第54卷增刊
2022年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 54 No. S
Oct. 2022

DOI:10.16356/j.1005-2615.2022.S.017

基于BERT的航天术语标准化

刘栋梁，张　嵩，张宁康，高　洋，林海波

（中国航天标准化研究所，北京 100071）

摘要：将航天术语标准化问题与深度学习相结合，发挥深度学习在文本建模上强大的语义表征能力，提出了一种基于BERT的航天术语标准化方法。首先，介绍了自然语言处理任务中的文本相似度计算方法和序列标准方法。然后，针对一定量的航天文本原词数据，使用Jaccard相似度计算得到候选标准词。最后，通过构建BERT的预训练模型获得预测标准词。实验结果表明，BERT在自然语言处理任务上具有强大优势，基于多个术语标准评测数据集上准确率达到89.93%，可以提高航天术语标准化水平和效果。

关键词：航天术语；标准化；深度学习；BERT；语义相似度

中图分类号：TP391.1　　**文献标志码：**A　　**文章编号：**1005-2615(2022)S-0109-06

Standardization of Aerospace Terms Based on BERT

LIU Dongliang, *ZHANG Song*, *ZHANG Ningkang*, *GAO Yang*, *LIN Haibo*

（China Aerospace Standardization Institute, Beijing 100071, China）

Abstract: Combining the aerospace terminology standardization with deep learning, and exerting the powerful semantic representation ability of deep learning in text modeling, we propose a BERT based aerospace terminology standardization method. First, the text similarity calculation method and the sequence standard method for natural language processing tasks are introduced. Second, for a certain amount of space text original word data, the candidate standard words are obtained by using Jaccard similarity calculation. Finally, the prediction standard words are obtained by constructing the pre-training model of BERT. The experimental results show that BERT has strong advantages in natural language processing tasks, and the accuracy rate on the evaluation data set based on multiple terminology standards reaches 89.93%. The proposed method can improve the standardization level and effect of aerospace terms.

Key words: aerospace terminology; standardization; deep learning; BERT; semantic similarity

近年来，中国航天领域各项重点工程加快开展，自主创新能力不断增强，整体实力跨入世界先进行列。2021年全球共146次航天发射中，中国占55次，居世界第一。“长征”火箭全年48次发射全胜，空间站建设五战五捷，“北斗三号”全球卫星导航系统顺利开通运行，“天问一号”再次拓展中国星际探索的新旅程。航天事业的不断发展离不开航天标准化行业的强力支撑。2021年10月，中共中央、国务院印发《国家标准化发展纲要》[1]阐述了中国标准化发展的重要使命和美好愿景。标准在保障航天各项工程顺利完成、开拓新技术领域、引导产业规模化发展、促进国际合作等方面发挥着不可替代的作用[2]。

航天术语标准化是整个航天标准化工作的重要组成部分，是航天技术交流与合作中统一概念的关键。但是由于航天产业的不断发展、航天队伍的不断壮大，标准宣传贯彻落实不到位，加之航天工作者个人书写习惯和术语表达多样性等因素，导致航天工作者在日常工作中形成的技术报告等文字材料中常常对同一概念会有不同的表述形式，因

收稿日期：2022-05-15；**修订日期：**2022-06-30

通信作者：刘栋梁，男，工程师，E-mail：907644846@qq.com。

引用格式：刘栋梁，张嵩，张宁康，等. 基于BERT的航天术语标准化[J]. 南京航空航天大学学报，2022，54(S)：109-114.
LIU Dongliang, ZHANG Song, ZHANG Ningkang, et al. Standardization of aerospace terms based on BERT[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(S): 109-114.

此,研究一种自动的航天术语标准化方法对于航天信息化建设和质量控制具有重要的现实意义。

本文将航天术语标准化问题与深度学习相结合,发挥深度学习在文本建模上强大的语义表征能力,通过构建BERT模型,提出了一种基于BERT的航天术语标准化方法。首先介绍了自然语言处理任务中的文本相似度计算方法和序列标准方法,针对一定量的航天文本原词数据,通过对比欧式距离法等4种不同计算方法下待标准化的原词与标准词间的相似度,对标准词进行赋分排序得到数个候选标准词。然后构建BERT的预训练模型,将航天语义原词和候选标准词相匹配,把航天语义原词和候选标准词的相似度求解问题转化为二分类问题,从而获得预测标准词。通过实验可以表明,BERT在自然语言处理任务上具有强大优势,可以提高航天术语标准化水平和效果。

1 相关方法

针对航天术语标准化的问题,早在1999年查朝晖等[3]就进行了相关研究和探讨,强化了术语和术语标准化的作用与地位,分析了当时中国术语标准化、尤其是航天术语标准化的现状,并分析了航天术语标准化工作中存在的航天工作者对术语标准化工作思想认识的不统一、术语标准体系不完善、内容重复交叉、缺乏统一协调管理等主要问题,对航天术语标准化工作提出了建设性提议。

此后历时二十多年的发展,航天术语标准化工作发展有了很大的跨越,多个术语的国家标准、国家军用标准、行业标准出台,大力推动了航天行业术语的规范化、统一化、体系化。随着信息技术的发展,术语标准化的研究与探索愈加丰富,尤其是与信息化相结合领域的探索。张嵩等[4]提出了一种基于语义相似度来计算航天标准相关性的评价方法,即通过计算语义相似度来度量航天领域标准名称、范围及主要内容的相关性,算例分析结果表明该方法可以有效评估标准间的关联程度;王洪东等[5]通过分析语言值有序对三元组及其性质,得到了语言值有序对三元组之间的相似度;杜迎雪等[6]基于最大长度区间思想,利用标准化区间权重向量的判定方法提出了一种区间权重向量标准化方法,并证明了该方法的有效性;崇伟峰等[7]通过数据预处理等四个模块实现了基于BERT蕴含分数排序的临床术语标准化系统并取得了很好地术语标准化效果。

通过对不同术语标准化相关研究的分析,可以发现不同的术语标准化方法都各有所长和不足。人工手动相结合的方式虽然匹配性高,精度更加准确,但是人工成本、时间成本大,工作效率低;机器学习的方法可以实现快速的捕捉语义信息,完成语义匹配,方式简单高效,但无法捕捉文本中深层的语义含义和信息,缺乏足够深度的表征能力;深度学习的方法通过构建适用的模型,可以较为完整的得到文本内的语义信息,较好的表述相关词语和含义,在文本分类、语义相似领域具有强大的表征能力。因此,本文主要通过深度学习的方法,构建BERT模型,通过航天术语文本相似度计算实验证明此方法的有效性。

2 基于BERT的术语标准化

2.1 文本相似度计算方法

在自然语言处理任务中,经常需要计算文本之间的相似度。比如,在对语料进行预处理时需要根据文本相似度将重复文本标注并删除。

在计算文本相似度时需要两个重要模块,分别是文本表示模型和相似度度量方法。文本表示模型将文本转换为计算机可识别的数值向量,相似度度量方法则根据数值向量计算文本之间的相似度。通过选取恰当的文本表示模型和相似度度量方法,即可构建出所需的文本相似度计算方案。

2.1.1 文本切分

n-gram是一种基于统计语言模型的算法,即对文本所有字符进行等量切分。“n”表示使用一个字符数为n的文本框,文本框逐字符地框过文本,每框到一个长度为n字符串就切分为一个gram,从而将文本中所有字符进行切分形成gram列表。

表1 n-gram文本切分示例

Table 1 n-gram text segmentation example

序号	n	名称	切分示例
1	1	Unigram	北/斗/导/航/用/户/设/备
2	2	Bigram	北斗/斗导/导航/航用/用户/户设/设备
3	3	Trigram	北斗导/斗导航/导航用/航用户/用户设/户设备
⋮	⋮	⋮	⋮

分词也是一种文本切分方式,它是将文本切分为一个个有句法意义的小单元。文字文本是序列数据,文字中的固定搭配,如词组或诗词,字符与字符之间会呈现出较高的相关度。由于分词切分法是在字符相关度较低的位置切分文本,与n-gram法相比,分词较完整地保留了文本信息。

特征构建是指将文本的内容以数值向量的形式表示。当用一个数值向量来描述文本在语义空

间中的位置时，用one-hot编码表示所有的term，然后删除文本中重复的term，再把得到的term的编码相加即可获得术语频率（Term frequency，TF）向量。

2.1.2　相似度计算

（1）欧式距离

$$\text{diatance}=\sqrt{(\boldsymbol{A}-\boldsymbol{B})\cdot(\boldsymbol{A}-\boldsymbol{B})^{\mathrm{T}}}=\sqrt{\sum_{i=1}^{N}(a_i-b_i)^2}$$

式中：$\boldsymbol{A}$、$\boldsymbol{B}$表示数值向量，即两个实例文本在欧式空间中的位置；$\boldsymbol{A}=(a_1,a_2,\cdots,a_i,\cdots,a_N)$；$\boldsymbol{B}=(b_1,b_2,\cdots,b_i,\cdots,b_N)$。

基于欧式距离的文本相似度为

$$\text{similarity}=\frac{1}{\text{distance}+0.001}$$

（2）余弦距离

$$\cos=\frac{\boldsymbol{A}\cdot\boldsymbol{B}^{\mathrm{T}}}{|\boldsymbol{A}|\cdot|\boldsymbol{B}|}=\frac{\sum_{i=1}^{N}a_i\cdot b_i}{\sqrt{\sum_{i=1}^{N}a_i^2}\cdot\sqrt{\sum_{i=1}^{N}b_i^2}}$$

基于余弦先距离的文本相似度计算公式为

$$\text{similarity}=1-\cos$$

（3）Jaccard相似度

Jaccard相似度用来比较两个集合之间的相似度。如有两个集合a和b，那么两者的Jaccard相似度的计算公式为

$$\text{similarity}=\frac{|a\cap b|}{|a\cup b|}=\frac{\text{len}(a\text{ and }b)}{\text{len}(a\text{ or }b)}$$

Jaccard相似度算法的本质是两个集合共有的元素越多，则判定二者越相似。

（4）海明距离

海明距离法通过比较不同文本的特征向量的每一个维度，给出在各个维度上的取值（0或1），再将所有取值相加。不相等的维度越多，即海明距离越大，则判定文档之间的差异度越大。海明距离的计算公式为

$$\text{distance}=\sum_{i=1}^{N}r_i$$

$$r_i=\begin{cases}1 & \forall a_i\neq b_i\\ 0 & \forall a_i=b_i\end{cases}$$

2.2　序列标注方法

2.2.1　BERT模型及输入表示

Google基于自注意力机制提出的BERT一种对文本语料库进行预训练的语言模型，它起源于预训练的上下文表示学习。与之前的模型不同，BERT是一种深度双向的、无监督的语言表示，且只使用纯文本形式的语料库进行预训练[8]。BERT的特点是在训练过程中会考虑到语料所在处的前后文。

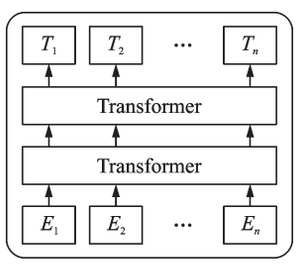

图1　BERT模型结构图

Fig.1　Structure diagram of BERT model

BERT通过Transformer结构搭建了一个多层双向的Encoder网络，从而实现深度双向语言表征。Transformer结构适合应用在多项NLP任务上，并具有快速训练的特点。

根据特定的需求，BERT模型的输入形式有单句或句对。对于任一个Token输入，它的嵌入词取决于其相应的词嵌入（Token embedding）、段嵌入（Segment embedding）和位置嵌入（Position embedding），如图2所示。

BERT中的Positional embedding对于每个时刻的位置并不是采用公式计算出来的，其原理也是类似普通的词嵌入一样，为每一个位置初始化了一个向量，然后随着网络一起训练。

BERT的输入表示有以下特点：

（1）对于中文模型，基于汉字的模型可以直接参与训练。

（2）模型输入需要使用一个起始Token，将其记为[CLS]，对应最终的Hidden state（即Transformer的输出），用其代表语料进行后续的分类工作。

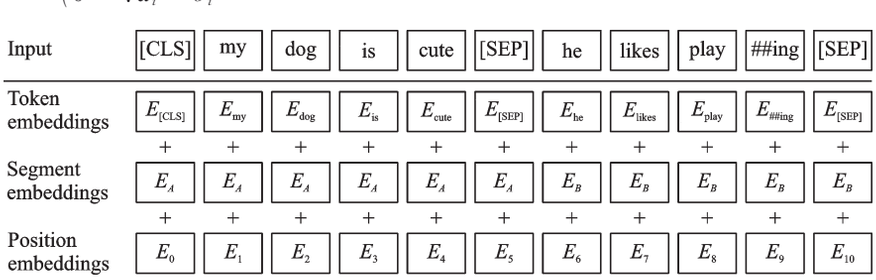

图2　BERT的输入表示

Fig.2　Input representation of BERT

(3)模型能处理句间关系。标记符[SEP]用来隔开两个句子,对于不同的语句,将学习到的Segment embeddings加到每个Token的Embedding上。

(4)每个单句(句对)输入,有一种(两种)Segment embedding。

2.2.2 基于掩盖的语言模型策略

为了使得模型能够有效的学习到双向编码的能力,BERT在训练中使用了基于掩盖的语言模型(Masked language model,MLM),即随机对输入序列中的某些位置进行遮蔽,然后通过模型来对其进行预测。如图3所示,其做法是随机掩盖掉输入序列中15%的Token(即用[MASK]替换掉原有的Token),然后在BERT的输出结果中取对应掩盖位置上的向量进行真实值预测。

更具体来讲,先选定15%的Token,然后将其中的80%替换为[MASK]、10%随机替换为其他Token、剩下的10%不变。最后取这15%的Token对应的输出做分类来预测其真实值。

由于MLM预测任务能够使得模型编码得到的结果同时包含上下文的语境信息,因此有利于训练得到更深的BERT网络模型。

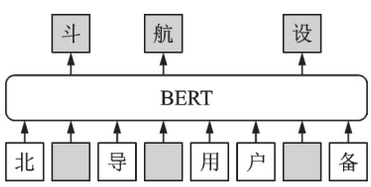

图3 BERT MLM策略示例

Fig.3 Example of BERT MLM policy

2.2.3 BERT序列标注

序列标注是自然语言处理(Natural language processing, NLP)的一个必要步骤,主要包括分词、词性标注、命名实体识别等子任务,而常通过fine-tuning对预训练后的BERT模型进行序列标注。常见的序列标注任务有3种:

(1)名词-实体识别(Named-entity recognition, NER)分辨出文本中的名词和实体(person人名,organization组织机构名,location地点名);

(2)词性标注(Part-of-speech tagging,POS)根据语法对token进行词性标注(noun名词, verb动词,adjective形容词);

(3)短语组块(Chunking,Chunk)将同一个短语的tokens组块放在一起。

本文将使用BERT for sequence classification,其由一个普通的BERT模型和一个单线性分类层构成。BERT模型对文本语料库进行预训练,单线性分类层对文本进行分类。当向模型输入数据时,整个预训练的BERT模型和未训练的单线性分类层将会一起被训练。

3 实 验

3.1 语料数据及预处理

本文基于GB/T 9390—2017《导航术语》[9]、GB/T 39267—2020《北斗卫星导航术语》[10]、GB/T 14733.6—2005《电信术语 空间无线电通信》[11]3个术语标准中的数据集进行试验,该评测针对一定数量的中文航天工作报告中的表述原词进行语义标准化。为了删除文本中的不规范的无关字符,首先需要对实验数据进行预处理;预处理后再计算原始词所对应的标准词的数量,而后分割具有多个对应标准词的原始词,进而使用相似度计算得到候选词的集合,以用于后期的预测。

中文文本中,BERT模型的最小输入单位为一个汉字,但中文的词性信息是根据词语进行标注的,为了适用于BERT的算法,需要将原中文文本拆分成一系列的汉字,并对每个汉字进行词性标注。具体进行编码的函数tokenizer.encode_plus包含步骤如表2所示。

表2 编码函数tokenizer.encode_plus步骤

Table 2 Steps of Encoding function tokenizer. encode_plus

序号	步骤操作
1	将句子分词为tokens
2	在两端添加特殊符号[CLS]和[SEP]
3	将tokens映射为下标IDs
4	将列表填充或截断为固定的长度
5	创建attention masks,将填充的和非填充tokens区分开来

3.2 评价指标

(1)正确率(Accuracy),即准确识别的正例(TP)与负例(TN)占总识别样本数(S)的比例。

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{S}$$

(2)召回率(Recall),即准确识别的正例(TP)占实际总正例的比例。其中,实际总正例为准确识别的正例与错误识别的负例之和。

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

(3)精确度(Precision),即准确识别的正例(TP)占识别出的正例的比例。其中,识别出的正例为准确识别的正例与错误识别的正例之和。

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

(4)F score，即R和P的加权调和平均，通过引入系数α对R和P进行加权调和。

$$F=\frac{(\alpha^2+1)P\cdot R}{\alpha^2(P+R)}$$

3.3　训练结果及结论

3.3.1　文本相似度计算实验

为了理解文本的相似度，对比分析前面提到的4种相似度计算方法，设计实验进行计算后得出结果如表3所示。

表3　使用不同方法生成的候选词前30个结果比较

Table 3　Comparison of the first 30 results of candidate words generated by different methods

方法	召回率/%	耗时/s
欧式距离	39.83	33.87
余弦距离	42.19	1 876.52
Jaccard	86.87	39.71
海明距离	84.26	23.68

从表3中可以得到以下结论：

(1)4种相似度计算方法中，Jaccard相似度系数计算方法优于其他3种，召回率最高，候选标准词生成匹配精确度高，且耗时较少，较为高效。

(2)其他3种相似度计算方法中，欧式距离作为其中最简单、最易理解的方式，相似度计算简单，受限制因素也较多，导致耗时虽然少，但召回率较低，导致生成的候选标准词匹配度较低；余弦距离对数据的绝对数值不敏感，计算时不考虑数据之间的共同评分项数量，导致效率低且效果差；海明距离的局部敏感性强，可以对两个相似的语义进行局部敏感计算，实验结果生成候选词匹配度略低于Jaccard相似度系数计算方法。

3.3.2　训练集构造及BERT-base参数设置

以6∶2∶2的比例指定训练集、验证集、测试集，本次实验将训练集、验证集、测试集的数据词条分为确定为3 000、1 000、1 000。

候选标准词的匹配是基于BERT模型的分类任务，需要构建包含正负例的训练集用于训练模型。

正例：＜原词，标准词，1＞

构建负例：计算原词与标准词词表中所有标准词的Jaccard相似度，取相似度值在前20% 的标准词组成候选词表，并在候选词表删除标准词，其余的候选词即为负例。

负例：＜原词，标准词，0＞

本文采取BERT-base参数模型，具体参数设置如表4所示，fine-tuning训练时，首先要使用Auto model for token classification加载预训练模型，随后将数据/模型/参数传入Trainer开始训练，最后使用evaluate方法进行评估，从而获取训练结果的精确度。

表4　BERT-base参数设置

Table 4　BERT-base parameter setting

序号	BERT参数	参数值
1	max_seq_length(Token序列的最大长度)	128
2	train_batch_size,(batch大小)	16
3	learning_rate(学习率)	2e−5
4	num_train_epochs(训练的epoch次数)	4

3.4　结果分析

针对训练集中出现正负例不均衡的问题，根据孙曰君等[12]实验结论，本文采用10倍正例加负例的组合方式，以提高BERT方法的性能，从而得出此条件下的精确度为90.29%、召回率为92.81%、正确率为89.93%、F score为91.05%。对于本文采用的基于BERT的航天术语标准化方法的预测结果，分别选取了4个预测正确的示例和2个预测错误的示例，如表5所示。

表5　预测结果样例

Table 5　Example of prediction results

原词	预测标准词	标准词	预测结果
地面导航无线电系统	陆基无线电导航系统	陆基无线电导航系统	正确
航空导航完全递减特性比	航道弯曲递减因子	航道弯曲递减因子	正确
改进零基准下滑天线	边带基准下滑天线	边带基准下滑天线	正确
几何式地理导航惯性系统	几何式惯性导航系统	几何式惯性导航系统	正确
运载体设备测量导航数据	运载体测量导航数据	运载体导出导航数据	错误
空中运载测量导航数据	空中运载导航数据	空中导出导航数据	错误

从实验结果数据可以看出，BERT训练速度快且准确度高，能按预期解决中文文本的序列标注问题，从而提高航天术语标准化的效果。根据表5的预测结果，从预测正确的示例来看，本文采用的基于BERT的航天术语标准化方法可以很好地捕捉到航天语义原词中的重点词组，能够通过不统一、不规范的航天领域语义原词中得到较为准确的标准词预测结果；通过分析预测错误的示例可以发

现,对于航天语义原词采用“位置+方式+用途+性质”组成的航天语义原词,在使用BERT模型进行术语标准化时,并没有充分考虑到位置、方式、用途、性质之间的互相匹配,且未得到正确修正,导致预测结果不准确,这也是此方法在后续应用过程中可以继续完善和改进的地方。

4 结 论

术语对保证技术概念和语言的准确具有头等重要的作用,航天术语标准化旨在通过规范化管理,实现多功能、高质量的信息化管理与质量控制,从而推动航天标准化工作的进步与发展。

本文发挥深度学习在文本建模上强大的语义表征能力,将航天术语标准化问题与深度学习相结合,对一定量的航天文本原词数据进行语义标准化。首先通过4种方法计算待标准化的原词与标准词之间的相似度,根据相似度对标准词进行排序得到候选标准词词表,然后将航天语义原词和候选标准词进行匹配,并将航天语义原词与候选标准词语义相似度的计算问题转化为二分类问题,从而获得预测标准词。

实验表明,BERT在自然语言处理任务上具有强大优势,在基于多个术语标准评测数据集上准确率达到89.93%,说明BERT模型方法可以提高航天术语标准化水平和效果。但另一方面,通过对预测样例的分析可以发现,对于较为复杂的航天语义原词,本文采用的方法仍会出现预测错误的情况,在未来应用此方法时,应予以注意,可以从航天术语本身的构成特点入手,增加相关语义特征,从而实现更好的预测效果。

参考文献:

[1] 中共中央国务院.国家标准化发展纲要[J].中国标准化,2021(21):9-16.
The CPC Central Committee and the State Council. The national standardization development outline[J]. China Standardization, 2021(21): 9-16.

[2] 魏永刚,代健,杨晓明.航天标准化的发展与展望研究[J].西北工业大学学报,2018,36(S1):28-32.
WEI Yonggang, DAI Jian, YANG Xiaoming. Research on the development and prospect of aerospace standardization[J]. Journal of Northwest University of technology, 2018,36(S1): 28-32.

[3] 查朝晖,刘海涛.航天术语标准化的研究与探讨[J].航天标准化,1999(3):3-7.
ZHA Zhaohui, LIU Haitao. Research and discussion on standardization of aerospace terminology[J]. Aerospace standardization, 1999(3): 3-7.

[4] 张嵩,杨晓明,田露.基于语义相似度计算的航天标准关联度评价[J].南京航空航天大学学报,2021,53(S1):153-156.
ZHANG Song, YANG Xiaoming, TIAN Lu. Relevance evaluation of aerospace standards based on semantic similarity calculation[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2021,53(S1): 153-156.

[5] 王洪东,侯雪辉,高蕴慧,等.基于语言值有序对三元组的多类型数据标准化方法[J].模式识别与人工智能,2019,32(3):278-286.
WANG Hongdong, HOU Xuehui, GAO Yunhui, et al. Multi type data standardization method based on language value ordered pair triples[J].Pattern Recognition and Artificial Intelligence, 2019,32(3): 278-286.

[6] 杜迎雪,常娟,刘卫锋.基于最大长度区间的区间权重向量标准化方法[J].数学的实践与认识,2020,50(7):8-16.
DU Yingxue, CHANG Juan, LIU Weifeng. Interval weight vector standardization method based on maximum length interval[J]. Practice and Understanding of Mathematics, 2020,50(7): 8-16.

[7] 崇伟峰,李慧,李雪,等.基于BERT蕴含推理的术语标准化系统[J].中文信息学报,2021,35(5):86-90.
CHONG Weifeng, LI Hui, LI Xue, et al. Term standardization system based on Bert implication reasoning [J]. Chinese Journal of Information Technology, 2021,35(5): 86-90.

[8] HO Q T, NGUYEN T, LE N, et al. FAD-BERT: Improved prediction of FAD binding sites using pre-training of deep bidirectional transformers[J]. Computers in Biology and Medicine, 2021, 131(2): 104258.

[9] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会.导航术语:GB/T 9390—2017[S].[S.l.]:[s.n.], 2017.

[10] 国家市场监督管理总局,国家标准化管理委员会.北斗卫星导航术语:GB/T 39267—2020[S].[S.l.]:[s.n.], 2020.

[11] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会[S].电信术语 空间无线电通信:GB/T 14733.6-2005[S].[S.l.]:[s.n.], 2005.

[12] 孙曰君,刘智强,杨志豪,等.基于BERT的临床术语标准化[J].中文信息学报,2021,35(4):75-82.
SUN Yuejun, LIU Zhiqiang, YANG Zhihao, et al. Standardization of clinical terms based on BERT[J]. Chinese Journal of Information Technology, 2021,35(4): 75-82.

(编辑:孙静)