文章编号：1003-0077(2021)05-0086-05

基于 BERT 蕴含推理的术语标准化系统

崇伟峰，李 慧，李 雪，任 禾，于 东，王晔晗

（云知声智能科技股份有限公司 人工智能实验室，北京 100096）

摘 要：临床术语标准化即对于医生书写的任一术语，给出其在标准术语集合内对应的标准词。标准词数量多且相似度高，存在 Zero-shot 和 Few-shot 等问题，给术语标准化带来了巨大的挑战。该文基于“中国健康信息处理大会”CHIP 2019 评测 1 中提供的数据集，设计并实现了基于 BERT 蕴含分数排序的临床术语标准化系统。该系统由数据预处理、BERT 蕴含打分、BERT 数量预测、基于逻辑回归的重排序四个模块组成。用精确率(Accuracy)作为评价指标，最终结果为 0.948 25，取得了评测 1 第一名的成绩。

关键词：BERT；术语标准化；蕴含推理

中图分类号：TP391 **文献标识码**：A

Term Normalization System Based on BERT Entailment Reasoning

CHONG Weifeng, LI Hui, LI Xue, REN He, YU Dong, WANG Yehan

(AI Labs, Unisound AI Technology Co. Ltd., Beijing 100096, China)

Abstract: The normalization of clinical terms is to assign a corresponding term in the standard term set to any term written by the doctor. This task is challenged by large amount of standard terms with high mutual similarity, as well as insufficient training data known as Zero-shot or Few-shot learning. This paper designs and implements a clinical term normalization system based on BERT entailment ranking. The system consists of four modules: data preprocessing, BERT entailment scoring, BERT quantity prediction, and logistic regression-based reordering. Tested in CHIP 2019 Track 1 "Evaluation of Chinese Clinical Term Normalization", it achieves a final accuracy of 0.948 25 as the top score in this campaign.

Keywords: BERT; term normalization; entailment

0 引言

临床术语标准化是医学统计中不可或缺的一项任务。DRGs(diagnosis related groups)中文翻译为疾病诊断相关分类。一直以来，DRGs 付费制度改革被认为是医保支付方式改革的重点。2019 年国家医疗保障局召开“疾病诊断相关分组 DRGs 付费国家试点工作启动”视频会议，公布了 30 个试点城市，以此推动 DRGs 付费方式的试点。DRGs 的指导思想是：通过统一的疾病诊断分类定额支付标准的制定，达到医疗资源利用标准化。在 DRGs 分组中，诊断和手术的术语标准化起着至关重要的作用[1]。临床上，关于同一种诊断、手术、药品、检查、化验、症状等往往会有成百上千种不同的写法。标准化(归一)要解决的问题就是为临床上各种不同说法找到对应的标准。

本文的主要目标是针对中文电子病历中挖掘出的真实语义实体进行语义标准化，即对于给定的任一手术原始词，找到对应的手术标准词(下文用原始词、标准词分别代表手术原始词和手术标准词)。本文所使用的数据集来自于“中国健康信息处理大会”CHIP2019 评测 1 的官方数据，所有的原始词均来自真实医疗数据，并以《ICD-9-CM-3-2017 协和临床版》作为标准词表。

通过对本任务以及数据的分析我们归纳出该任务的四个难点：①标准词数量特别大，标准词表中共有 9 866 个不同的编码、9 467 个不同的标准词。

收稿日期：2020-06-15 **定稿日期**：2020-11-28

一般来讲，标准词数量越多，任务的难度越大。②有些标准词之间相似度非常高，比如“硬脊膜外病损切除术”和“硬脊膜下病损切除术”之间只有一字之差，而且这个字在通用的自然语言处理任务中通常不是特别重要。③Few-shot 和 Zero-shot 的问题，统计评测数据的开发集中共有 480 个标准名称，其中有 111 个没有出现在训练集中，占比达到 23%，另外还有 66 个标准名在训练集中只出现了一次。由此可见，Few-shot 和 Zero-shot 问题确实比较严重。④标准词的数量是不确定的。医生写的原始词中有可能出现多种分隔符，比如“＋”“，”等，但是根据分隔符的数量并不能确定标准名的数量。比如原始手术词“宫腔镜检查术＋分段诊断性刮宫术”中包含了分隔符“＋”，但是对应的标准名只有一个“宫腔镜诊断性刮宫术”。另外“经皮肾镜碎石取石”中没有分隔符，但是其对应的标准名却是两个，“经皮肾镜碎石术(PCNL)”和“经皮肾镜取石术”。这种情况给该任务带来了巨大挑战。图 1 为一些难点的举例。

手术原始词	手术标准词	
右肾上腺巨大肿瘤切除术	肾上腺病损切除术	原始词表达方式多样 原始可能对应多个标准词
左股骨肿瘤刮除植骨内固定术	股骨内固定术##股骨病损切除术##股骨移植术	
LC	腹腔镜下胆囊切除术	
经皮肾镜碎石取石	经皮肾镜碎石术（PCNL）##经皮肾镜取石术	原始词的分隔符不同 分隔符划分的数量与标准数量有差异
全肠脏切除+回肠肠脏术	小肠段分离术##根治性肠脏切除术##可控回肠肠脏术	
宫腔镜检查术+分段诊断性刮宫术	宫腔镜诊断性刮宫术	
经腹左肾根治术，左肾上腺切除术	单侧肾切除术##单侧肾上腺切除术	
1、左股骨粗隆骨折ORIF2、股骨头钻孔减压	股骨骨折切开复位内固定术##股骨减压术	

图 1　手术术语标准化难点举例

对于术语标准化问题，有研究者使用编辑距离计算相似度进行处理[2]，也有研究者利用卷积神经网络模型进行处理[3]。而 BERT(bidirectional encoder representations from transformers)[4]最近在众多的自然语言处理任务中表现非常出色。其主要是借助大规模语料的预训练，得到网络参数后，在下游任务中使用相同的网络结构，对参数进行调整，从而直接解决各种自然语言处理任务。在我们的任务中，多个难点都可以通过 BERT 来解决或者缓解。比如 Zero-shot 和 Few-shot 问题可以借助 BERT 通过预训练得到的知识来弥补。标准词多的问题可以通过利用 BERT 将多分类问题转换成蕴含判断的二分类问题。由于 BERT 中学到的表征是上下文相关的，即使两个文本相似度很高，也可能有不同的表征。标准词不能根据分隔符的数量来确定，我们将其转换为一个机器学习问题，使用基于 BERT 的微调来解决问题。通过上述分析，我们构建了一套以 BERT 为核心的术语标准化系统。

接下来的内容中将采用如下方式组织：第 1 节介绍系统的整体架构，第 2 节详细介绍每个模块，第 3 节介绍系统的整体表现，最后第 4 节进行总结和展望。

1　系统框架

基于上文分析的难点，本文构建了术语标准化系统。该系统由数据预处理、BERT 蕴含打分、BERT 数量预测、基于逻辑回归的重排序四个模块组成，如图 2 所示。首先对手术原始词进行必要的预处理，方便后续的模块处理；然后对预处理后的手术原始词和所有的标准名计算蕴含分数，同时对手术原始词和标准名利用命名实体识别(named entity recognition，NER)解析出其中的部位和术式，计算手术原始词和标准词中的对应成分(部位和术式)的相似度，将蕴含分数和两个相似度分数作为重排序模块的特征，对所有候选项进行排序。同时，将标准化后的手术原始词作为输入进行数量分类，得到标准名的数量，最后取重排序后得分最高的 k 个作为标准名输出。各个模块的详细介绍见第 2 节。

2　方法

本节将按照顺序介绍数据预处理、BERT 蕴含打分、基于逻辑回归的重排序、BERT 数量预测等四个模块。

2.1　数据预处理

预处理的目的主要是消除输入中与任务无关的信息，同时增强对任务有正向作用的信息。本次评测中我们使用的预处理操作主要有：①规范化操作，包括将所有的全角字符转为半角字符，将所有数字标准化为阿拉伯数字，去掉输入中自带的编码。②修改英文数据对应的标准名，如原始词“PICC”对

应的标准词为“经外周静脉穿刺中心静脉置管术”，修改标准词为“经外周静脉穿刺中心静脉置管术(PICC)”。③修改包含“不”的标准名。如“手指骨折开放性复位术不伴内固定”修改为“手指骨折开放性复位术”。预处理中对标准词做的修改，会在最后给出结果时进行还原。

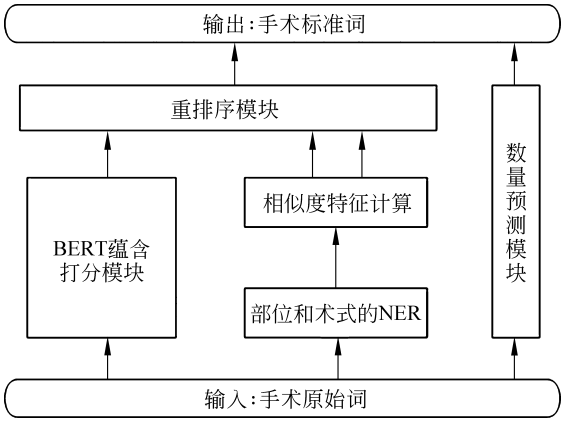

图 2 系统整体架构图

2.2 BERT 蕴含打分

蕴含打分模块主要是对给定的手术原始词计算其与每一个标准名的蕴含分数。这个过程我们主要利用 BERT 来实现。BERT 采用预训练—微调(pretraining-finetuning)的模式解决自然语言处理任务(NLP)[4]，其提及在 11 个自然语言处理(natural language processing，NLP)任务中获得了新的最先进(state-of-the-art)结果。采用新的掩码语言模型(masked language model，MLM)作为训练目标，具体来讲以 15%的概率用掩码字符(mask token)([MASK])随机地对每一个训练序列中的 token 进行替换，然后预测出[MASK]位置原有的单词。这样训练出来的模型能生成深度的双向语言表征，词语的表征和上下文背景有关。为了能够处理如问答、自然语言推断等任务，需要理解两个句子之间的关系，而 MLM 任务倾向于抽取 token 层次的表征，因此不能直接获取句子层次的表征。为了使模型能够理解句子间的关系，BERT 使用了预测是否下一句(next sentence predict，NSP)任务来预训练，简单来说就是预测两个句子是否连在一起。具体的做法是：在语料库中挑选出句子 A 和 B 组成句子对，50%的句子对是原文中连贯的，50%是两个句子随机配对的，把句子对输入到 BERT 模型中，进行二分类的预测。通过 MLM 和 NSP 任务预训练得到的模型，在微调阶段既可以处理单个句子的分类任务，又可以处理句子对分类任务。

本研究使用 BERT 计算蕴含分数的具体实现方法如图 3 所示。

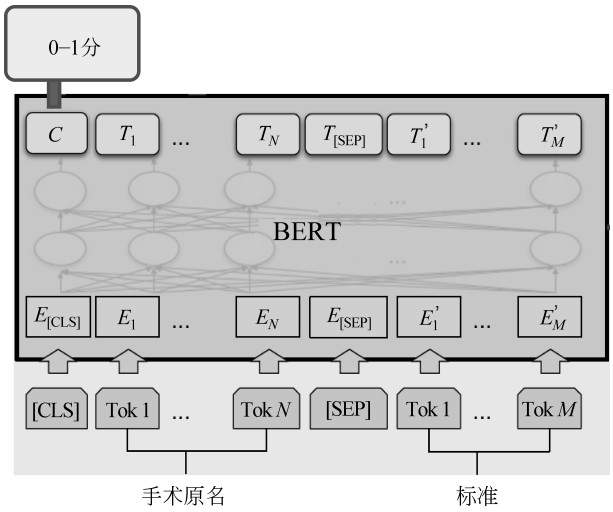

图 3 BERT 蕴含打分模块

将手术原始词和标准名拼接起来作为输入，进行 0-1 分类，0 代表错误的标准名，1 代表正确的标准名，将分类为 1 的概率作为蕴含分数。为了实现这个功能，我们需要构造微调 BERT 模型的数据集。对任意一条评测数据中手术原始词 O_i 和对应的标准词 S_i，构造数据集的方法主要包括如下 3 步：①将(O_i，S_i，1)加入数据集。②对于任意的标准词 S_j，如果 $S_j \neq S_i$，且 O_i 和 S_j 的相似度或者 S_i 和 S_j 的相似度大于给定的阈值，则将(O_i，S_j，0)加入数据集。本方案中，相似度计算方法使用最长公共子字符串相似度，计算如式(1)所示。

$$\frac{\text{len}(\text{LCS}(\text{str1},\text{str2}))}{\max(\text{len}(\text{str1}),\text{len}(\text{str2}))} \tag{1}$$

其中，LCS 表示最长公共子字符串。在这里相似度阈值取 0.75。在构建负例时，我们尝试加入随机负例，模型指标并没有提升。为了加快模型训练速度，在本研究中未加入随机负例。

数据集构建完成后，采用不同的参数训练多个模型，将多个模型的蕴含分数求平均值作为最后的分数。通过实验发现集成 5 个模型后能够达到最高指标。其超参设置如表 1 所示。

表 1 不同模型的超参设置

模型编号	学习率	Epoch	数据量	优化器
1	2e-5	3	4 000	Adam
2	1e-5	3	5 000	Adam

续表

模型编号	学习率	Epoch	数据量	优化器
3	2e-5	3	5 000	Adam
4	3e-5	3	5 000	Adam
5	2e-5	5	5 000	Adamax

2.3　基于逻辑回归的重排序

由于数据量以及数据分布的问题，单纯利用端到端的模型，并不能解决术语标准化中的所有问题。为了更充分利用数据，我们对数据集中的手术原始词和标准词进行了命名实体识别(named entity recognition，NER)，将术语中的手术部位和手术术式抽取出来[5]。进行 NER 的具体方式是用 BERT 进行序列标注任务的微调，其训练数据的构造过程如下：①先根据词表进行回标。②人工修正，得到 1 000 条训练数据。③使用 1 000 条训练数据对 BERT 进行微调。④对 500 条数据进行预测并修正其中错误，得到 1 500 条训练数据，训练最终模型，得到部位和术式后，利用 BERT 模型得到部位和术式的向量表示，然后分别计算手术原始词和标准词之间部位的相似度以及手术原始词和标准词之间术式的相似度。相似度用两个向量之间的余弦相似度来表示。

将 BERT 蕴含分数(用 S_E 表示)、部位相似度(用 S_P 表示)、术式相似度(用 S_O 表示)一起作为逻辑回归模型的特征，如式(1)、式(2)所示。

$$\alpha = a_1 * S_E + a_2 * S_P + a_3 * S_O \tag{2}$$

$$y = \frac{1}{1 - e^{-\alpha}} \tag{3}$$

因为 4 000 条训练集上的蕴含分数和相似度分数已经用来训练，会出现分数和测试集中不一致情况，所有逻辑回归参数使用 1 000 条开发集进行估计。得到参数后对 BERT 蕴含分数的 top30 的结果重新打分。NER 结果示意图如图 4 所示。

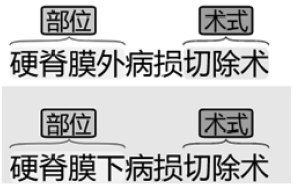

图 4　NER 结果示意图

2.4　基于 BERT 的数量预测

数量预测模块主要是确定需要输出多少个标准名。图 5 为数据集中每个原始词对应的标准词的数量分布。

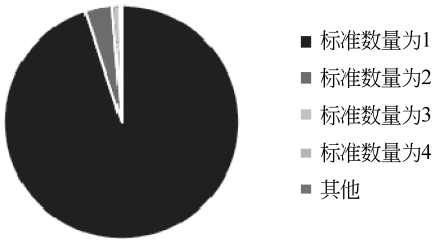

图 5　标准手术名称数量分布

图 5 中其他标签的数量仅为 1，因此我们将这个问题抽象成一个类别为{1，2，3，4}的多类别分类问题。对于真实标准名数量超过 4 的情况，我们直接根据手术原始词中分隔符的数量来确定。

在预测标准名数量时，采用如下方法构造数据集：①将原始数据集中的手术原始词加入到数量预测的数据集中，标准名的数量作为类别标签；②将 ICD-9-CM-3 协和 2017 版编码表中带有“伴”的标准名加入到数量预测数据集中，类别标签为 1；③将 CHIP 数据中拆分出来手术原始词没有分隔符、标准名数量为 2 的手术原始词加入到数量预测数据集中，类别标签为 2；手术原始词有分隔符，但是只有一个标准名的手术原始词中，类别标签为 1；④利用前三步得到的数据进行数据拼接，使得数据集中标签为 1、2、3、4 的实例(case)的数量大体相当。

完成数据集的构建后，利用 BERT 微调参数，进行数量预测。同时结合之前处理标准名超过 4 的规则，确定最终标准名的数量。

3　结果

我们基于谷歌发布的中文基本(Base)版本的模型进行微调。该模型总共有 12 层，Transformer 中隐变量的维度为 768，多头注意力机制中注意力头数为 12，总参数数量超过 110M。在微调过程中我们将出现在手术名称中但不包含在 BERT 词表中的字符加入到词表中。

表 2 是数量预测模块的结果，基于 BERT 的数量预测模块在开发集上的精确率(Acc)达到了 0.988，在盲测集上的 Acc 达到 0.991，明显高于采用规则和启发式方法判定标准名个数的指标。通过分析数据发现，之所以测试集上的指标比开发集上指标高，是因为测试集中标准名数量为 1 的比例更大，数量预测更容易。

表 2 数量预测模块的结果

数量预测方法	Dev	Test
简单规则	0.945	—
默认数量为 1	0.950	—
基于 BERT 分类	0.988	0.991

表 3 展示了加入不同的模块后本文系统的表现。如果用基于 BERT 的蕴含分数直接选择数量预测出来 top k 的结果，通过调整训练模型的参数，本文单个模型的最好指标 Acc＝0.92；将 5 个不同的模型集成后，Acc 指标提高到 0.932；加入部位和术式的相似度后，并利用逻辑回归重排序后开发集上 Acc 达到 0.941；最终的系统在测试集上的 Acc 为 0.948 25，在所有 19 支参赛队伍中名列第一。

表 3 手术术语标准化结果正确率(Acc)指标

术语标准化方法		Dev	Test
BERT 蕴含判断	单个模型	0.920	—
	集成模型	0.932	—
回归重排序		0.941	0.948 25

4 结论与展望

我们提出一套基于 BERT 蕴含推理的术语标准化系统，包括数据预处理、BERT 蕴含打分、基于逻辑回归的重排序、BERT 数量预测四个模块。这套系统在 CHIP-2019 评测任务 1“临床术语标准化任务”中取得了第一名的好成绩。该系统不但适用于手术临床术语的标准化，也能较好地迁移到其他类型的术语标准化任务中，比如疾病或者检查、检验等。

本系统中 BERT 蕴含打分和 BERT 数量预测两个模块是核心模块，其余模块主要是为了提升性能指标。在优化系统的过程中，我们发现一些有用的结论：术语标准化问题可以定义成一个蕴含推理的问题，这个问题可以用 BERT 较好地解决。在这个转换过程中，负例的构造尤其关键。另外，BERT 能够通过手术原始词较好地预测标准词的数量，这展示了 BERT 强大的模型表征能力。在使用 BERT 微调参数完成任务的过程中，对数据进行适当的增强能够比较显著地提升术语标准化的精确率。手术原始词和标准词的手术部位相似度以及手术原始词和标准词的术式相似度能够起到一定作用。

在以后的工作中，我们希望能将这套系统扩展到其他版本的手术术语标准化以及诊断术语标准化中。由于当前国内各个省使用的 ICD 版本差异性比较大，国家层面的医疗统计和政策执行困难，使得基于病历的临床医学研究工作受到制约。因此，我们的任务是研发出能够适应各个版本的手术和诊断术语标准化系统，并能够在不同版本之间进行转换，这是一件对医疗行业意义重大的事情，我们会为此持续不断地投入精力。另外，已知部位和术式在手术术语标准化中起到一定作用，后续我们还会尝试分析其他成份，比如入路、内镜、疾病性质等对手术术语标准化所起的作用。

参考文献

[1] 秦安京. 疾病分类编码准确是诊断相关分组(DRGs)的保障[J]. 中国病案，2007，8(7)：10-11.

[2] 赵亚辉. 临床医疗实体链接方法研究[D].哈尔滨：哈尔滨工业大学硕士学位论文，2017.

[3] Li Haodi，Cheng Qingcai，Tang Buzhou，et al. CNN-based ranking for biomedical entity normalization[J]. BMC Bioinformatics，2017，18(11)：80-91.

[4] Jacob Devlin，Ming-Wei Chang，Kenton Lee，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810.04805，2018.

[5] 刘爱民.病案信息学[M].北京：人民卫生出版社，2014：206-209.

崇伟峰(1985—)，硕士，算法工程师，主要研究领域为语义理解、知识图谱。

E-mail：chongweifeng@unisound.com

李雪(1994—)，通信作者，硕士，算法工程师，主要研究领域为语义理解、海量数据查询。

E-mail：xlee9403@163.com

李慧(1990—)硕士，算法工程师，主要研究领域为自然语言理解、数据挖掘。

E-mail：lihui@unisound.com