

文章编号: 1003-0077(2021)03-0088-06

## 基于 BERT 的手术名称标准化重排序算法

陈漠沙, 仇 伟, 谭传奇

( 阿里巴巴 机器智能技术实验室, 浙江 杭州 311121)

**摘 要:** 临床术语标准化是医学文本信息抽取中不可或缺的一项任务。临床上对于同一种诊断、手术、药品、检查、化验、症状等, 往往会有多种不同的写法, 术语标准化(归一)要解决的问题就是为临床上各种不同的说法找到对应的标准名称。在检索技术生成候选答案的基础上, 该文提出了基于 BERT(bidirectional encoder representation from transformers) 对候选答案进行重排序的方法。实验表明, 该方法在 CHIP2019 手术名称标准化数据集上单模型准确率达到 89.1%、融合模型准确率达到 92.8%, 基本满足实际应用标准。同时该方法具备较好的泛化能力, 可应用到其他医学种类术语的标准化任务上。

**关键词:** 手术名称标准化; Lucene 检索; BERT

**中图分类号:** TP391

**文献标识码:** A

### A BERT Based Reordering Method for Clinical Operation Term Normalization

CHEN Moshu, QIU Wei, TAN Chuanqi

(Machine Intelligence Technology, Alibaba, Hangzhou, Zhejiang 311121, China)

**Abstract:** Clinical term normalization is an indispensable task in clinical text information extraction. There are often various ways of writing about the same clinical term like diagnosis, operation, medicine, examination, laboratory test, symptom, etc., and term normalization is to find the corresponding standard name for different clinical terms. Based on the candidate answers generated by information retrieval tools, this paper proposes a method of reordering candidates based on BERT (Bidirectional Encoder Representation from Transformers). The experimental results show that the accuracy of single model and fusion model achieves 89.1% and 92.8%, respectively.

**Keywords:** clinical operation term normalization; Lucene information retrieval; BERT

## 0 引言

临床术语标准化任务是医学文本信息抽取中不可或缺的一项任务。患者病历详细记录了患者的临床病史和疾病进展, 包括但不限于症状、疾病、诊断和药物等。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说法, 一方面可以帮助研究人员对电子病历进行后续的统计分析和挖掘; 另一方面也可以促进 AI 技术在医学应用系统的落地, 如“CDSS(临床决策诊疗系统)”“DRGs(诊断相关分组管理系统)”等。

近年来医院信息化建设水平的提升以及电子病历的普及, 为大规模应用机器学习算法解决医学领域的研究任务提供了便利。针对医学术语归一(标准化)任务, 学术界也开展了一系列评测任务, 包括: ShARe/CLEF-2013 任务 1<sup>[1]</sup>、SemEval-2014 任务 7<sup>[2]</sup>、SemEval-2015 任务 14<sup>[3]</sup> 和 N2C2-2019 任务 3<sup>[4]</sup>, 以上评测任务均是英文领域。针对中文数据集, 在目前所知范围内, CHIP-2019“手术名称标准化”任务<sup>①</sup>应该是第一个中文医学术语归一化评测任务。这些评测任务和数据集均有效地推动了相关技术的发展。

术语归一任务的研究大致经过了几个阶段: 基

① <http://cips-chip.org.cn/evaluation>

于规则的方法<sup>[5-6]</sup>、基于机器学习的方法<sup>[7]</sup>和基于深度学习的方法<sup>[8-9]</sup>。本文在使用检索技术生成候选答案的基础上,借鉴了语言模型 BERT 的思路,在重排序阶段使用 BERT<sup>[10]</sup>模型对候选答案进行打分重排序,单模型取得 89.1% 的准确率,最终提交的融合模型在测试集上取得 92.8% 的准确率,基本达到实际应用标准。本文采用的方法具备很好的泛化能力,可应用在其他类型的医学术语标准化任务上。

## 1 相关工作

早期的研究方法均是围绕基于规则的方法开展的,其中代表性的工作包括 Ghiasvand 等<sup>[5]</sup>在 SemEval-2014 任务 7<sup>[2]</sup>上提出的基于编辑距离特征来生成候选集的方法,该方法首先通过训练数据的每一条实体及其在 UMLS 系统对应的标准术语学习到 554 种编辑距离模式,该模式之后被应用到测试集来增强候选答案的覆盖,在 SemEval 任务上取得了最佳性能;Kang 等<sup>[6]</sup>在生物领域文本上提出了 5 种规则来提升疾病术语的归一化性能。

针对实体归一化任务,常见的机器学习解决思路是利用学习排序(learning to rank<sup>①</sup>)的方法来生成最终的答案,在医学 NLP 领域的代表性工作是 Leaman 等<sup>[7]</sup>提出的利用 pairwise 排序学习的方法,该方法利用向量空间来表示实体以及数据库中的标准术语,通过学习相似矩阵来完成给定实体和候选答案之间的匹配映射。

深度学习时代,研究者对实体归一任务提出了更广泛的解决思路,其中代表性工作是 Luo 等<sup>[8]</sup>提出的多视图 CNN 模型,该方法利用多任务共享网络结构同时学习出院记录的诊断信息和手术信息标准化,在中文数据集上取得了不错的性能;随着语言模型的兴起,预训练语言模型在多项 NLP 任务上均刷新了最优结果,针对医学术语归一任务, Ji 等<sup>[9]</sup>提出的利用 BERT 进行重排序的思路,在多个标准数据集上均取得了最佳性能。

本文在前述方法的基础上,基于通用的检索+重排序框架,利用给 BERT 输入对(实体词,候选词)进行打分重排序,在 CHIP 评测任务上取得了不错的性能效果。

## 2 任务描述与数据统计

### 2.1 任务描述

针对中文电子病历中挖掘出的真实手术实体进

行语义标准化。具体来说,给定一手术原词,要求系统给出其对应的手术标准词。其中,所有手术原词均来自于真实医疗数据,并以《ICD9—2017 协和临床版》手术词表为标准进行标注。

### 2.2 数据统计

表 1 展示了本次评测任务的部分样例数据,数据集中“手术原词”和“归一化标准词”之间存在 4 种匹配对应关系:

- (1) “一对一”关系:一个手术原词对应一个归一化标准词。
- (2) “一对多”关系:一个手术原词对应多个归一化标准词。
- (3) “多对一”关系:多个手术原词对应一个归一化标准词。
- (4) “多对多”关系:多个手术原词对应多个归一化标准词。

表 1 手术术语标准化任务示例

手术原词	归一化标准词
右肾上腺巨大肿瘤切除术	肾上腺病损切除术
右眼硅油取出联合人工晶体Ⅱ期植入术	玻璃体硅油取出术+眼内人工晶状体二期置入
右叶甲状腺切除+左叶甲状腺部分切除术	单侧甲状腺切除伴他叶部分切除术
右眼白内障超声乳化抽吸术+人工晶体植入术	置入人工晶状体+白内障晶状体乳化和抽吸

数据集的详细统计信息如表 2 所示。通过观察表 2 可以发现,1 个手术原词最多可以对应到 7 个标准词。因此,相较于单纯的“一一对应”标准词归一化,本次评测中涉及到的任务更具难度和挑战。

表 2 评测数据分布统计

	训练集	验证集	测试集
数据条数	4 000	1 000	2 000
最大长度	144	120	93
平均长度	12.33	12.31	12.35
标准词平均个数	1.07	1.06	1.06
标准词最大个数	7	4	5
“一一对应”比例	0.918	0.913	0.927

① [https://en.wikipedia.org/wiki/Learning\\_to\\_rank](https://en.wikipedia.org/wiki/Learning_to_rank)

### 2.3 评测指标

任务以准确率 (accuracy) 作为最终评估标准, 准确率的定义为: 给出正确的手术原词加手术标准词的组合/待预测手术原词的总数, 形式化描述为: 对于一条手术原词  $S_i$ , 其对应  $N$  个归一化标准词。模型预测输出了  $M$  个归一化标准词, 则该条数据的得分如式(1)所示。

$$S_i = \frac{\text{Count}(N \cap M)}{\text{Max}(\text{Count}(N), \text{Count}(M))} \quad (1)$$

总得分如式(2)所示。

$$S = \frac{1}{k} \sum_{i=1}^k S_i \quad (2)$$

其中,  $k$  表示测试集中的手术原词条目数。

## 3 方法

本文的模型整体架构如图 1 所示。首先, 针对编码文件和标注文件分别建立索引; 然后, 对于给定的“手术原词”, 分别查询两个索引文件, 生成候选答案; 最后, 将各候选答案与手术原词一起, 输入 BERT 模型打分, 并基于 BERT 模型给出的分数输出“标准词”。接下来, 本文将分别介绍候选答案生成、候选答案打分、后处理和模型投票部分。

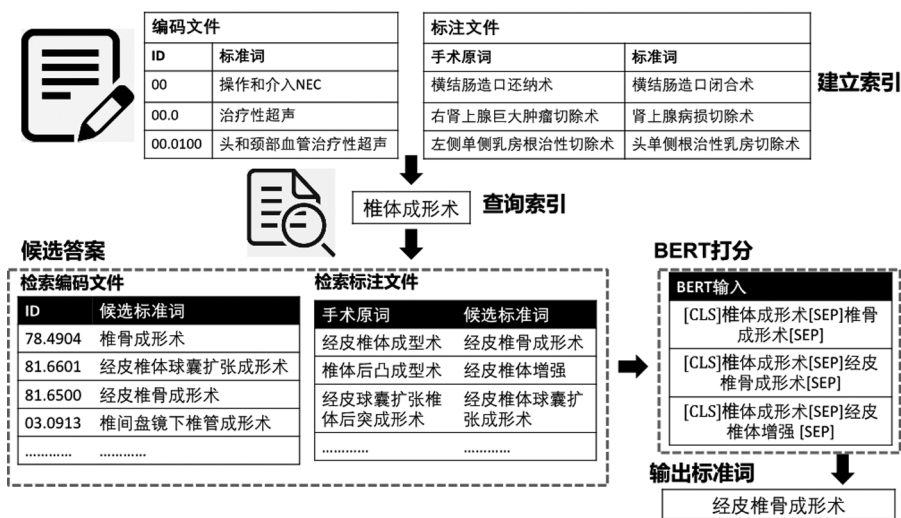


图 1 模型整体架构图

### 3.1 候选答案生成

本文使用 Lucene<sup>①</sup> 工具生成候选答案, Lucene 是一套用于全文检索和搜索的开放源码程式库, 其默认排序基于 TF-IDF 和向量空间模型, 可以方便而快速地找到与被检索短语在文字上相似的目标结果。本文设计了两种检索方式, 分别建立了索引。第一种方式是检索“编码—标准词”, 其目标是直接查找与待归一化的“手术原词”最相近的“标准词”。第二种方式是检索“标注历史”, 其目标是在标注数据上查找与待归一化的“手术原词”最相近的数据, 并取该条数据的归一化“标准词”作为候选答案。本文通过结合上述两种检索方式, 各取检索得分排名前 20 的“标准词”作为候选答案, 可以达到超过 99% 的覆盖率。

### 3.2 候选答案打分

通过 Lucene 检索得到候选标准词后, 本文基于 Transformer 框架对候选标准词进行打分。由于手术原词和标准词不易拆分, 如上文例子中提到的“多对一”情况, 将手术原词拆分后就无法对应到正确的标准词了。因此, 本文将该任务定义为: 某一个归一化标准词是否被包含在手术原词之中。即将整个手术原词和一个归一化标准词作为一条输入, 判别该归一化标准词是否应该出现在最终输出之中。本文使用 BERT 模型作为打分模型, 如图 2 所示。

对于“手术原词”和一个候选“标准词”, 本文按照 BERT 模型的规范, 将其按字分词并排列成“[CLS] 手术原词 [SEP] 标准词 [SEP]”的形式,

① <http://lucene.apache.org/>

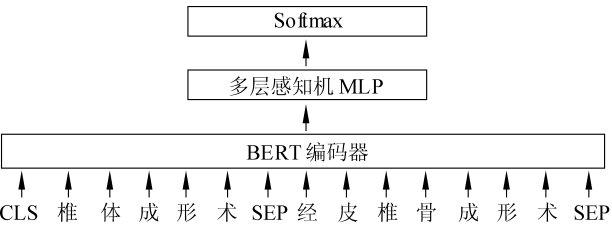


图 2 BERT 打分模型

输入 BERT 编码器。然后将 BERT 编码器的结果，即“[CLS]”位置处的表示，输入多层感知机中，得到二维向量。最后，通过 Softmax 归一化为 0 到 1 之间的概率。

3.3 后处理

针对模型的输出结果，本文采用了两种后处理方式：①若原词中不包含分隔符号“+”，则只输出得分最高的一个手术标准词；②若所有候选得分均未超过选定的阈值，则从候选标准词中输出得分最高的一个手术标准词。

3.4 模型投票

本文在提交结果时使用集成模型，按上述模型设定，共训练了 10 组模型。这 10 组模型的区别在于训练集的不同，因为评测任务参赛选手自由划分训练集和验证集，我们将验证数据集随机等分成三份，在某些实验组上会将验证集的两份加入到原始训练集中，作为最终的模型训练数据，将剩余的一份验证集作为最终的验证集。为了集成这 10 组模型，本文通过投票的方式来集成不同模型的输出，得到最终输出结果。对于每一个输入的手术原词，10 组模型分别输出各自的“标准词”预测结果，统计所有预测结果中各个“标准词”的出现次数，将出现次数最多即得票数最高的“标准词”作为最终预测结果，若有多个候选“标准词”的得票数相同，则这些候选词均会被输出作为预测结果。

4 实验结果及分析

4.1 实验结果

对于训练数据的构造和采样，本文尝试了多种方式，结果如表 3 所示。基于在验证集上的结果，最终本文选择“BERT-Base-Chinese<sup>①</sup>”初始化的“BERT-取前 20 候选+10 倍正例”的模型设定。

表 3 不同采样策略下的模型准确率

模型	准确率
BERT—取 Top10 候选	85.6
BERT—取 Top20 候选	87.5
BERT—取 Top20 候选+10 倍正例	89.1
BERT—取 Top20 候选+正确答案同子类的其他条目作为负例	88.6
RoBERTa—取 Top20 候选+10 倍正例	87.4

4.2 实验分析

4.2.1 候选答案生成

在候选答案生成中，我们结合了两种检索方式，表 4 展示了在验证集上两种方法的检索覆盖率。

表 4 检索方式覆盖率统计

Top	1	2	5	10	20	50
编码检索	49.5	57.4	68.8	77.4	86.0	91.7
历史检索	71.1	78.3	83.1	85.9	88.1	88.7
编码检索+历史检索	86.8	91.5	95.4	97.8	99.2	99.6

可以观察到，使用单一检索方式，都无法覆盖到全部数据。“编码检索”的缺点在于：对于与标准词在字面上差异较大的手术原词，无法检索到候选答案。“历史检索”的缺点在于：未在标注历史中出现过的归一化“标准词”，无法得到正确的候选答案。而通过结合上述两种方式，各取检索得分排名前 20 的“标准词”，就可以达到超过 99% 的覆盖率，基本上解决了候选答案生成问题。

在验证集上各取排名前 50 的检索结果也未能覆盖的几个例子如表 5 所示，从中可以观察到，这几个例子的“手术原词”和“归一化标准词”在字面上差距较大，并且存在无法解析的缩写，未来需要医学相关知识或更多的标注数据，以帮助模型得到更好的检索结果。

表 5 检索召回失败举例

手术原词	归一化标准词
左侧化疗泵置入+右侧脑室腹腔分流术	颅腔或组织的导管置入术

<sup>①</sup> [https://storage.googleapis.com/bert\\_models/2018\\_11\\_03/chinese\\_L-12\\_H-768\\_A-12.zip](https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip)



续表

手术原词	归一化标准词
LC	腹腔镜下胆囊切除术
原发性肝癌 HIFU 术	肝病损超声刀治疗
化疗泵置入 + 脑室腹腔分流术	颅腔或组织的导管置入术

#### 4.2.2 模型集成

本文用不同的训练数据训练了 10 组模型进行集成,为简单起见用数字 1、2、3 来表示验证数据集被随机分成的三部分,10 组模型在验证集上的结果如表 6 所示。

表 6 不同 stacking 策略下模型准确率

序号	训练集	验证集	准确率/%
1	训练集	验证集(全)	89.1
2	训练集+验证集(1,2)	验证集(3)	89.0
3	训练集+验证集(1,2)	验证集(3)	88.6
4	训练集+验证集(1,2)	验证集(3)	89.4
5	训练集+验证集(1,3)	验证集(2)	88.6
6	训练集+验证集(1,3)	验证集(2)	88.7
7	训练集+验证集(1,3)	验证集(2)	89.4
8	训练集+验证集(2,3)	验证集(1)	88.7
9	训练集+验证集(2,3)	验证集(1)	88.9
10	训练集+验证集(2,3)	验证集(1)	89.3

#### 4.3 实验参数设置

本文方法 BERT 中的参数均使用“BERT-Base-Chinese”中的原始参数。隐藏层的维度为 768, dropout 参数设置为 0.1, batch-size 大小设置为 64, 选择 Adam 作为优化器,学习率设置为 0.000 01, 训练轮次设置为 40。两种检索方式均取排名前 20 的检索结果作为候选答案。

### 5 错误分析

针对模型在验证集上预测错误的例子,我们进行了人工分析和归纳,模型预测的错误大致可以分为两类。

(1) 模型对“术式”“操作”和“部位”的特征学习得不充分,这类错误占据了 85.6%,如对手术原词“腹腔镜下胆囊取石术”,模型预测的最高分是“胆道

镜下胆管取石术”,而对应的标准术语应该是“腹腔镜下胆囊切开取石术”,在这个例子中,模型没有区分出“腹腔镜”和“胆道镜”这两个操作原语。

(2) 模型预测出来的标注词数量错误,这类错误占据了 3.8%,主要出现在手术原词和标准词个数不一致的情况下,模型未能很好地学习到数量上的映射。

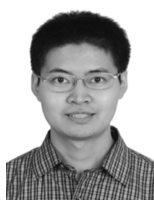
### 6 总结

在 CHIP-2019 评测任务 1 上,本文提出了一种比较通用的解决术语归一的框架:检索+重排序,检索部分采用了开源的 Lucene 对训练数据和标准答案集做检索生成候选答案;重排序部分本文采用了当下流行的 BERT 语言模型来辅助(手术原词,候选词)打分排序。本文提出的方法在测试集上单模型达到了 89.1%、融合模型达到 92.8% 的性能,证明了该方法的有效性。本文提出的方法具备较强的通用性,可同步迁移到医学领域其他类型术语的标准化任务上。同时本文提出的方法也有一定的局限性,如“错误分析”一节中所提到的,未能够利用医学领域知识,因此如何在模型中引入医学领域知识是未来重点的突破方向之一。此外,本文用到的 BERT 模型是官方开源的基础模型,本次任务由于时间关系以及医学文本的隐私性,我们未能对语言模型进行 fine-tuning 工作,医学领域的语言模型也是后续可尝试的方向。

### 参考文献

- [1] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, et al. Overview of the share/clefe health evaluation lab 2013[C]//Proceedings of International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013: 212-231.
- [2] Sameer Pradhan, Noémie Elhadad, Wendy Chapman, et al. SemEval-2014 task 7: Analysis of clinical text [C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 54-62.
- [3] Noémie Elhadad, Sameer Pradhan, Sharon Gorman, et al. SemEval-2015 task 14: Analysis of clinical text [C]//Proceedings of the 9th International Workshop on Semantic Evaluation, 2015: 303-310.
- [4] Luo YF, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization[J]. Jour-

- nal of Biomedical Informatics, 2019, 22: 103-132.
- [5] Ghiasvand O, Kate R J. UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 828-832.
- [6] Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text[J]. JAMIA, 2012, 20(5): 876-881.
- [7] Leaman R, Doğan RI, Lu Z. DNorm: Disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29: 2909-2917.
- [8] Luo Y, Song G, Li P, et al. Multi-task medical concept normalization using multi-view convolutional neural network[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence.
- [9] Zongcheng Ji, Qiang Wei, Hua Xu. BERT-based ranking for biomedical entity normalization[J]. arXiv preprint arXiv: 1908.03548, 2019.
- [10] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J].arXiv preprint arXiv: 1801.04805, 2018.



陈漠沙(1987—), 通信作者, 硕士, 高级算法专家, 主要研究领域为自然语言处理。

E-mail: chenmosha.cms@alibaba-inc.com



谭传奇(1992—), 博士, 算法专家, 主要研究领域为自然语言处理。

E-mail: chuanqi.tcq@alibaba-inc.com



仇伟(1987—), 硕士, 算法专家, 主要研究领域为自然语言处理。

E-mail: qiuwei.cw@alibaba-inc.com