# 融合BERT和阻塞过滤的国家电网公共数据模型实体映射技术

李雨霏[1],郝保聪[1],楼轶维[2*],杨诗语[1],高士杰[3],张鹏宇[1]

1. 国家电网有限公司大数据中心,北京 100053
2. 北京大学计算机学院,北京 100871
3. 北京中电普华信息技术有限公司,北京 100085

**摘要**　针对目前国家电网公共数据模型SG-CIM(state grid-common information model)难以实现自动更新迭代和挖掘新元素效率较低等问题,提出了一种基于知识图谱和BERT(bidirectional encoder representations from transformers)模型的SG-CIM模型自动映射技术。在现有SG-CIM模型的基础上,构建出SG-CIM知识图谱和数据表知识图谱;通过研究基于BERT模型和阻塞过滤的实体映射技术,在2个知识图谱之间建立映射关系;对文本方法映射效果进行实验分析,结果表明在自制数据集上微调后BERT模型的精确度在88%以上。

**关键词**　知识图谱;SG-CIM模型;BERT模型;阻塞过滤;实体对齐;实体映射

国家电网公司提出的标准数据统一模型——SG-CIM模型为电网领域的信息化交互和建模建立了统一的规范[1-2],SG-CIM模型有Logic模型和Physic模型两种数据存储模式,该数据模型主要由一级主题域、二级主题域、类及类的关系这3个层面构成,这些层面主要是依据电网处理信息的对象功能不同而建立[2]。由于目前SG-CIM模型不仅数量庞大,而且涉及门类极多,所以无法对所有数据模型进行全方位的定义,这就导致SG-CIM模型与数据表之间的映射定位技术还没有非常完善,极大影响了模型的适用性,因此研究一种统一数据模型映射技术,对于SG-CIM模型的进一步改进和发展有着重要意义。

知识图谱(knowledge graph,KG)作为组织和表达知识的重要工具和手段,在智能问答[3-5]、推荐系统[6]、垂直搜索等场景中有着广泛的应用,知识图谱旨在采用图的结构来建模和记录世界万物之间的关联关系和知识,以实现更加精准的对象级搜索[7]。因此构建出SG-CIM知识图谱和数据表知识图谱,并在两者之间建立映射关系,能够提高模型

收稿日期:2023-02-06;修回日期:2023-03-09

基金项目:国网大数据中心科技项目(SGSJ0000SJJS2200040)

作者简介:李雨霏,高级工程师,研究方向为大数据应用技术等,电子信箱:15101537383@126.com;楼轶维(通信作者),博士研究生,研究方向为大数据应用技术,电子信箱:cyfqylyw@gmail.com

引用格式:李雨霏, 郝保聪, 楼轶维, 等. 融合BERT和阻塞过滤的国家电网公共数据模型实体映射技术[J]. 科技导报, 2023, 41(15): 113-123; doi: 10.3981/j.issn.1000-7857.2023.15.012

本身的质量和迭代速度,提高对未收录新实体、新关系、新属性的查找和发掘的效率。

然而,已有研究成果主要面向电网以外的其他领域,直接在SG-CIM模型和数据表之间进行实体对齐的研究尚未见报道,因此研究一种适用于SG-CIM模型的自动映射技术的需求尤为迫切。

## 1 研究思路

在2个不同的知识图谱之间建立映射关系,实际上就是对2个KGs进行实体对齐。实体对齐也称为实体消岐或实体匹配[8],主要目的是判断来源于不同知识图谱的2个实体是否指向现实世界中的同一真实对象,其本质是解决多源知识库之间的异构问题。主要用到的技术有词嵌入和文本语义相似度计算。在文本相似度计算方面,诸多研究者提出了行之有效的文本相似度计算方案,文献[9]提出Word2Vec,通过该模型获取对应词向量后进行余弦相似度的计算,但该模型无法表征一词多义的特殊情形。文献[10]充分考虑到句子之间的交互信息,利用卷积神经网络[11](convolutional neural network,CNN)模型进行文本相似度计算,取得了较好的效果。文献[12]提出了基于LSTM(long short-term memory,LSTM)模型进行文本匹配的方法,该模型通过注意力机制计算单词间的相似度,不依赖于句子向量,但缺点是该模型无法进行并行计算,导致计算效率低下。2018年,Google在文献[13]中提出了BERT模型,该模型主要基于Transformer模型的双向编码器实现,进一步增强了词向量的泛化能力,能充分考虑词向量或句向量的上下文信息,也解决了一词多义问题并实现了并行计算。

实体对齐算法主要分为2个方向:传统的实体对齐算法和基于表示学习的实体对齐算法[14],传统的实体对齐算法主要是基于实体相似度理论来判断实体是否对齐,文献[15]提出使用TF-IDF(term frequency-inverse document frequency)算法,通过给实体间距离设置阈值的方式来获取对齐实体。文献[16]通过提出HistSim和DisNGram算法解决了异构数据源中的实体对齐问题,HistSim通过计算实体对相似度,修剪不对齐实体,DisNGram算法从字符级别出发,通过计算字符相似度来筛选对齐实体。文献[17]将实体对齐看做成是一个二分类问题,即匹配和不匹配,但实体对齐和二分类问题的关键区别在于:在实体对齐任务中,数据集中不匹配的数量要远高于匹配的数量。基于此,他们提出了结合阻塞过滤和主动学习(active learning,AL)[17]的实体对齐方法,该方法能够平衡匹配数据和不匹配数据的数量。文献[18]提出使用BERT模型来解决与模式无关的实体对齐问题,最终通过BERT分类器获得预测标签。

在相关工作的基础上,将知识图谱技术应用到电网领域,提出一种基于知识图谱和深度语言模型BERT的SG-CIM模型自动映射技术,从SG-CIM模型中选取2~3个模型、场景变化较大的典型主题域,研究该模型与数据表的映射技术。首先通过抽取Logic模型和Physic模型中的实体、属性和关系,构建出SG-CIM知识图谱和数据表知识图谱;然后,针对Word2Vec分词模型无法解决一词多义的特殊情形且有可能导致文本相似度计算结果不够准确等问题,并结合SG-CIM模型实际,使用BERT模型进行中文分词。由于SG-CIM模型数据量大,直接遍历所有数据文件会产生极大冗余,因此引入阻塞过滤机制进行数据清洗,并结合BERT模型提出一种实体文本映射技术,最终在其对应的两个知识图谱之间建立映射关系。

## 2 模型设计

### 2.1 总体设计

基于知识图谱和BERT深度语言模型设计的实体映射技术主要由知识图谱的构建和映射关系的建立2部分组成。主要处理过程如下:首先在SG-CIM的Logic模型和Physic模型的基础上,通过知识抽取、本体构建、知识存储等分别构建出SG-CIM知识图谱和数据表知识图谱,之后构建数据集微调BERT模型,研究基于阻塞过滤机制和BERT模型的实体映射技术,在2个知识图谱之间建立映射关系。主要技术路线如图1所示。

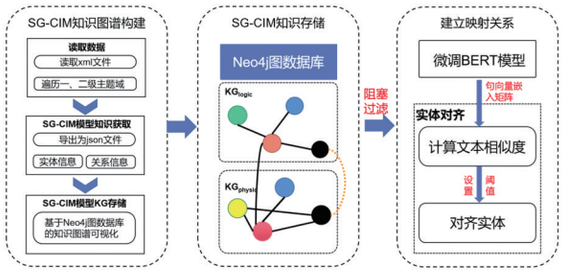



图1　本文方法技术路线

## 2.2　SG-CIM知识图谱和数据表知识图谱构建

构建知识图谱主要有2种方式,一种是自上而下构建,这种方法要求首先要有高质量数据集,之后从中抽取出本体和模式信息;另一种是自下而上构建,即从开放链接数据源或非结构化文本中提取出置信度较高的知识作为知识图谱构建的基础[19]。由于在SG-CIM模型知识图谱构建过程中,将从SG-CIM模型的xml文件中抽取出的一级主题域、二级主题域、实体相关描述等相关信息导出成为了格式严谨、质量较高的json数据集,且本体构建也并不复杂,因此SG-CIM知识图谱和数据表知识图谱适合选择自上而下的方式进行构建。

对SG-CIM知识图谱和数据表知识图谱的总体构建流程如图2所示,首先是遍历Logic模型和Physic模型中所有一级主题域和二级主题域,获得实体概念、规则和关系表示,抽取出其实体集和关系集,构建出Logic模型数据集和Physic模型数据集;然后根据构建好的数据集来完成知识图谱本体的构建;最后将整理后的数据导入到Neo4j图数据库中进行存储。

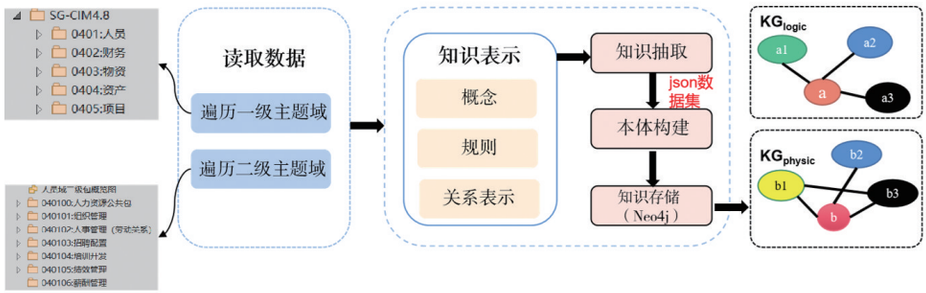



图2　SG-CIM知识图谱构建流程

### 2.2.1　构建数据集

SG-CIM模型主要以EAP文件的格式存放,可以使用Enterprise Architect软件打开该类格式文件。在进行数据读取之前,首先要将EAP文件转换成xml文件格式,然后遍历xml文件的树形结构找到实体名称、关系、属性的存放路径,之后抽取出实体中英文名称、实体所处一二级主题域、实体间关系及属性描述信息等,并对抽取出的数据进行规范化处理,将实体及其属性信息整理成一个json串,进而得到实体集,将实体及其关系整理得到关系集;最后将处理完成的数据存储到JSON文件中。构建出的Logic模型和Physic模型实体集中的单个实体由实体编号id、实体所处一级主题域area1、实体所处二级主题域area2、实体中文名称ch、实体英文名称en、实体相关描述des、实体属性信息attr等构成。

### 2.2.2　本体构建(ontology building)

本体构建[20]主要有人工构建和采取自动化、半自动化构建2种方式,但前者需要耗费大量人工,且构建效率低下,因此SG-CIM逻辑模型和物理模型的构建采用自动、半自动化方式。

采用自动、半自动化方式构建本体由预处理、术语抽取、概念抽取、关系抽取和本体形成等子任务组成,以SG-CIM的Logic模型本体构建为例,总结本体构建步骤如下:(1)首先进行数据预处理,对输入数据进行术语和概念的抽取;(2)对输入数据进行电网领域特定关系的抽取;(3)将抽取的概念、关系进行归纳总结,形成目标本体。SG-CIM模型的实体本体定义如表1所示。

### 2.2.3　知识存储

Neo4j图数据库非常适用于基于图结构的知识图谱的存储[19],使用Python提供的py2neo工具包来

表1 SG-CIM模型的实体本体定义

| 属性名称 | 属性值 |
|---|---|
| name | [实体中,英文名称] |
| area1 | 实体所处一级主题域 |
| area2 | 实体所处二级主题域 |
| des | 实体相关描述 |
| attr | [属性列表] |

完成SG-CIM知识图谱和数据表知识图谱中节点和关系的创建,使用Cypher语句对某个特定节点进行增删改查操作。本研究构建的SG-CIM知识图谱和数据表知识图谱中实体和关系种类可概括为两类节点和两类关系,其中节点分为实体节点和属性节点,关系分为实体与属性间关系和实体与实体间关系,节点和关系数量如表2所示。

表2 节点和关系数量

| 模型 | 实体节点 | 属性节点 | 实体-属性关系 | 实体-实体关系 |
|---|---|---|---|---|
| Logic模型 | 6759 | 87587 | 87520 | 3886 |
| Physic模型 | 4796 | 115085 | 169722 | 2348 |

构建出的部分知识图谱实体节点及其关系如图3所示。

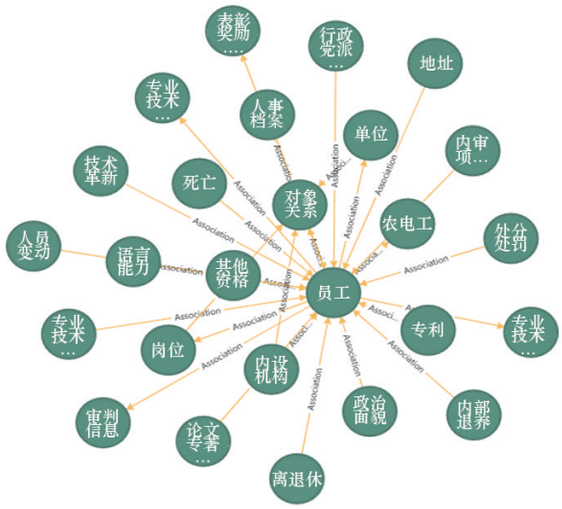



图3 SG-CIM知识图谱部分节点和关系

## 3 混合模型理论基础

### 3.1 BERT预训练语言模型

BERT模型的基本结构如图4所示。

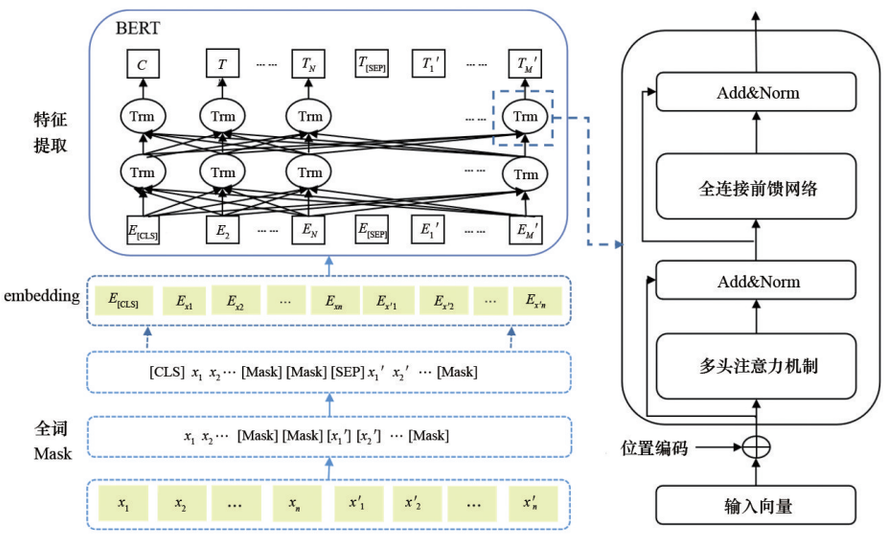



图4 BERT模型结构

BERT模型由屏蔽语言建模(masked language model,MLM)和下一句预测(next sentence prediction,NSP)2个无监督任务进行联合预训练[19]。在前期处理句向量时,会首先以字符为单位把输入序列切分成若干子词,如图4中$x_1$、$x_2$、$x'_n$所示,之后会随机选择15%的子词进行屏蔽(Mask)。考虑到所用语料均为中文,按照中文分词习惯,选择了全词Mask[21]的方法,在中文预训练任务中,如果一个完整词的部分子词被随机Mask,那么整个词都将被Mask掉。具体Mask过程如表3所示。

表3　BERT模型全词Mask

| 类型 | 文本 |
|---|---|
| 原始文本 | 青岛是一座美丽的城市。 |
| 分词文本 | 青岛 是 一座 美丽 的 城市 。 |
| 原始Mask输入 | 青[Mask] 是 一座 美[Mask] 的 [Mask] 市 。 |
| 全词Mask输入 | [Mask][Mask] 是 一座 [Mask][Mask] 的 [Mask][Mask] 。 |

对于经过全词Mask后的输入序列,会增加[CLS]和[SEP]标记,[CLS]用于分类任务,表示句子的开始,只在输入序列的头部进行显示,[SEP]表示句子的结尾,用来区分前后2个句子[22]。之后再经过embedding层获取词嵌入$E_{x1}$、$E_{x2}$、$E_{x'n}$等,经过Trm即Transformer编码器处理之后,完成特征提取和BERT模型的微调,之后即可将BERT模型应用到命名实体识别、文本分类、文本相似度计算等任务中。

使用BERT模型计算文本相似度的一大优势在于BERT模型内叠加了若干Transformer编码器,而Transformer的多头注意力(multi-head attention)部分,可以使模型在不同的表示子空间里学习到相关信息,能充分考虑其他实体对当前实体的影响,提高对新实体挖掘的效率。其框架结构如图5所示。

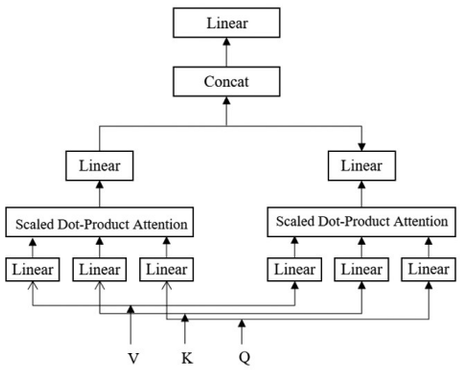



图5　多头注意力机制结构

多头注意力机制可表达为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

$$head_i=\mathrm{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}\right) \tag{2}$$

$$\mathrm{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Concat}\left(head_1,head_2,\cdots,head_h\right)\boldsymbol{W}^{o} \tag{3}$$

式中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为对原始输入序列经过线性变换后得到的词向量;$\boldsymbol{W}_i^{Q}$,$\boldsymbol{W}_i^{K}$,$\boldsymbol{W}_i^{V}$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的权重矩阵;i为多头注意力编号,取1~$h$,$h$为多头注意力头数;dk为K的维度。令$\boldsymbol{Q}$=$\boldsymbol{K}$=$\boldsymbol{V}$即为自注意力机制[23]。

除了注意力子层外,每一个编码器和解码器中还包含一个全连接前馈网络(FFN)[24],该层由2个线性变换组成,中间是一个ReLU(rectified linear unit)激活函数,全连接前馈网络可以表示为:

$$FFN(\boldsymbol{x})=\max\left(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{4}$$

式中,$\boldsymbol{x}$为多头注意力机制的输出;$\boldsymbol{b}$为偏置向量。

### 3.2　基于语义的实体映射模型

用$G_1$=($E_1$, $R_1$, $A_1$, $T_1$),$G_2$=($E_2$, $R_2$, $A_2$, $T_2$)表示两个知识图谱[25],其中$E_i$、$R_i$、$A_i$、$T_i$($i$=1,2)分别对应知识图谱中的实体集、关系集、属性集和事实三元组,可以将实体对齐的形式化定义表示为

$$\mathrm{Align}\left(G_1,G_2\right)=\left\{\left(E_1\in G_1,E_2\in G_2,\mathrm{sim}\in[0,1]\right)\right\} \tag{5}$$

式中,sim表示对实体对相似性的度量,sim越大表示两个实体越可能是等价实体对。

实体对齐的目的[25]可以被形式化为$S=\left\{e_1,e_2\in E_1\times E_2\middle|e_1\leftrightarrow e_2\right\}$,$\leftrightarrow$表示来自$G_1$的实体$e_1$和来自$G_2$的实体$e_2$有相同的语义关系,($e_1$, $e_2$)是一个等价实体对,即$S$为一个等价实体集合。

SG-CIM知识图谱和数据表知识图谱中存在大量表述不同但指向同一对象的实体,研究一种实体对齐算法,是在2个知识图谱之间建立映射关系的关键。本文实体对齐采用二分类的思想,首先利用阻塞过滤方式,筛选出候选实体对,然后用BERT模型分词代替传统的Word2Vec的分词方法获取文本的句向量嵌入矩阵,之后利用式(6)计算矩阵的余弦距离

$$\mathrm{similarity}=\cos\left(A,B\right)=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}\left(A_i\right)^2}\sqrt{\sum_{i=1}^{n}\left(B_i\right)^2}} \tag{6}$$

实体对齐框架如图6所示。

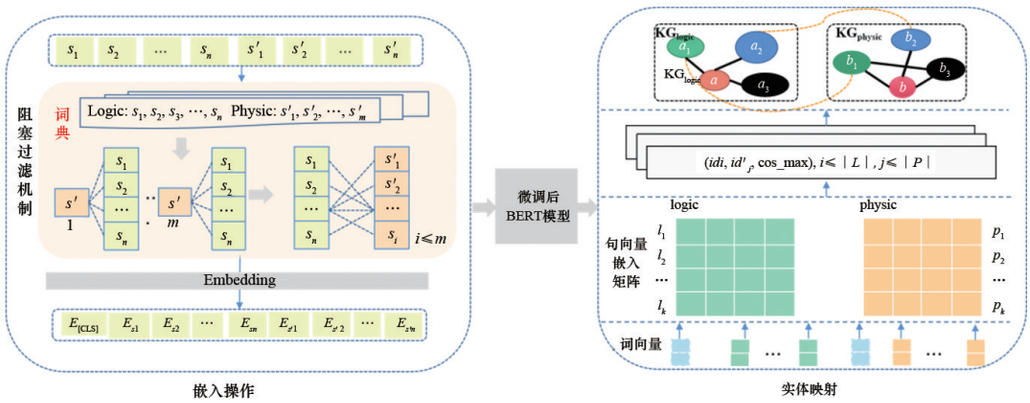



图6 实体对齐框架

由于SG-CIM知识图谱和数据表知识图谱中节点数量不一致,在表现形式上,2个知识图谱是异构的,所以一般基于模式固定的实体对齐方法可能无法应用于本文的知识图谱实体对齐[18]。因此在考虑SG-CIM模型结构特点的基础上,提出结合阻塞过滤及BERT深度语言模型进行实体对齐的方法。本文方法较之前的研究成果有以下3点不同:(1)输入格式可以有多种,如JSON文件以及XML文件等,最终都可以实现SG-CIM模型和应用数据表间映射关系的建立;(2)本文使用的阻塞过滤算法为构建词典筛选SG-CIM知识图谱和数据表知识图谱中名称相同的实体作为预对齐实体,也即在对数据进行预处理时,过滤掉那些明显与现有实体节点无法匹配的实体,比如数据表知识图谱中的"项目域"及其相关节点与SG-CIM知识图谱中任何实体节点都不匹配,那么就先过滤掉该"项目域"及相关节点。经过这一步的处理,会过滤掉大量冗余节点,减少算法的时间复杂度;(3)将实体对齐转化成二分类问题,输出结果分为相似和不相似2种情况,对应的二分类函数为

$$\mathscr{B}: E_1 \times E_2 \rightarrow \{\text{Similar, Not Similar}\} \quad (7)$$

实体对齐算法如算法1所示,解释如下。

1)对知识图谱进行数据预处理,即从数据源中抽取出实体节点相关信息构成json格式数据集,根据实体名称分别从两个数据集中获取部分实体相关描述一致的文本对,并将标签label设置为1,表示相似,将部分描述不一致的文本随机组合构成不相似文本对,标签label设置为0。

算法1 实体对齐算法

| 算法1 EntityMappingBERT() |
|---|
| 输入:多数据源待对齐实体集合MultiS-UnalignedSet |
| 输出:0或1 |
| for 实体对EP=(l,p) in MultiS-UnalignedSet |
| if ∃ l.name = p.name then |
| T=Triplet(*l*, *p*, label=1) //实体名称相label=1 |
| else |
| random(l,p)//随机分配实体对 |
| T= Triplet(*l*, *p*, label=0) //实体名称不相似label=0 |
| Finetune-BERT(T) |
| if ∀p.name≠l.name then |
| delete p //过滤掉不相关实体 |
| A= BERT(*l*) //BERT获取句向量 |
| B= BERT(*p*) |
| Similarity = cos(A,B) //计算余弦相似度 |
| if Similarity > 相似度阈值Θ then |
| 返回1 |
| else |
| 返回0 |
| end for |

2)将标记好的文本对以三元组($l_i$, $p_j$, label)的形式输入至BERT模型进行微调训练,如图7所示。其中$(l_i, p_j) \in L \times P$, $i \leqslant |L|$, $j \leqslant |P|$, $l_i$表示SG-CIM知识图谱中实体的相关描述,$p_j$为数据表知识图谱中实体相关描述,$L$为SG-CIM知识图谱的实体集,$P$为数据表知识图谱中的实体集。

3)分别按顺序从SG-CIM知识图谱和数据表知识图谱的json格式数据集中抽取出每一实体节

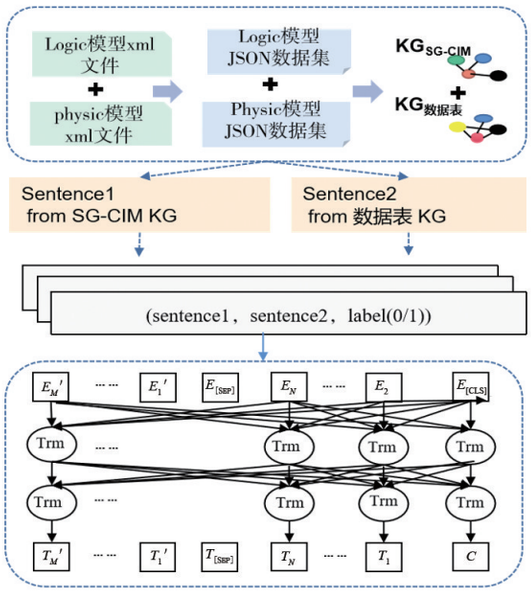



图7　微调BERT模型

点的文本描述,利用阻塞过滤机制清洗掉不相关实体后得到输入序列,将其输入至微调后的BERT模型获取句向量嵌入矩阵。

4)利用余弦距离计算出相似度最大的文本对,同时分别获取文本在2个json格式数据集中的id并输出为三元组($id_i$, $id'_j$, cos_max)的格式,cos_max是两文本之间的余弦相似度最大值。

5)根据相似度分布情况设置阈值,若相似度大于阈值,则表明其对应的两实体对相似,可以进行对齐,若相似度小于阈值,则说明该实体对不相似,即无法进行对齐,实体对齐完成后,文本映射关系建立完成。

# 4　实验分析

## 4.1　数据集

实验采用的数据集是从SG-CIM的Logic模型和Physic模型中抽取出来的实体相关描述,在该数据集中,从2个模型中分别抽取出一句有关单个实体的描述组成文本对,将部分文本对语义相同或相似描述标记为正例,标签标记为1,将语义不同的部分文本对标记为0,视为负例。数据集特征如表4所示。

表4　数据集描述

| 模型 | 实体数量 | 关系数量 | 训练数据 | 测试数据 | 验证数据 |
|---|---|---|---|---|---|
| Logic模型 | 6759 | 3802 | 7000 | 2000 | 1000 |
| Physic模型 | 4976 | 2348 | 7000 | 2000 | 1000 |

## 4.2　实验环境与参数配置

实验基于Tensorflow搭建,使用的操作系统为Ubuntu 20.04,CPU为Intel(R) Xeon(R) Silver 4216 CPU@2.10GHz,GPU为NVIDIA GeForce RTX A2000,python为3.7版本,Tensorflow使用1.15版本。

实验所使用的模型为BERT模型,选取的是12层的Transformer,学习率设置为$2\times10^{-5}$,max_seq_len取值为128,batch_size取值为32,dropout取值为0.5。

## 4.3　评估指标

本文采用的评估指标为精确度$P$、召回率$R$和$F_1$值来评估本文设计方法的有效性,各评估指标计算方法为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{8}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{9}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{10}$$

式中,TP为正例预测为正例的实体数;FN为将正例识别为负例的实体数;FP为将负例识别为正例的实体数。

## 4.4　实验结果分析

利用微调后BERT模型获得文本描述的句向量,并通过实体对齐方法筛选出了相似度最大文本对,最终在2个知识图谱之间建立了映射关系,文本映射的散点图如图8所示。

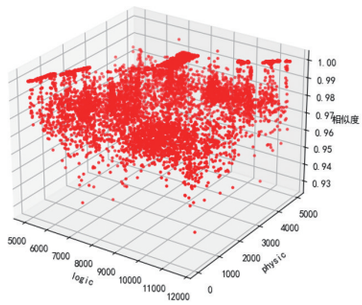
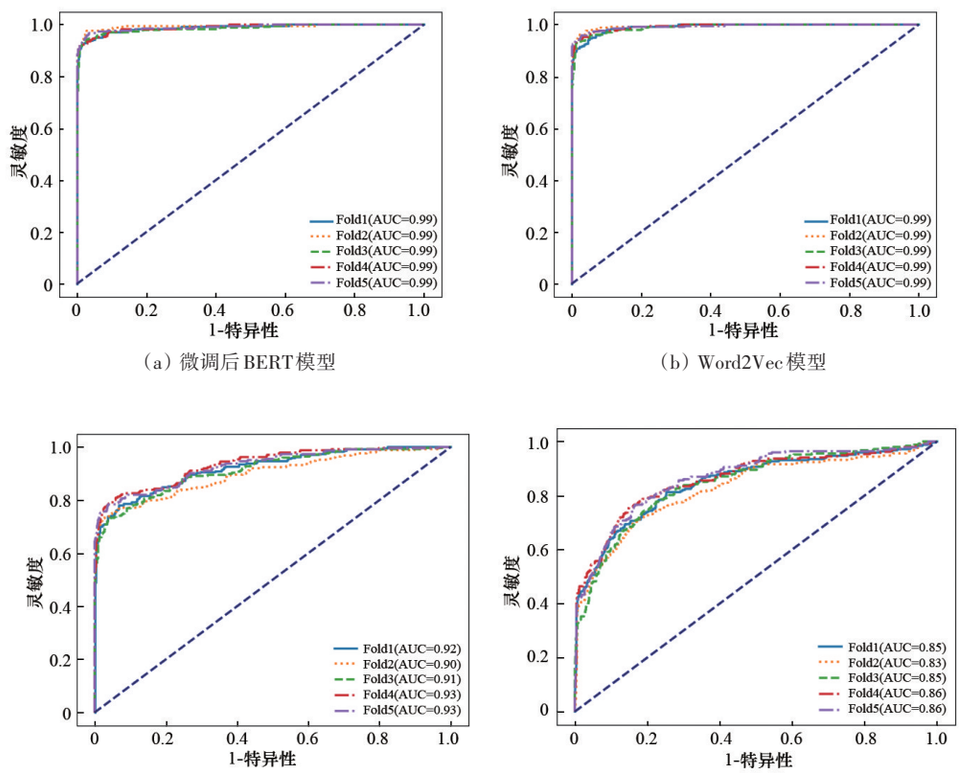



图8　相似度最大文本对映射关系

图8展示的是所有相似度最大文本对,即映射关系的散点图,其中logic表示SG-CIM知识图谱中节点对应的id,physic表示数据表知识图谱中节点对应的id,“相似度”轴表示文本对相似度最大值。以点(4977,7,1)为例,在SG-CIM知识图谱中,id为4977的节点对应中文名称ch为“员工”,所处一级主题域area1为“人员”,所处二级主题域area2为“人力资源公众包”,在数据表知识图谱中,id为7的实体节点中文名称ch为“员工”,实体所处一级主题域area1为“人员”,二级主题域area2为“人力资源公共包”,2个节点的相似度为1。

为了验证本文所用模型的有效性,分别将微调后BERT模型与Word2Vec、LSI(latent semantic indexing)模型及LDA(latent dirichlet allocation)模型进行5折交叉验证和对比实验。

AUC(area under curve)是ROC[26](receiver operating characteristic)曲线的线下面积,是评估模型性能的重要指标,一般取为0.5~1.0,越接近于1.0说明模型效果越理想。利用测试集绘制出微调后模型及其他对比模型的ROC曲线如图9所示,其中Fold1~Fold5分别对应第1~5折交叉验证的ROC曲线,根据AUC值可以判断出微调后的BERT模型性能优于LDA、LSI模型。

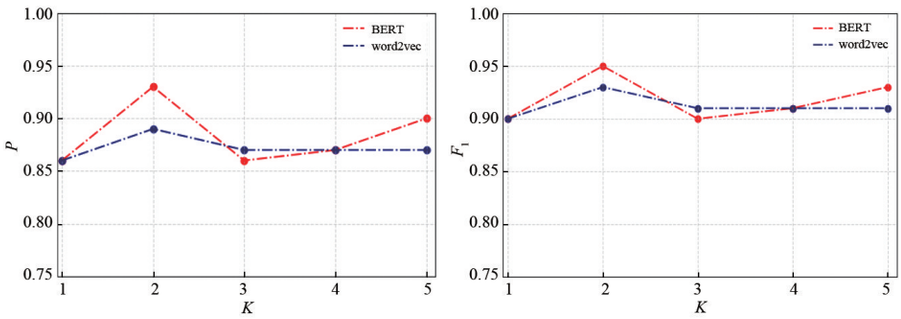



(a) 微调后BERT模型 (b) Word2Vec模型

(c) LSI模型 (d) LDA模型

图9 各模型的ROC曲线

此外,由于BERT模型和Word2Vec模型AUC值无明显差异,为了对比微调后BERT模型和Word2Vec模型的性能,又绘制出了两者的精确度$P$、$F_1$的对比图,如图10所示。横坐标表示$K$折交叉验证中$K$的取值,$K$折交叉验证指的是将原始数据集划分成$K$个大小相等的子集,然后依次取其中一个子集用作验证集,其余$K$-1个子集用作训练集,最后把$K$次验证结果的平均值作为最终模型性能评估指标,由于本文对两模型使用5折交叉验证进行评估,所以横坐标$K$的取值为1~5。




(a) $P$值　(b) $F_1$值

图10　BERT、Word2Vec精确度$P$、$F_1$值对比

BERT的精确度和召回率分别为0.884、0.92，Word2Vec模型的精确度和召回率分别为0.872、0.91，证明本文所使用的微调后BERT模型的性能要优于Word2Vec。将微调后BERT模型的阈值设置为0.954，将大于阈值的文本对判断为相似，小于阈值的文本对判断为不相似，得到测试集预测标签predict_label，如表5所示。

实验证明，在自制数据集上利用BERT模型和基于阻塞过滤的实体映射技术在2个知识图谱之间建立映射关系，使实验取得了比较好的效果。

表5　部分文本对相似度计算结果及其预测标签

| Logic模型数据 | Physic模型数据 | 相似度 | 真实标签 | 预测标签 |
|---|---|---|---|---|
| 描述分包商资质； | 记录分包商资质； | 0.943 | 1 | 0 |
| 合同评价星级类型 | 技术规范物料编码关联关系。 | 0.905 | 0 | 0 |
| 描述正式试卷的基本信息； | 用于记录正式试卷的基本信息； | 0.988 | 1 | 1 |
| 检修计划审核信息 | 技改大修检修设备信息； | 0.963 | 0 | 1 |
| 竞价记录用来保存回收商竞价信息。 | 竞价记录将记录回收商的竞价信息 | 0.960 | 1 | 1 |

## 5　结论

在现有SG-CIM模型的基础上，提出了一种基于知识图谱和BERT的实体映射技术。针对传统的分词模型无法解决一词多义的问题，使用微调后的BERT模型获取句向量嵌入矩阵，有效解决了一词多义的特殊情形。对于不同来源且数据量较大的实体对齐问题，引入了阻塞过滤机制，最终通过计算文本对之间的余弦距离，在SG-CIM知识图谱和数据表知识图谱之间建立了映射关系，并通过实验对比，验证了本文所提方法的有效性。本文方法为后续SG-CIM模型的新实体、新属性和新关系的发掘奠定了基础。

## 参考文献(References)

[1] 杨帅. 基于SG-CIM的配电网生产管理系统的研究与应用[D]. 北京: 华北电力大学, 2018.

[2] 徐尧强, 舒乔晔, 黄昭, 等. 基于公共信息模型的电力项目管理模型设计[J]. 能源工程, 2021(4): 76-80.

[3] HAO YI M,WU Y,CHEN L,et al. Intelligent question answering system based on domain knowledge graph[P]. 2022 3rd International Conference on Artificial Intelligence and Education: IC-ICAIE 2022, 2022.

[4] ZHOU H J, SHEN T T, LIU X L, et al. Survey of knowledge graph approaches and applications[J]. Journal on Artificial Intelligence, 2020, 2(2): 89-101.

[5] 曲克童. 基于深度迁移学习的电力知识图谱智能问答[D]. 北京: 华北电力大学(北京), 2022.

[6] Sajisha P S, Anoop V S, Ansal K A. Knowledge graph-based recommendation systems: The state-of-the-art and some future directions[J]. International Journal of Machine Learning and Networked Collaborative Engineering, 2019, 3(3): 159-167.

[7] 陈烨, 周刚, 卢记仓. 多模态知识图谱构建与应用研究综述[J]. 计算机应用研究, 2021, 38(12): 3535-3543.

[8] 闻涛. 面向知识图谱的实体对齐和知识补全[D]. 杭州: 杭州电子科技大学, 2019.

[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. New York: ACM, 2013: 3111-3119.

[10] Hu B T, Lu Z D, Li H, et al. Convolutional neural network architectures for matching natural language sentences[DB/OL]. arXiv Preprint: 1503.03244, 2015.

[11] Yoon K. Convolutional neural networks for sentence classification. 2014[DB/OL]. arXiv Preprint: CL/1408.5852, 2014.

[12] Wang S H, Jiang J. Learning natural language inference with LSTM[DB/OL]. arXiv Preprint: 1512.08849, 2015.

[13] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[DB/OL]. arXiv Preprint: 1810.04805, 2018.

[14] 张富, 杨琳艳, 李健伟, 等. 实体对齐研究综述[J]. 计算机学报, 2022, 45(6): 1195-1225.

[15] Cohen W W, Richman J. Learning to match and cluster large high-dimensional data sets for data integration[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 475-480.

[16] Song D Z, Luo Y, Heflin J. Linking heterogeneous data in the semantic web using scalable and domain-independent candidate selection[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 143-156.

[17] Arasu A, Götz M, Kaushik R. On active learning of record matching packages[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. New York: ACM, 2010: 783-794.

[18] Teong K S, Soon L K, Su T T. Schema-agnostic entity matching using pre-trained language models[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 2241-2244.

[19] 李家瑞, 李华昱, 闫阳. 面向多源异质数据源的学科知识图谱构建方法[J]. 计算机系统应用, 2021, 30(10): 59-67.

[20] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[21] Cui Y M, Che W X, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[22] 谢腾, 杨俊安, 刘辉. 基于BERT-BiLSTM-CRF模型的中文实体识别[J]. 计算机系统应用, 2020, 29(7): 48-55.

[23] 杨晨. 基于神经网络的短文本语义相似度计算方法研究[D]. 成都: 电子科技大学, 2020.

[24] 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. New York: ACM, 2017: 6000-6010.

[25] Zeng K S, Li C J, Hou L, et al. A comprehensive survey of entity alignment for knowledge graphs[J]. AI Open, 2021, 2: 1-13.

[26] Carrington A M, Manuel D G, Fieguth P W, et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 329-341.

# An entity mapping technology of national grid public data model integrating BERT and congestion filtering

LI Yufei[1], HAO Baocong[1], LOU Yiwei[2*], YANG Shiyu[1], GAO Shijie[3], ZHANG Pengyu[1]

1. Big Data Center of State Grid Corporation of China, Beijing 100053, China
2. School of Computer Science, Peking University, Beijing 100871, China
3. Beijing Zhongdian Puhua Information Technology Co., Ltd., Beijing 100085, China

**Abstract** Aiming at the problems of current SG-CIM (state grid-common information model) such as difficult to achieve automatic update iteration and low efficient mining of new elements, an SG-CIM model automatic mapping technology based on BERT model and blocking filtering is proposed. On the basis of the existing SG-CIM, an SG-CIM knowledge map and data table knowledge graph are constructed at first. Secondly, by studying the entity alignment method based on BERT model and blocking filtering, the mapping relationship between the two knowledge graphs is established. Finally, the effectiveness of the proposed method is verified by experimental analysis of the text mapping effect. Results show that the accuracy of BERT model after fine-tuning on a self-made data set is more than 95%. This method lays a foundation for subsequent mining of new elements and automatic updating iteration of SG-CIM.

**Keywords** Knowledge graph; SG-CIM model; BERT model; entity alignment; Establish Mapping

（责任编辑　刘志远）