文章编号：1003-0077(2021)03-0094-06

CHIP2019 评测任务 1 概述：临床术语标准化任务

黄源航[1]，焦晓康[2]，汤步洲[1,3]，陈清财[1,3]，闫 峻[2]，

(1. 哈尔滨工业大学 计算机科学与技术学院，广东 深圳 518055；
2. 医渡云(北京)技术有限公司，北京 100191；3. 鹏城实验室，广东 深圳 518055)

摘 要：第五届中国健康信息处理会议(China Conference on Health Information Processing，CHIP2019)组织了中文临床医疗信息处理方面的三个评测任务，其中任务 1 为临床术语标准化任务。该任务的主要目标是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。评测数据集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了标注。共有 56 支队伍报名参加了评测，最终有 20 支队伍提交了 47 组结果。该评测以准确率作为最终评估标准，提交结果中最高准确率达到 94.83%。

关键词：中国健康信息处理会议；临床术语标准化；自然语言处理

中图分类号：TP391 **文献标识码**：A

Overview of the CHIP2019 Shared Task Track1: Normalization of Chinese Clinical Terminology

HUANG Yuanhang[1], JIAO Xiaokang[2], TANG Buzhou[1,3], CHEN Qingcai[1,3], YAN Jun[2],

(1. School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China; 2. Yidu Cloud (Beijing) Technology Co., Ltd, Beijing 100191, China; 3. Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China)

Abstract: The 5th China Conference on Health Information Processing held a shared task including three tracks on Chinese clinical medical information processing. The first track is normalization of Chinese clinical terminology that assigns standard terminologies to surgical entities extracted from Chinese electronic medical records. All surgical entities in the Track1 dataset were collected from real medical data and annotated with standard surgical terminologies of "ICD9-2017 Clinical Edition". A total of 56 teams signed up for the track, and eventually 20 teams submitted 47 system runs. Accuracy is used to measure the performances of all systems, and the highest accuracy of all submitted system runs reached 0.9483.

Keywords: China Conference on Health Information Processing; normalization of Chinese clinical terminology; natural language processing

0 引言

随着信息技术的快速发展，计算机技术在医疗领域得到了广泛的应用。如何利用计算机相关技术处理海量的临床医疗数据是诸多学者一直在研究的热点问题。其中，临床术语标准化是临床医疗信息处理领域的一个重要研究方向。

临床上，由于医疗人员的记录风格存在差异，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。比如，在中文临床医疗文本中，“先天性脊柱侧弯”可以表述为“先天性脊柱侧凸”，也可以表述为“先天性脊柱侧弯畸形”；在英文临床医疗文本中，“heart attack”“MI”和“myocardial infarction”都可以代表“心肌梗塞”的含义。临床术语标准化就是要为临床上各种不同表述找到

收稿日期：2020-06-15 **定稿日期**：2020-08-21

基金项目：国家自然科学基金(61876052)；国家自然科学联合重点基金(U1813215)；广东省自然科学基金(2020KZDZX1222)；深圳市基础研究项目(JCYJ20190806112210067)

对应的标准表述。有了术语标准化的基础，研究人员才可以对临床医疗文本进行后续的统计分析。目前部分医疗机构采用人工方式将临床术语手动规范化为标准术语，但是由于临床术语专业性较强，并且表述方式过于多样，这种方式对工作人员专业知识要求较高，所需人力成本巨大，得到的标准化结果也往往不够准确。

第五届中国健康信息处理会议（CHIP2019）组织了与中文临床医疗信息处理相关的三项评测任务，其中评测任务 1 为临床术语标准化任务。该评测任务鼓励参赛者使用计算机技术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，即给定一手术原词，将其自动映射为手术词表中对应的手术标准词。本次评测数据集由医渡云（北京）技术有限公司提供，其中的手术原词全部来自真实医疗数据。训练数据由专业人员依据《ICD9-2017 协和临床版》手术词表对手术原词进行了人工标注，将手术原词手动映射为手术词表里的手术标准词，标注样例如表 1 所示，其中多个标准词用“＃＃”分隔。参赛队伍需要构建系统将测试数据中的手术原词映射到给定手术词表里的手术标准词。本次评测以准确率（accuracy）作为评估指标。最终，排名第一的参赛队伍提交结果的准确率为 94.83％。本文将对此次评测任务中的数据、各支队伍的提交结果以及评价指标进行分析和总结。

表 1　CHIP2019 评测任务 1 标注数据样例

原始词	标准词
横结肠造口还纳术	横结肠造口闭合术
右肾上腺巨大肿瘤切除术	肾上腺病损切除术
左侧单侧乳房根治性切除术	单侧根治性乳房切除术
经皮三叉神经半月节射频热凝术	三叉神经半月节射频热凝术
右肾探查，右肾根治术	肾探查术＃＃单侧肾切除术

1　相关工作

临床术语表述方式的不统一给医疗信息的整合、交换和共享等工作带来了诸多障碍。因此，开展临床术语标准化的相关研究有助于推动医疗领域数字化、信息化建设，实现高效率的全社会医疗资源共享。国外对于临床医学术语标准化的探索起步较早，目前已经做了许多研究工作。MetaMap 是美国国立医学图书馆建立的一个实现生物医学文本到一体化医学语言系统（unified medical language system，UMLS）概念映射的在线工具，它能标记出生物医学文本所包含的 UMLS 超级叙词表（Metathesaurus）中的医学概念。Aronson[1] 对 MetaMap 的文本映射基本原理进行了描述，即对于医疗文本，MetaMap 使用基于规则的方法，通过计算文本中的名词短语与检索 Metathesaurus 得到的候选词之间的匹配程度来查找并返回与此文本相关的 Metathesaurus 概念。然而，这种简单的字符串匹配方法对数据要求较高，泛化能力不强。Leaman 等人[2] 提出了一个利用机器学习方法对医疗文本中的疾病名称进行标准化的模型 DNorm。DNorm 模型使用机器学习中的文档对排序学习（pairwise learning to rank）技术对文本中发现的疾病名称和知识库中的实体概念进行相似度打分并排序，最终返回分数最高的候选概念或能够在词表中完全匹配的候选概念作为疾病名称标准化后的标准概念。DNorm 当时在公开数据集 NCBI 上达到了最好的效果，但该模型在计算相似度时并没有深入挖掘文本中所包含的语义信息。随着深度学习技术的发展和计算性能的大幅提升，神经网络被广泛应用于医疗信息处理领域。Limsopatham 和 Collier[3] 提出使用卷积神经网络（convolutional neural network，CNN）或者长短期记忆网络[4]（long short-term memory，LSTM）对社交媒体中的文本进行编码，把每个医学概念看作一个类别，将编码后的文本表示经过分类器映射到对应的医学概念上。这是深度学习技术首次被应用到医学术语标准化任务中，相比传统的字符串匹配或者机器学习方法，深度学习技术能够更好地利用文本中所包含的语义信息。近些年来，国际上组织了多个与临床术语标准化相关的评测任务，比如 CLEF（Conference and Labs of the Evaluation Forum）eHealth 2017[5]、eHealth 2018[6] 和 eHealth 2019[7] 中的多语言信息抽取任务，SMM4H（social media mining for health）2019[8] 中的药物副作用抽取以及标准化任务，BioNLP（Biomedical Natural Language Processing Workshop）2019[9] 中的药品和化学实体标准化子任务。

由于我国医疗信息化发展进程相对滞后，且医疗术语相关编码体系建设起步较晚，目前国内关于中文临床术语标准化开展的研究较少。CHIP2019

评测任务 1 是国内首个聚焦于中文临床术语标准化工作的评测，旨在利用前沿的深度学习和自然语言处理技术，推动临床术语标准化的相关研究工作。

2 评测数据

CHIP2019 评测任务 1 数据集中包含的所有手术原词均是来自三甲医院的真实医疗数据，由医渡云(北京)技术有限公司提供。训练集和验证集中分别包含了 4 000 条和 1 000 条手术原词，对每条手术原词以《ICD9-2017 协和临床版》手术词表为标准进行了标注，形成＜原始词，标准词＞对。《ICD9—2017 协和临床版》手术词表为层级结构，层级越深，标准词表述越具体，因此对于手术原词，标注原则为尽量查找层级深的标准词，无法准确对应标准词时再查找上级标准词。

测试集中包含了 2 000 条手术原词，要求参赛者给出其对应的手术标准词。表 2 展示了评测数据集的具体统计信息。除了训练集、验证集和测试集，此次评测还提供了《ICD9-2017 协和临床版》手术词表，里面包含了 9 867 个手术标准词。

表 2 CHIP2019 评测任务 1 数据集统计信息

	训练集	验证集	测试集	手术词表
数据量	4 000	1 000	2 000	\
最大手术原词长度	122	102	80	\
最小手术原词长度	2	2	3	\
平均手术原词长度	12.36	12.34	12.35	\
手术原词对应最多标准词个数	7	4	5	\
手术原词对应最少标准词个数	1	1	1	\
手术原词对应平均标准词个数	1.07	1.06	1.06	\
最大手术标准词长度	24	24	27	37
最小手术标准词长度	3	4	3	1
平均手术标准词长度	9.14	9.24	9.16	9.16

训练集、验证集和测试集的数据分布基本保持一致。长度不超过 20 的手术原词在训练集中占比约为 94%，在验证集和测试集中将近 95%，可以看出此次评测数据集中的手术原词均为短文本。由于训练集、验证集和测试集里面出现的手术标准词均来自《ICD9-2017 协和临床版》手术词表，因此手术标准词长度分布和手术词表基本一致。由于数据集中的每条手术原词可能对应多个手术标准词，如手术原词“经皮肾镜碎石取石术(左侧)”对应的手术标准词为“经皮肾镜碎石术(PCNL)”和“经皮肾镜取石术”。这也是本次评测任务的难点之一。在训练集、验证集和测试集中，只对应一个手术标准词的手术原词占比大约为 95%，即大部分手术原词对应单个标准词。关于数据集中手术原词对应手术标准词个数的具体统计信息如图 1 所示。

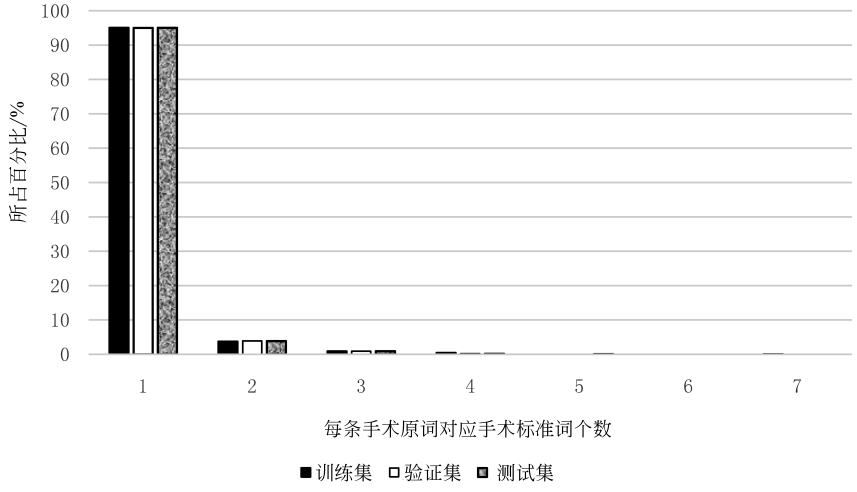

图 1 CHIP2019 评测任务 1 手术原词对应标准词个数统计信息

3　评估指标

CHIP2019 评测任务 1 以准确率(accuracy，A)作为最终评估标准。本任务中，准确率定义：给出正确的手术原词加手术标准词的组合除以待预测手术原词的总数。具体计算如式(1)所示。

$$A=\frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{\max(|P_i|,|G_i|)} \quad (1)$$

对于测试集中的第 i 条手术原词，预测出的手术标准词集合为 P_i，真实对应的手术标准词集合为 G_i，N 是测试集中手术原词的个数。需要说明的是，计算 P_i 和 G_i 的交集时，遵循严格匹配原则，即预测出的手术标准词必须与手术原词真实对应的某个手术标准词完全匹配，否则不属于交集。

4　评测结果

CHIP2019 评测任务规定每支参赛队伍最多提交三份结果，取准确率最高值作为该支队伍的最终结果。本次评测共收到了 47 份提交结果，其中准确率最高为 94.83%，最低为 35.11%，平均值为 79.75%。表 3 为对评测任务 1 排名前十参赛队伍系统的简要描述，包括所使用的技术方法简述以及是否使用外部资源(此评测任务允许使用公开数据资源和选手个人/组织的内部资源，不允许执行任何人工标注)。

表 3　CHIP2019 评测任务 1 排名前十队伍系统信息

排名	队伍名	技术方法简述	外部资源
1	YZS-cwf	应用深度学习模型 BERT。术语归一化模块采用深度学习模型 BERT 打分；数量预测模块采用深度学习模型 BERT 做多标签分类。根据数量预测模块的结果，取 top$k$ 个标准名称作为输出	无
2	ALBB-cms	IR＋Rerank 的 pipeline，IR 使用的是 Lucene 工具，Rerank 部分使用的是基于 BERT-base-Chinese 的简单 MLP 网络	无
3	ZKYF-xl	使用深度学习方法进行词条标准化。具体为：①利用 Transformer 网络结构，对给定的训练数据＋标准词表进行端到端的翻译模型训练(共训练字到字、字到词、词到词三种)；②利用给定的训练数据对中文 BERT 预训练模型进行 finetune，训练出相似度打分模型；③对于给定的待标准化原词，首先利用步骤①中训练的三种端到端翻译模型分别进行标准词以及其所隶属类别的输出，然后利用输出结果结合文本相似度算法 MetricLCS 在标准词库中进行相似词候选集筛选，最后将筛选出的候选集交由步骤②训练的相似度打分模型进行重打分，选出得分最高词作为最终标准词进行输出	无
4	DLLGDX-syj	用简单的相似度计算方法挑选候选词，使用 BERT 中文模型进行相似度计算，结合了人工规则对多标签数据进行处理	无
5	XHYXY-yfh	首先依据训练集和手术词表建立标准词到手术原词的映射表，依据映射表进行采样用于训练 BERT 模型；预测时，计算测试集手术原词与映射表手术原词的相似度，排序取 top5。然后，利用 BERT 模型判断 top5 对中两个手术原词的标签是否为 1，为 1 则取映射表手术原词对应的标准词为最终答案返回	无
6	GR-hwj	①把整个问题看作是 Seq2Seq 的类似问题，然后借鉴 VAE 的思想对模型的隐层进行 finetune，训练得到一个模型；②把它看成一个分类问题，将手术词表中所有标准词构建成一个对应标签的数据点进行数据增强，然后构建一个分类模型 BERT＋LSTM；③将上述两个模型的结果进行简单的线性组合	无
7	SRBL-plb	基于 n-gram 相似度等规则匹配方法筛选训练样本，使用 BERT 网络进行训练，使用了 5 折交叉验证方法，最终使用 5 个模型的平均分数	无
8	SXDX-djj	Seq2Seq＋attention＋后处理(如果输入在手术词表中，则跳过预测)	无
9	YXKXY-lxy	综合了特征分析、编辑距离、同义词词典和文本相似度的语义相似度算法。(外部资源为中文医学主题词表 CMeSH、中文一体化医学语言系统 CUMLS、ICD9CM 英文版。)	有
10	YXKXY-zwq	使用相似度＋BERT，用 bert as service 生成原始词的句向量，计算相似度匹配	无

可以看出，本次评测排名前十的大部分参赛队伍都使用了预训练语言模型 BERT[10]（bidirectional encoder representation from transformers），证明了利用大规模语料进行预训练得到的语言模型在自然语言处理领域的有效性。除了“GR-hwj”和“SXDX-djj”两支队伍，其余八支队伍均将标准化任务当成了文本相似度任务进行处理。本质上，临床术语标准化任务是语义相似度匹配任务的一种。但是由于原词表述方式过于多样，并且标准词词表通常规模较大，单一的匹配模型很难获得很好的效果。部分队伍除了文本匹配模块，还加入了筛选匹配候选词模块，即通过相似度计算等手段为每个手术原词筛选若干手术标准词作为匹配候选，再将＜手术原词，手术标准词＞对输入到文本匹配模型进行关系判断。相似度计算是较为常见的筛选候选方法，“YZS-cwf”采用多标签分类模型获取匹配候选；“ALBB-cms”引入信息检索技术，使用检索工具 Lucene 得到匹配候选；“ZKYF-xl”借鉴了生成模型的思想，利用 Transformer[11] 训练了端到端的翻译模型辅助相似度计算。“GR-hwj”将临床术语标准化任务分别当作生成问题和分类问题进行处理，融合了生成模型和分类模型的结果。“SXDX-djj”则完全将临床术语标准化任务当作生成问题，以 Seq2Seq[12] 模型为基础构建系统进行标准词预测。

测试集中一共有 2 000 条手术原词，其中 1 901 条手术原词对应单个手术标准词，99 条手术原词对应多个手术标准词。表 4 是评测任务 1 排名前十队伍的系统在测试集上的评估结果，每列结果最高数值已加粗表示。其中“Acc”为在测试集整体 2 000 条数据集上计算的准确率结果，“Acc-single”为在测试集中对应单个手术标准词的 1 901 条手术原词上计算的准确率结果。“Acc-multiple”则是在测试集中对应多个手术标准词的 99 条手术原词上计算的准确率结果。可以看出：此次评测排名前十队伍对于单个手术标准词的预测准确率较高，前五支队伍均接近 0.9 或者达到 0.9 以上，前十队伍中“Acc-single”最高值与最低值之差为 0.217 0。对于对应多个手术标准词的情况，前十队伍系统预测结果比对应的单个手术标准词差，最高值为 0.888 9，最低值为 0.491 3，二者相差 0.3976，说明各支队伍系统在预测多个手术标准词时性能差异相对较大。

针对测试集中的 2 000 条手术原词，其中前十支队伍均没有预测正确的手术原词，一共有 52 条，可以将这些手术原词看作系统普遍预测错误的数据样例。通过分析这些数据样例，发现不易准确预测的手术原词可以主要归为以下三类：

（1）手术原词对应多个手术标准词，这与上文展示的各支队伍对于多标准词的预测结果相符。

（2）手术原词对应的标准词在训练集出现次数很少甚至没有出现过，这给模型的训练增加了难度，模型无法准确地学习到这些标准词的相关信息。

（3）手术原词对应的手术标准词在手术词表中有相似的标准词，模型在预测时难以有效地分辨这些相似标准词的区别。

表 4 CHIP2019 评测任务 1 排名前十队伍系统评估结果

排名	队伍名	Acc	Acc-single	Acc-multiple
1	YZS-cwf	**0.948 3**	**0.951 3**	**0.888 9**
2	ALBB-cms	0.927 2	0.938 8	0.703 6
3	ZKYF-xl	0.913 4	0.927 9	0.634 7
4	DLLGDX-syj	0.888 5	0.902 2	0.627 1
5	XHYXY-yfh	0.885 2	0.896 9	0.660 3
6	GR-hwj	0.849 4	0.865 9	0.532 5
7	SRBL-plb	0.841 8	0.864 3	0.409 4
8	SXDX-djj	0.827 7	0.847 7	0.443 1
9	YXKXY-lxy	0.775 8	0.786 2	0.577 3
10	YXKXY-zwq	0.722 2	0.734 3	0.491 3

5 结语

临床术语标准化是医疗信息处理领域中的一个重要研究方向。在如今互联网、大数据迅速发展的时代背景下，术语标准化有助于整合和利用规模庞大的、分散的、非结构化的医疗信息数据。随着人工智能的兴起，自然语言处理相关技术的应用逐渐渗透到医疗领域。如何利用自然语言处理等深度学习技术处理临床术语标准化问题，是 CHIP2019 评测任务 1 关注的重点。

本文是对 CHIP2019 评测任务 1 的简要概述。本次评测吸引了来自企业、高校和研究机构的 56 支队伍报名参加，一共接收了 47 组结果，最高准确率达到了 94.83%。参赛队伍大多数以预训练语言模型 BERT 为基础构造了系统。相比于未引入预训练模型的系统，这些以预训练语言模型为基础的系统取得了较好的标准化效果。大多数系统对于对应

单标准词的手术原词预测效果较好，对于对应多标准词的手术原词预测效果相对较差。通过分析预测错误的数据，总结了出错的主要类型，这是临床术语标准化任务的主要难点，也是未来研究工作中应该关注的重点。

参考文献

[1] Aronson A R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program [C]//Proceedings of the American Medical Informatics Association Symposium, 2001: 17-21.

[2] Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.

[3] Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1014-1023.

[4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] Neveol A, Robert A, Anderson R, et al. CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French[C]//Proceedings of the Workshop of the Cross Language Evaluation Forum, 2017:1-17.

[6] Neveol A, Robert A, Grippo F, et al. CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian[C]//Proceedings of the Workshop of the Cross Language Evaluation Forum, 2018: 1-18.

[7] Dorendahl A, Leich N, Hummel B, et al. Overview of the CLEF eHealth 2019 multilingual information extraction[C]//Proceedings of the Workshop of the Cross Language Evaluation Forum, 2019: 1-9.

[8] Weissenbacher D, Sarker A, Magge A, et al. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019[C]//Proceedings of the 4th Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, 2019: 21-30.

[9] Agirre A G, Marimon M, intxaurrondo A, et al. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track [C]//Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, 2019: 1-10.

[10] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.

[11] 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, 2017: 5998-6008.

[12] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, 27: 3104-3112.

黄源航(1993—)，硕士研究生，主要研究领域为自然语言处理。
E-mail: 18S051003@stu.hit.edu.cn

汤步洲(1984—)，通信作者，博士，副教授，博士生导师，主要研究领域为人工智能，自然语言处理，医学信息学。
E-mail: tangbuzhou@gmail.com

焦晓康(1992—)，硕士研究生，主要研究领域为医疗文本的自然语言处理。
E-mail: xiaokang.jiao@yiducloud.com