

基于模糊自然语言处理的铁路CTC接口文本智能测试方法

角远韬, 李润梅, 王剑

(北京交通大学自动化与智能学院, 北京 100044)

摘要: 模糊自然语言处理将模糊理论应用在自然语言处理(NLP)的任务中, 随着大模型与人工智能的不断发展, 有关文本数据的研究不断深化。铁路调度集中控制(CTC)系统作为大型复杂系统, 各子系统、服务器软件间的接口数据均以日志文本格式存储与传输。由于其具有文本数量多、文本类型杂等特点, 提出了一种模糊自然语言处理的方法, 解决CTC系统接口数据的人工测试难题。模糊C均值(FCM)聚类算法将日志文本分为不同的标签类别, 并将其作为NLP任务中命名实体识别的标签输入, 在传统BiLSTM-CRF模型上引入BERT进行文本编码, 更准确地理解文本之间的关系并提高文本识别的精确度。根据前序训练模型, 研发了铁路CTC系统日志文本接口测试的智能验证工具, 其可以改善目前CTC系统的人工测试现状, 帮助测试人员进行接口测试验证, 提升测试工作的智能化、自动化水平。

关键词: 自然语言处理; 模糊文本聚类; 铁路调度集中控制系统; 命名实体识别; 智能测试

中图分类号: U284.5

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202419

Intelligent testing method for railway CTC interface data based on fuzzy natural language processing

JIAO Yuantao, LI Runmei, WANG Jian

School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China

Abstract: Fuzzy natural language processing applies fuzzy theoretical knowledge to the task of natural language processing (NLP). With the continuous development of large language model and artificial intelligence, research on text data continues to deepen. As a large and complex system, the interface data between various subsystems and server software are stored and transmitted in log text format. Due to its large number of texts and miscellaneous text types, a fuzzy NLP method was proposed to solve the problem of manual testing the interface data of centralized traffic control (CTC) system. The fuzzy C-means (FCM) clustering algorithm divided the log text into different label categories, which was used as the label input for named entity recognition in NLP tasks, and BERT was introduced on the traditional BiLSTM-CRF model for text encoding, which understood the relationship between texts more accurately and improved the accuracy of text recognition. An intelligent verification tool for log-text interface testing of railway CTC system was presented based on an improved training model, which enhanced the current manual testing process of CTC system, assisted testing staff in verifying the interface testing, and increased the level of intelligence and automation in testing work.

Key words: natural language processing, fuzzy text clustering, railway centralized traffic control system, named entity recognition, intelligent testing

收稿日期: 2024-04-28; 修回日期: 2024-05-27

通信作者: 李润梅, rmli@bjtu.edu.cn

基金项目: 国家重点研发计划(No.2022YFB4300500); 中国国家铁路集团有限公司科技研究开发计划(No.L2022X002)

Foundation Items: National Key R&D Program of China(No.2022YFB4300500), Technological R&D Program of China state Railway Group Co., Ltd. China Railway Project(No.L2022X002)

0 引言

随着人工智能的推动,大语言模型的发展取得了快速进步,以自然语言为范式的文本数据的研究工作不断深入^[1]。命名实体识别(named entity recognition, NER)作为其中的关键任务之一,在理解文本内容方面扮演着至关重要的角色^[2]。其核心任务是对文本进行类别标注,通过标注后的数据识别其他相似文本中的实体,从而完成命名实体识别工作,即命名实体与实体识别。

目前命名实体识别在自然语言处理(natural language processing, NLP)领域得到了广泛应用,涵盖了舆情分析^[3]、医学诊断^[4]以及文学研究^[5]等多个领域。黄炜等^[6]提出基于双向循环神经网络—条件随机场(bidirectional long short-term memory-conditional random field, BiLSTM-CRF)的涉恐信息模型进行命名实体识别,但无法处理字符级别的文字编码。于祥钦等^[7]利用卷积神经网络(convolutional neural network, CNN)获取字符级别特征的向量,输入BiLSTM-CRF模型中进行生物医学领域的命名实体识别,但缺乏对上下文的理解能力。谷歌公司提出的基于Transformer的双向编码器(bidirectional encoder representations from transformer, BERT),使模型预训练的能力得到了显著提升,尤其在文本编码与上下文理解方面表现出了很大优势,为解决上述问题提供了新的思路和方法^[8]。

铁路调度集中控制(centralized traffic control, CTC)系统作为铁路大型系统,包含十几个子系统,各个系统之间的数据传输与软件日志均以自然语言文本形式储存,日志文本数据的接口测试工作至关重要^[9]。然而,目前测试工作仍然依赖人工完成,效率低且容易出错。结合NLP的命名实体识别方法,可以智能化地完成文本识别与接口测试,从而提升铁路CTC系统接口测试的准确率与效率^[10]。在CTC系统中,服务器软件产生的大量日志文本需要进行重点测试,其内容涉及不同的类别。因此,确定文本实体的命名类别至关重要,需要利用文本聚类方法进行数据挖掘与分析。

文本聚类作为数据处理的常见算法之一,其目的是寻找大量相似的数据,并将相似的数据分成一类。目前主流的文本聚类方法包括层次聚类方法、划分聚类方法与模糊聚类方法等。刘昕等^[11]提出了多层次文本聚类方法,解决了大规模文本数据中潜

在知识发现的问题。蔡莉等^[12]提出基于网格划分思想的聚类算法,解决了现有算法不适应性的问题,增强了分布式计算能力。刘雅萱等^[13]提出一种自适应模糊算法,提高了多语义文本的聚类准确度。

本文将模糊文本聚类方法与命名实体识别任务结合,构建模糊自然语言处理模型,提出一种智能化的铁路CTC系统服务器软件的文本日志接口测试方法,其创新点如下。

(1) 模糊C均值(fuzzy C-means, FCM)聚类算法生成日志文本的高频测试类别,作为命名实体识别任务的训练标签实体,有效替代传统的命名实体识别任务中的标签类别。

(2) 改进传统的BiLSTM-CRF算法,结合大模型中的BERT进行文本编码,更高效、快速地生成并验证接口测试的文本内容。

(3) 研发铁路CTC系统服务器日志文本接口测试的智能验证工具,辅助测试人员完成测试用例文本数据部分的验证,从而改善人工测试工作的现状,提升测试工作的智能化与自动化水平。

1 铁路CTC服务器的接口测试

1.1 铁路CTC系统发展介绍

铁路CTC系统是高速铁路运输组织中典型的大型分布式复杂系统,其结构庞大,如图1所示。不同子系统对应一个或多个子软件,维持整个CTC系统运转的软件有近十几个,因此保障各软件间正常的信息传输至关重要。铁路CTC系统需要具备与其他车站通信的服务器、临时限速服务器(temporary speed restriction server, TSRS)、铁路综合数字移动通信系统(global system for mobile communication-railway, GSM-R)、中国列车控制系统(China train control system, CTCS)、区间闭塞设备、铁路数字相邻的调度集中/列车调度指挥系统、信号集中监测系统、无线调车机车信号和监控系统、运输信息集成平台等外部系统连接的接口,各系统软件间接口的信息文本传输是CTC测试工作的重点^[14]。

中国的调度集中系统引入日本的分散自律概念后,不断优化体系结构,形成了最初的CTC 1.0系统。随着中国铁路建设的进一步发展,CTC系统的功能得到不断扩充,逐渐发展为CTC 2.0系统。

2016年,中国铁路总公司发布了《调度集中系统技术条件》(Q/CR 518-2016),规定了CTC系

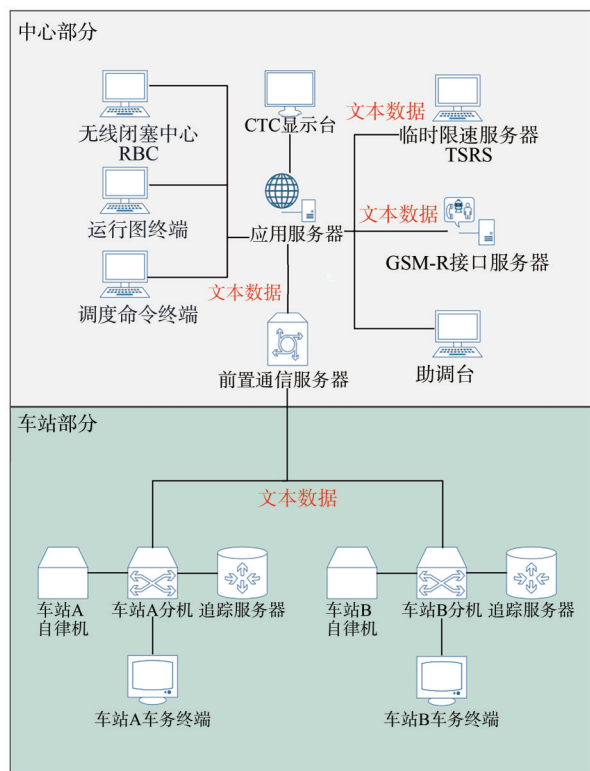


图1 铁路CTC系统结构

统的特点，形成CTC 3.0系统，其在功能上不断完善，系统架构越来越庞大。

CTC系统在实际投入使用前，需要进行完整的测试，测试工作分为若干阶段：CTC单元测试、CTC集成测试、CTC功能测试、CTC接口测试与回归测试。CTC系统以往版本的接口测试数据均以文本形式储存，扩充了多个其他功能软件接口的CTC 3.0系统迫切需要一种高效的测试方法，以测试CTC系统与其他铁路列车控制系统以及数据库的数据传输和数据交换，进而验证其可靠性。接口测试的重点对象为服务器软件接收其他系统信息而生成的日志文本数据^[15]。

1.2 服务器的日志文本分析

铁路调度集中系统的服务器包含应用服务器和通信服务器，均采用双机热备系统，各分为两台主备机^[16]。

应用服务器的主要功能包括接收自律机发送的站场等信息并以日志文本存储；响应其他终端软件的操作命令，通过数据库访问接口以文本形式保存存储列车运行计划等信息。

通信服务器起到连接应用服务器与其他车站终端通信服务的作用，以统一的方式接受客户端的连

接，并根据其他软件数据帧中的实体目的地址进行消息转发，转发数据均以文本形式存储，包括中文自然语言描述、特定数字编码以及特定英文符号。

CTC服务器软件在进行调度业务办理时产生的日志文本如图2所示，这一日志文本是后续进行接口测试以及信息传输正确性验证的重要参考依据，图2下划线标识部分为重点识别的日志文本数据。目前服务器软件日志文本的接口测试工作，仍然以人工手动测试验证为主，包括设计测试用例、划分测试场景、确认输入操作、明确预期结果与手动查找对应软件服务器的日志。

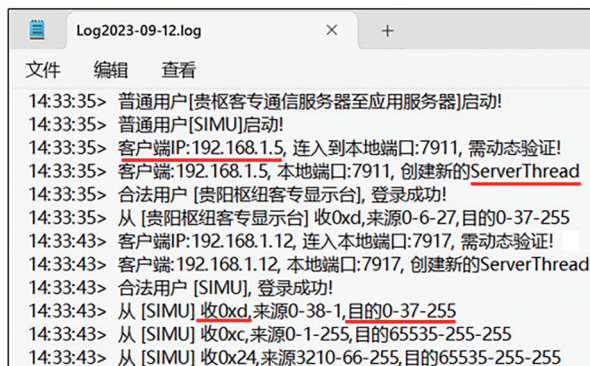


图2 CTC服务器软件的日志文本

CTC接口测试的用例示例集如图3所示。按照测试用例的要求，在图2所示的日志文本上进行文本搜索与验证，该过程耗费大量人力、物力与时间。由于服务器存储的文本数量较多，一分钟就会存储上百条数据，该接口文本日志的查找、定位、判断以及撰写测试报告的效率低下，且容易出现错误判断等工作失误。

因此，提高测试工作中日志文本数据的验证效率，采用更智能寻优的测试方法代替人工测试模式已经成为铁路CTC系统接口测试研究的关键。

2 模糊自然语言处理模型构建

铁路CTC系统服务器软件接口的日志文本中最为关键的问题是如何用机器智能判定，取代人工验证接口测试，即确定研究对象的日志文本后，需要选择合适的自然语言处理模型。

本节将从文本聚类与命名实体识别两个部分进行分析研究。其中，文本聚类将采用模糊C均值聚类算法对服务器的日志进行聚类，算法输出高频文本的聚类类别作为后续工作的实体标签。命名实体

输入	预期结果
1) G网接口配置ProcLogOutInfo=1, SendLogOutInfo=1; 2) 设置应用服务器日志记录G网接口通道 0x0E详细信息; 3) 使用GRIS仿真软件发送K无线车次号信息, 时刻T1; 4) 使用GRIS仿真软件发送K无线车次注销信息, 时刻T2;	1. T1时, G网接口添加K无线车次号信息, 车次显示有黄框; 2. T2时, 查看应用服务器日志, 存在无线 车次注销信息0x0E-0x04, 验证信息中 功能码、站码、车次号长度、车次号、 机车类型、机车号长度、机车号等内容 填写正确;

图3 CTC接口测试的用例集示例

识别根据聚类标签进行数据标注、模型训练与接口测试生成验证, 自然语言模型采用BERT, 结合传统BiLSTM-CRF模型进行实体识别。模糊自然语言处理模型的流程如图4所示。

2.1 文本聚类算法

文本聚类是数据挖掘常用的算法之一, 旨在从大量的文本文件中发现相似的自然语言数据, 并揭示数据之间的潜在特征。常见的文本聚类方法包括划分聚类方法^[17]、层次聚类方法^[18]与模糊聚类方法^[19]。其中典型的划分聚类方法有K-means聚类方法, 即首先划分一个初始簇, 然后根据距离函数不断更新簇里的数据, 最终使每个簇里的相似度尽可能高。但K-means聚类方法是一种硬分配方法, 意味着每个数据点只能属于一个簇, 忽略了数据点可

能属于多个簇的情况。层次聚类对给定数据集进行层次性分解, 常见的层次聚类算法包括DIANA (divisive analysis) 算法、BIRCH (balanced iterative reducing and clustering using hierarchies) 算法。但层次聚类计算复杂高且具有不可逆性, 当数据集发生微小变化时, 需要重新运行整个聚类算法以获得新的层次结构。

与传统的聚类算法不同, 模糊文本聚类算法基于模糊集合理论, 模糊集合具有更灵活的隶属关系, 允许元素部分属于一个或多个集合。在模糊集合中, 元素的隶属度可以是[0,1]的任意值, 更具灵活性。模糊聚类方法通过构造隶属度函数, 对文本数据进行隶属度计算, 选择隶属度大的类别划分聚类。典型的模糊聚类方法有模糊C均值聚类算法。

因此, 相比其他文本聚类方法, 模糊文本聚类算法能够更好地处理本文的研究对象, 即铁路CTC系统服务器日志文本这一类中英文与数字交杂的模糊文本数据, 为后续文本分类与命名实体识别提供有效数据。

2.2 模糊C均值聚类算法

模糊C均值聚类算法是一种软聚类方法。FCM聚类算法将模糊理论用于数据聚类, 是一种基于隶属度的聚类算法, 每个数据点都有一个隶属度矩阵表示其对每个聚类的隶属度, 其数学表达如下。

(1) 隶属度矩阵 U

在FCM聚类算法中, 每个样本 x_i 与每个聚类中心 C_j 之间的隶属度用一个矩阵 $U=[u_{ij}]$ 来表示。其中, u_{ij} 表示样本 x_i 属于聚类 j 的隶属度, C 表示聚类中心, c 表示聚类数, m 是模糊度参数, $J_m(U, C)$ 为最小化目标函数, 用来衡量样本点到聚类中心的

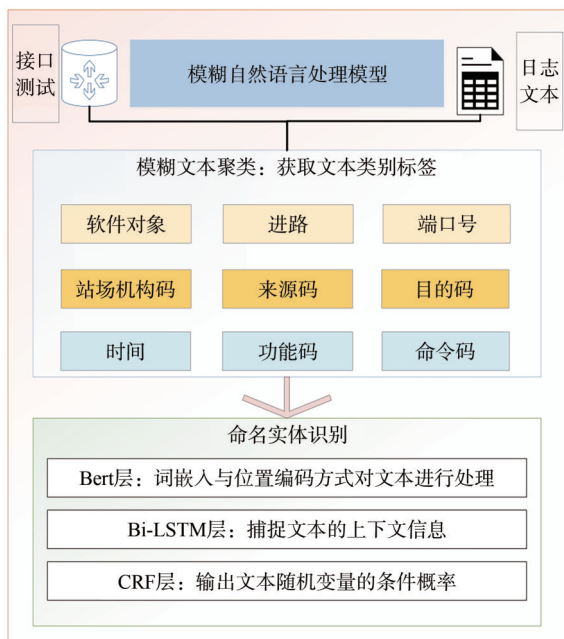


图4 模糊自然语言处理模型的流程

平方距离和隶属度之间的加权距离。隶属度矩阵满足以下约束：

$$0 \leq u_{ij} \leq 1 \quad (1)$$

$$J_m(U, C) = \sum_{j=1}^c u_{ij} = 1 \quad (2)$$

(2) 最小化目标函数 $J_m(U, C)$

给定数据集 $X = \{x_1, x_2, \dots, x_n\}$ 以及聚类数 c ，FCM 聚类算法致力于找到 c 个聚类中心 C_j ，使目标函数最小化，目标函数如式 (3) 所示。

$$J_m(U, C) = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m \cdot \|x_i - C_j\|^2 \quad (3)$$

其中， $x_i - C_j$ 是样本 x_i 与聚类中心 C_j 之间的欧氏距离。

(3) 更新聚类中心 C

每次迭代都需要更新聚类中心 C ，聚类中心 C_j 更新为：

$$C_j = \frac{\sum_{i=1}^n (u_{ij})^m \cdot x_i}{\sum_{i=1}^n (u_{ij})^m} \quad (4)$$

(4) 更新隶属度矩阵 U

每次迭代过程都需要更新隶属度矩阵 U ，具体如下。

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - C_j\|}{\|x_i - C_k\|} \right)^{\frac{2}{m-1}}} \quad (5)$$

(5) 终止条件

FCM 聚类算法的迭代在满足以下任意一个终止条件时停止：达到最大迭代次数；目标函数 $J_m(U, C)$ 的变化小于特定阈值；隶属度矩阵 U 的变化小于特定阈值。

模糊理论以一种模糊的方式处理不确定的文本数据，铁路 CTC 系统服务器日志文本内容的模糊性质可通过模糊数学的模型来表达，有利于文本数据的搜集、分类与管理，提高接口测试水平。

2.3 命名实体识别任务

命名实体识别本质上是一个序列标注模型，根据命名实体类别进行文本标注，再从相似文本中找出相似实体进行识别，即命名实体与实体识别。

目前，NER 任务在自然语言处理领域的许多场景有着广泛应用，比如构建领域知识库、机器翻译、情感分析等。早期的 NER 任务主要采用基于统计学的方法，包括隐马尔可夫模型 (hidden Mar-

kov model, HMM) 与条件随机场 (conditional random field, CRF) 等，该方法容易实现全局最优结果，但训练时间过长^[20]。随着神经网络的飞速发展，基于深度学习的 NER 方法应运而生，如深度学习与条件随机场相结合的方法。深度学习可采用卷积神经网络 (convolutional neural network, CNN)、循环神经网络 (recurrent neural network, RNN) 以及长短期记忆 (long short-term memory, LSTM) 网络等典型深度学习模型。基于深度学习的 NER 方法针对文本数据的模型训练速度得到了较大提升，但模型上下文理解与泛化能力不足^[21]。

随着人工智能与大语言模型的快速发展，GPT、BERT 等基于 Transformer 的注意力机制的模型被不断应用到自然语言处理任务中。目前 NER 任务最流行的方法为 BiLSTM-CRF 模型，在此基础上，利用 BERT 模型强大的文本表示和理解能力进行文本编码，从而更好地研究文本数据^[22]。

2.4 BERT+BiLSTM-CRF

本节提出基于 BERT 和 BiLSTM-CRF 的铁路 CTC 系统服务器日志的命名实体识别模型，该模型包含 3 层：BERT 层、Bi-LSTM 层与 CRF 层。其中，BERT 层用于初步对接口文本进行语义信息位置编码，将自然语言文字转化为计算机可识别的数字。Bi-LSTM 层则根据位置编码结果进一步提取文本的上下文特征，从而更好地理解文本表示。最后将上下文特征输入 CRF 层，结合文本聚类输出的序列标签对其状态进行约束优化，最终输出全局最优序列。

(1) BERT 层

BERT，即基于 Transformer 的双向编码器，是一种深度学习语言表示模型，BERT 拥有强大的语言表征能力和特征提取能力，特别擅长处理长文本序列数据。

BERT 使用注意力机制取代了传统的 RNN 层，以生成向量表示，引入词嵌入与位置编码对文本进行处理，可以提升对文本上下文关联关系的学习，从而更好地处理接口文本数据。词嵌入的向量包括字向量与标识符向量，其中，<CLS> 代表一个序列的开始，<SEP> 用于分隔两个测试用例，<EOS> 代表一个序列的结束，<PAD> 为补全未达到设定长度的句子的字符，通常为一个或多个字符。处理系统生成的文本数据时，一个长句通常被拆分为 2 个

短句, 分别用 sent1、sent2 表示。

$$\text{Sents} = \langle \text{CLS} \rangle + \langle \text{sent1} \rangle + \langle \text{SEP} \rangle + \langle \text{sent2} \rangle + \langle \text{EOS} \rangle + \langle \text{PAD} \rangle + \dots + \langle \text{PAD} \rangle \quad (6)$$

BERT 层用位置编码将上下文文本连接起来, 选择模型自带的字典 (vocab.txt) 对所有测试用例进行逐字符位置编码, 用正余弦函数表示当前单词在语句中的绝对位置, 正余弦函数乘积得到单词在语句中的相对位置。第 $2i$ 和 $2i+1$ 维的位置编码值分别如式 (7)、式 (8) 表示。

$$\text{PE}_{(\text{pos}, 2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \quad (7)$$

$$\text{PE}_{(\text{pos}, 2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \quad (8)$$

其中, d_{model} 为词向量编码的位置, i 指词向量的第 i 维, $2i$ 和 $2i+1$ 分别代表偶数维度和奇数维度, pos 为当前词语在序列中的位置。

(2) Bi-LSTM 层

Bi-LSTM 是对传统 LSTM 模型中的 RNN 结构进行改进的方案, 用于增强对文本上下文关系特征的学习。该方案利用引入的记忆单元与门限机制控制文本信息的遗忘、输入和输出, 有效解决了 RNN 中的梯度消失与爆炸问题。双向 LSTM (Bi-LSTM) 的简单表达式可以通过拼接正向和反向隐藏状态得到:

$$h_t^{\text{Bilstm}} = [h_t^{\text{forward}}; h_t^{\text{backward}}] \quad (9)$$

其中, h_t^{Bilstm} 是双向 LSTM 在时间步长 t 上的输出, h_t^{forward} 是正向 LSTM 在时间步长 t 的隐藏状态, h_t^{backward} 是反向 LSTM 在时间步长 t 的隐藏状态, “;” 表示拼接操作, 该操作将正向和反向的隐藏状态在特定时间步上按照维度进行连接, 形成一个更丰富的表示。

(3) CRF 层

条件随机场是模型在随机给定一组输入变量的情况下输出的条件概率, 其本质上是一种判别式概率无向图模型。CRF 模型对给定的 Bi-LSTM 输出文本 x , 其预测的标签序列为 y 的概率如下:

$$P(y|x) = \frac{\exp(s(x, y))}{\sum_y \exp(s(x, y'))} \quad (10)$$

$$s(x, y) = \sum_i P_E(x_i, y_i) + P_T(y_{i-1}, y_i) \quad (11)$$

其中, P_E 为 Bi-LSTM 输出给定文本的概率, P_T 为对应 CRF 模型的转移概率, 即转移矩阵对应的数值, i 表示在序列中的位置, y 表示整个序列的标签, y'

为输出文本 x 下所有可能的标签序列。

3 CTC 接口日志文本的智能测试实验

为处理铁路 CTC 系统服务器软件的日志文本文件, 完成服务器软件与其他终端软件的接口测试工作, 实验部分包含以下步骤。

首先, 模糊 C 均值聚类方法将 CTC 系统的日志文本数据作为输入, 经过初始化文本参数、设置迭代次数、构建隶属度矩阵、划分和最终输出文本聚类中心等步骤, 获得服务器日志文本中最高频且能代表文本类型的关键词语, 将其作为聚类中心。

其次, 将文本聚类中心作为下一步骤命名实体识别任务中的分类标签, 对 BERT 层、Bi-LSTM 层与 CRF 层进行训练, 测试最终模型的识别准确率等评价指标。

最后, 利用 Pyqt5 开发可用于辅助接口测试日志文本识别的工具, 用于实现高效能、智能化的接口测试。

3.1 获取模糊聚类中心

为处理铁路 CTC 系统服务器软件的日志文本文件, 完成服务器软件与其他终端软件的接口测试工作, 本节利用模糊文本聚类方法对调度集中系统的日志文本数据进行分析。文本模糊聚类过程如下。

(1) 选取服务器日志文本文件

本次选取应用服务器软件和通信服务器的日志 2024 年 3 月 12 日的文件, 该日的测试前期准备比较完备, 进行了列车计划下达、进路办理、列车接发车以及其他核心调度业务, 软件生成较多日志文本, 便于接口测试。

(2) 初始化文本

调用 jieba 库进行自然语言分词, 根据词频—逆文档频率 (term frequency-inverse document frequency, TF-IDF) 生成 TF-IDF 矩阵, 用于衡量一个词语对于一个文档集或语料库中某一份文档的重要程度, 便于后续进行模糊文本聚类。

①词频 (term frequency, TF): 词语 w 在文档 d 中的出现频率。其中, $\text{count}(d, w)$ 为单词 w 在文档 d 中出现的次数; $\text{count}(d, *)$ 为文档 d 的总词数。

$$\text{TF}(d, w) = \frac{\text{count}(d, w)}{\text{count}(d, *)} \quad (12)$$

②逆文档频率 (inverse document frequency, IDF): 若一个词语在多个文档中出现, 则意味着该词语的重要程度不高, 反之, 则重要程度高。其

中, $N(w)$ 为词语 w 在 N 个文档中出现的次数, N 为语料库中的文档总数。

$$\text{IDF}(w) = \log \frac{N}{N(w)} \quad (13)$$

(3) 模糊文本聚类

根据 FCM 算法, 依次构建隶属度矩阵、架构目标函数并更新聚类中心, 最终输出文本聚类后的聚类中心。以上算法利用 python 的 skfuzzy 库进行模糊文本聚类, 利用 numpy 库进行文本数据的处理, 最终根据日志文本的内容性质将其划分为中文自然语言、英文+数字、长文本数字和短文本数字 4 类, CTC 系统日志文本模糊聚类结果如图 5 所示, 主成分 1、主成分 2 是通过主成分分析 (principal components analysis, PCA) 方法提取的新特征。PCA 将高维数据投影到低维空间上, 保留数据方差信息, 由于数据中心化的数据特性, 数据均值均被中心化后趋于零, PCA 则根据数据的方差结构进行投影, 数据的主成分也因此集中在正方向, 即横轴纵轴均不出现负值。

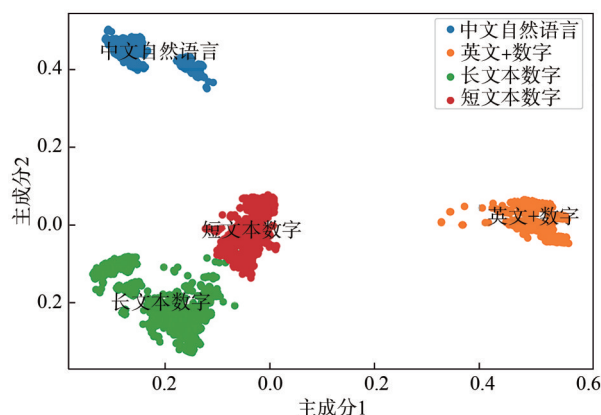


图5 CTC系统日志文本模糊聚类结果

在输出模糊文本聚类的基础上, 进一步统计文档高频命名实体的名称类别, 用于替代传统 NER 模型中的人名、地名、机构名等。结合 CTC 系统日志文本的命名实体, 将 4 类聚类结果进一步划分为 9 类命名实体标签, 见表 1。

3.2 CTC 接口测试文本的命名实体识别

本节将模糊文本聚类输出的 9 类命名实体标签作为训练的标注对象, 利用标注工具 LabelStudio 进行文本标注, 根据模糊文本聚类后的类别名称对训练数据进行类别标注, 标注采用 BMEO 规则 (B 为词首、M 为词中、E 为词尾、O 为单字), 在 LabelStudio 工具上编辑后自动导出 BMEO 后缀类别,

表1 9类命名实体标签

命名实体中文	命名实体英文	原文本样例
软件对象	NAME	中文语言描述
进路	RACE	中文语言描述
端口号	COUT	短数字:6 001
站场机构码	EDU	短数字:3 214
来源码	ORG	长数字:2 371-66-255
目的码	LOC	长数字:65 535-255-255
时间	TITLE	长数字:15:51:25
功能码	PRO	英文+数字:DeviceCode=1
命令码	ID	英文+数字:0xa7

比如 B-NAME、M-NAME、E-NAME 等标签。

训练的模型采用基于 BERT 和 BiLSTM-CRF 的铁路 CTC 系统服务器日志的命名实体识别模型, 模型的评价指标如下。

(1) 精度

精度又称查准率。针对判定结果, 预测为正例的样本中真正是正例的样本所占的比率, 其数值越高, 模型对“预测为正”的判断越可信。

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (14)$$

其中, TP 表示模型正确地将正例样本预测为正例样本的数量, FP 表示模型错误地将负例样本预测为正例样本的数量。

(2) 召回率

召回率为被正确判定的正例样本占总样本中正例样本的比率。召回率越高越好, 数值越高意味着模型对“实际为正”的样本的误判越少, 漏判的概率越低。

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (15)$$

其中, FN 表示模型预测结果为负例样本, 但实际为正例样本的数量。

(3) F1-score

F1-score 是精度和召回率的加权调和平均值。数值越接近 1, 模型查全率与查准率的均衡效果越好。

$$\text{F1-score} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{precision} + \text{recall}} \quad (16)$$

本次共训练验证了 Log2024-03-12.log 日志文件中的 15 162 条数据, precision 的平均值为 0.9599, Recall 的平均值为 0.9598, F1-score 的平均值为 0.9597, 9 类标签包含的所有 BMEO 评估指标如图 6 所示, 其中 3 项指标的平均数值超过了 0.95%, 3 项评估指标的个别标签最低值也高于 85%。因

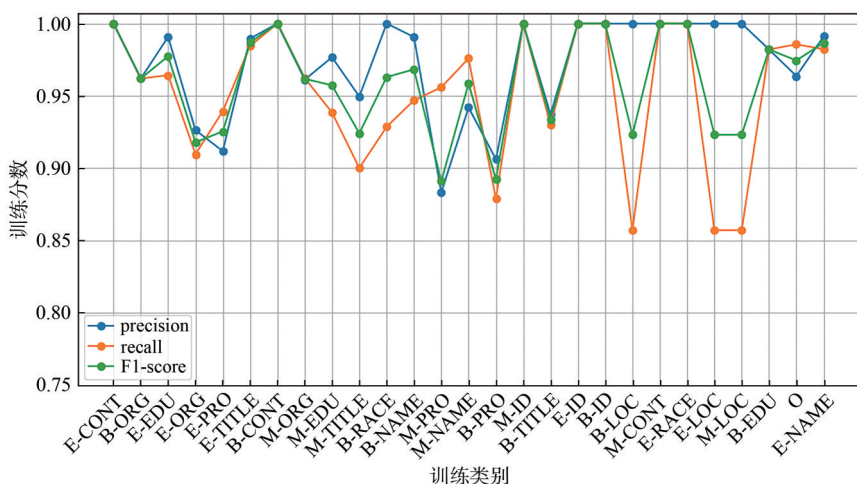


图6 模型评估指标效果

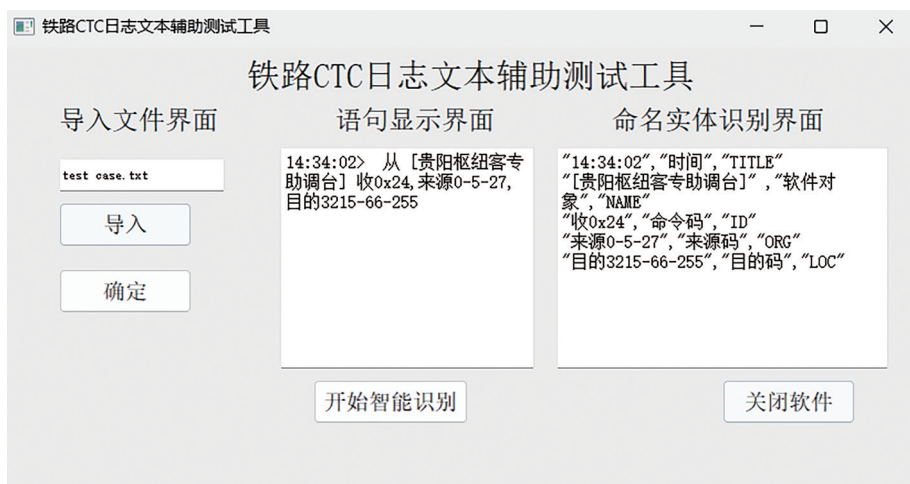


图7 铁路CTC系统日志文本辅助测试工具的界面

此, 模型具有较完善的训练效果。

在训练后的模型的基础上, 利用Pyqt5工具开发可视化软件, 该软件是可用于服务器接口测试中日志文本识别的辅助工具, 能够实现导入接口测试的日志文本、显示接口测试的日志文本语句定位以及智能识别接口测试日志文本的名称与内容等功能, 铁路CTC系统日志文本辅助测试工具的界面如图7所示。

4 结束语

CTC系统服务器软件生成的日志文本文件包含大量反映调度系统指令传输、功能发送、操作记录等运行状态的信息, 是系统投入实际使用前进行接口测试验证的重要依据。目前这一测试验证工作全部由人工完成, 存在易出错、效率低且工作量过大等问题。

本文为改善铁路CTC系统接口日志文本依赖人工测试的现状, 提出模糊自然语言处理的模型, 将模糊文本聚类方法与命名实体识别模型结合, 通过模糊C均值聚类算法生成高频文本类别, 并将其作为实体标签, 在传统NER任务的BiLSTM-CRF模型上引入BERT进行文本词嵌入与位置编码, 对系统的接口日志文本进行高效、准确的实体识别。基于此模型, 应用Pyqt5开发了一种用于铁路CTC系统日志文本接口测试的智能验证工具, 可辅助测试人员进行日志文本分析, 智能化、高效化地完成系统的接口测试工作。

参考文献:

- [1] 王飞跃, 缪青海, 张军平, 等. 探讨AI for Science的影响与意义: 现状与展望[J]. 智能科学与技术学报, 2023, 5(1): 1-6.
WANG F Y, MIAO Q H, ZHANG J P, et al. The DAOs to AI for science by DeSci: the state of the art and perspective[J]. Chinese Journal

- of Intelligent Science and Technology, 2023, 5(1): 1-6.
- [2] 王丽客, 孙媛, 刘思思. 基于多级注意力融合机制的藏文实体关系抽取[J]. 智能科学与技术学报, 2021, 3(4): 466-473.
WANG L K, SUN Y, LIU S S. Tibetan entity relation extraction based on multi-level attention fusion mechanism[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(4): 466-473.
 - [3] SHISHEHGARKHANEH M B, MOEHLER R C, FANG Y H, et al. Transformer-based named entity recognition in construction supply chain risk management in Australia[J]. IEEE Access, 2024, 12: 41829-41851.
 - [4] AKHTYAMOVA L, MARTÍNEZ P, VERSPOOR K, et al. Testing contextualized word embeddings to improve NER in Spanish clinical case narratives[J]. IEEE Access, 2020, 8: 164717-164726.
 - [5] ZHU P, CHENG D W, YANG F Z, et al. Improving Chinese named entity recognition by large-scale syntactic dependency graph[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 979-991.
 - [6] 黄炜, 黄建桥, 李岳峰. 基于BiLSTM-CRF的涉恐信息实体识别模型研究[J]. 情报杂志, 2019, 38(12): 149-156.
HUANG W, HUANG J Q, LI Y F. Research on entity identification model of terrorism-related information based on BiLSTM-CRF[J]. Journal of Intelligence, 2019, 38(12): 149-156.
 - [7] 于祥钦, 王香, 李智强, 等. 基于字符级特征自适应的生物医学命名实体识别[J]. 小型微型计算机系统, 2023, 44(9): 1876-1883.
YU X Q, WANG X, LI Z Q, et al. Biomedical named entity recognition based on character level feature adaptation[J]. Journal of Chinese Computer Systems, 2023, 44(9): 1876-1883.
 - [8] HAN X Y, WANG L. A novel document-level relation extraction method based on BERT and entity information[J]. IEEE Access, 2020, 8: 96912-96919.
 - [9] XU W, DAI X Y, ZHAO C, et al. Parallel testing for centralized traffic control systems of intelligent railways[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(9): 4249-4262.
 - [10] FENG G, WEI X. Research on function test method of intelligent high-speed railway centralized traffic control system[C]//Proceedings of the 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC). Piscataway: IEEE Press, 2021: 104-108.
 - [11] 刘昕, 席永轲, 何杰, 等. 基于多层次聚类的文本知识挖掘[J]. 计算机工程与设计, 2021, 42(1): 106-112.
LIU X, XI Y K, HE J, et al. Text knowledge mining based on multi-level clustering[J]. Computer Engineering and Design, 2021, 42(1): 106-112.
 - [12] 蔡莉, 王浩宇, 周君, 等. 一种改进的自适应网格划分的分布式聚类算法[J]. 小型微型计算机系统, 2023, 44(4): 731-736.
CAI L, WANG H Y, ZHOU J, et al. Improved distributed clustering algorithm for adaptive mesh partitioning[J]. Journal of Chinese Computer Systems, 2023, 44(4): 731-736.
 - [13] 刘雅萱, 武娇, 顾兴全, 等. 基于多语义文本表示的自适应模糊聚类算法[J]. 科学技术与工程, 2023, 23(11): 4709-4717.
LIU Y X, WU J, GU X Q, et al. Adaptive fuzzy clustering algorithm based on multi-semantic text representation[J]. Science Technology and Engineering, 2023, 23(11): 4709-4717.
 - [14] XU W, ZHANG N, WANG T. Design and realization of a TDCS/CTC-based train reception and departure operation control-management integrated system[C]//Proceedings of the 2017 Chinese Automation Congress (CAC). Piscataway: IEEE Press, 2017: 6311-6317.
 - [15] LI Z, XU W, ZHANG T, et al. Research and application of automated test for background program of railway CTC system[C]//Proceedings of the 2021 China Automation Congress(CAC). Piscataway: IEEE Press, 2021: 1975-1978.
 - [16] XU W. Design and implementation of FZy-CTC dispatching command release system for Han Huang line of Beijing Railway Bureau[C]//Proceedings of the 2016 IEEE International Conference on Service Operations and Logistics, and Informatics(SOLI). Piscataway: IEEE Press, 2016: 261-266.
 - [17] GUAN J Y, LI S, HE X X, et al. SMMP: a stable-membership-based auto-tuning multi-peak clustering algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 6307-6319.
 - [18] LIU Z, CAO J, LU Y, et al. A hierarchical clustering method of SOM based on DTW distance for variable-length seismic waveform[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.
 - [19] NASIBOV E N, ATILGAN C. A note on fuzzy joint points clustering methods for large datasets[J]. IEEE Transactions on Fuzzy Systems, 2016, 24(6): 1648-1653.
 - [20] NOORZADEH S, FAGHIHROOHI S, ZAREI M. A comparative analysis of HMM and CRF for early prediction of sepsis[C]//Proceedings of the 2019 Computing in Cardiology Conference(CinC). Piscataway: IEEE Press, 2019: 1-4.
 - [21] WANG Q, ZENG L. Chinese symptom component recognition via bi-directional LSTM-CRF[C]//Proceedings of the 2018 10th International Conference on Advanced Computational Intelligence(ICACI). Piscataway: IEEE Press, 2018: 45-50.
 - [22] WANG X G, PAN X B, CHEN C, et al. A named entity recognition model based on BERT model and lexical fusion in the financial regulation field[C]//Proceedings of the 2023 5th International Conference on Frontiers Technology of Information and Computer(ICFTIC). Piscataway: IEEE Press, 2023: 477-482.

[作者简介]



角远韬(2000—), 男, 北京交通大学自动化与智能学院硕士生, 主要研究方向为智能测试方法、铁路调度集中控制。



李润梅(1975—), 女, 博士, 北京交通大学自动化与智能学院副院长、教授, 主要研究方向为二型模糊集理论、无人车控制关键技术、交通大数据处理与预测。



王剑(1978—), 男, 博士, 北京交通大学自动化与智能学院党委副书记、副院长、教授, 主要研究方向为列车运行控制系统、铁路卫星导航、调度控制一体化。