.
DAI Zhihui, ZHANG Fuze, YANG Xin, et al. Intelligent comparison method for fault record waveform data based on the FDTW algorithm[J]. Power System Protection and Control, 2023, 51(23): 82-91.

[5] 严敬汝, 臧谦, 赵宇皓, 等. 基于配网录波特征库的故障识别与保护定值整定及实现[J]. 电力科学与技术学报, 2023, 38(2): 248-254.
YAN Jingru, ZANG Qian, ZHAO Yuhao, et al. Implementation of distribution network fault identification and protection setting based on characteristic recording data map[J]. Journal of Electric Power Science and Technology, 2023, 38(2): 248-254.

[6] 陈泗贞, 梁竞雷, 卢迪勇, 等. 基于 COMTRADE 模型的电力系统多源故障数据融合分析方法[J]. 电力科学与技术学报, 2019, 34(3): 92-100.
CHEN Sizhen, LIANG Jinglei, LU Diyong, et al. Multi-source fault data comprehensive analysis method for power system based on COMTRADE model[J]. Journal of Electric Power Science and Technology, 2019, 34(3): 92-100.

[7] 谢坤, 余华武, 陈福锋, 等. 一种兼容 Comtrade 格式用于间隔层故障诊断的新型智能终端[J]. 电力系统自动化, 2016, 40(4): 111-114.
XIE Kun, YU Huawu, CHEN Fufeng, et al. A novel smart terminal with Comtrade analysis and bay level online fault diagnosis function[J]. Automation of Electric Power Systems, 2016, 40(4): 111-114.

[8] YAO Zhiqing, LI Danyang, LI Zhiyong, et al. Relay protection mirror operation technology based on digital twin[J]. Protection and Control of Modern Power Systems, 2023, 8(4): 913-926.

[9] 叶远波, 程晓平, 张兆云, 等. 电力系统故障区域录波自动分析关键技术[J]. 中国电力, 2022, 55(4): 93-99.
YE Yuanbo, CHENG Xiaoping, ZHANG Zhaoyun, et al. Key technology of automatic analysis of fault area wave recording of power system[J]. Electric Power, 2022, 55(4): 93-99.

[10] 吴楠, 史明明, 朱卫平, 等. 暂态录波型故障指示器的单相接地故障研判应用[J]. 南方电网技术, 2021, 15(1): 61-68.
WU Nan, SHI Mingming, ZHU Weiping, et al. Application on transient recording fault indicator for single-phase grounding fault diagnosis[J]. Southern Power System Technology, 2021, 15(1): 61-68.

[11] 巫聪云, 徐晓峰, 钟洁, 等. 继电保护多源异构信息智能化处理关键技术研究[J]. 云南电力技术, 2019, 47(5): 53-56.
WU Congyun, XU Xiaofeng, ZHONG Jie, et al. Key technologies and applications of intelligent multi-source heterogeneous data processing for relay protection[J].

Yunnan Electric Power, 2019, 47(5): 53-56.

[12] 戴志辉, 杨鑫, 刘悦, 等. 基于增量学习优化的故障录波文件通道名称识别方法[J]. 电力系统保护与控制, 2023, 51(4): 148-156.
DAI Zhihui, YANG Xin, LIU Yue, et al. Recognition method of fault recorder file channel name based on incremental learning optimization[J]. Power System Protection and Control, 2023, 51(4): 148-156.

[13] 王冠南, 郭丽娟, 彭曙蓉, 等. 基于正则表达式和 Jaccard 系数的智能变电站录波通道同源匹配[J]. 浙江电力, 2024, 43(1): 20-27.
WANG Guannan, GUO Lijuan, PENG Shurong, et al. Homologous matching of recording channels in intelligent substations based on regular expression and Jaccard similarity coefficient[J]. Zhejiang Electric Power, 2024, 43(1): 20-27.

[14] 范卫东, 冯晓伟, 董金星, 等. 基于历史数据语义相似度的智能变电站虚端子自动连接[J]. 电力系统保护与控制, 2020, 48(17): 179-186.
FAN Weidong, FENG Xiaowei, DONG Jinxing, et al. Automatic matching method of a virtual terminal in intelligent substation based on semantic similarity of historical data[J]. Power System Protection and Control, 2020, 48(17): 179-186.

[15] RONG Lu, DING Yijie, WANG Mengyao, et al. A multi-modal ELMo model for image sentiment recognition of consumer data[J]. IEEE Transactions on Consumer Electronics, 2024, 70(1): 3697-3708.

[16] YAN Ziming, XU Yan. Real-time optimal power flow with linguistic stipulations: integrating GPT-agent and deep reinforcement learning[J]. IEEE Transactions on Power Systems, 2024, 39(2): 4747-4750.

[17] 田波, 张越, 蒙飞, 等. 电网故障处置信息自适应理解框架及关键技术[J]. 中国电力, 2024, 57(7): 188-195.
TIAN Bo, ZHANG Yue, MENG Fei, et al. Adaptive Understanding framework and Key technology of power grid fault disposal information[J]. Electric Power, 2024, 57(7): 188-195.

[18] 叶远波, 李端超, 谢民, 等. 基于知识图谱的二次设备测试自动配置方法[J]. 电力系统保护与控制, 2022, 50(12): 162-171.
YE Yuanbo, LI Duanchao, XIE Min, et al. Automatic configuration method of secondary equipment test based on a knowledge graph[J]. Power System Protection and Control, 2022, 50(12): 162-171.

[19] 谢庆, 蔡扬, 谢军, 等. 基于 ALBERT 的电力变压器运维知识图谱构建方法与应用研究[J]. 电工技术学报, 2023, 38(1): 95-106.
XIE Qing, CAI Yang, XIE Jun, et al. Research on construction method and application of power transformer operation and maintenance knowledge map based on ALBERT[J]. Transactions of China Electrotechnical Society, 2023, 38(1): 95-106.

[20] 量度继电器和保护装置 第 24 部分: 电力系统暂态数据交换(COMTRADE)通用格式: GB/T 14598. 24—2017[S]. 北京: 中国标准出版社, 2017.
Measuring relays and protection equipment—part 24: common format for transient data exchange (COMTRADE) for power systems: GB/T 14598. 24—2017[S]. Beijing: Standards Press of China, 2017.

[21] 常风然, 萧彦, 张洪, 等. 智能变电站录波方案的探索与实践[J]. 电力自动化设备, 2011, 31(1): 109-112.
CHANG Fengran, XIAO Yan, ZHANG Hong, et al. Research and practice of wave recording scheme for smart substation[J]. Electric Power Automation Equipment, 2011, 31(1): 109-112.

[22] 赵俊华, 文福拴, 黄建伟, 等. 基于大语言模型的电力系统通用人工智能展望: 理论与应用[J]. 电力系统自动化, 2024, 48(6): 13-28.
ZHAO Junhua, WEN Fushuan, HUANG Jianwei, et al. Prospect of artificial general intelligence for power systems based on large language model: theory and applications[J]. Automation of Electric Power Systems, 2024, 48(6): 13-28.

[23] ZHANG Dongwen, XU Hua, SU Zengcai, et al. Chinese comments sentiment classification based on word2vec and SVMperf[J]. Expert Systems with Application, 2015, 42(4): 1857-1863.

[24] 刘中硕, 郑少明, 陶畅, 等. 继电保护装置缺陷文本专业词典构建及其语言特性分析[J]. 中国电力, 2023, 56(7): 146-155.
LIU Zhongshuo, ZHENG Shaoming, TAO Chang, et al. The construction of the professional dictionary of relay protection defect text in a regional power grid and its natural language characteristics analysis[J]. Electric Power, 2023, 56(7): 146-155.

收稿日期：2024-07-02；　修回日期：2024-09-07

作者简介：

戴志辉(1980—)，男，博士，教授，研究方向为电力系统保护与控制；E-mail: daihuadian@163.com

张富泽(1998—)，男，硕士研究生，研究方向为电力系统保护与控制。E-mail: zhangfuze520@163.com

(编辑　张　颖)