融合相似度算法与预训练模型的中文电子病历实体映射方法研究*

冯凤翔 任慧玲 李晓瑛 王巍洁 王 勖 张 颖

（中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100020）

〔摘要〕 采用自标注中文电子病历标准数据集，融合相似度算法与预训练模型并分别应用于实体映射的候选实体生成和实体消歧阶段，对不同相似度算法和预训练模型的性能进行比较分析。提出基于别名间相似性改进药物类实体映射效果的方法，结合 Jaccard 相似度算法与 BERT 预训练模型，高效实现海量中文电子病历实体映射任务。

〔关键词〕 实体映射；实体标准化；相似度算法；电子病历；BERT 模型

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.05.008

Study on Chinese Electronic Medical Record Entity Mapping Method by Fusing Similarity Algorithms and Pre－trained Models

FENG Fengxiang, REN Huiling, LI Xiaoying, WANG Weijie, WANG Xu, ZHANG Ying, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕 The self－annotated Chinese electronic medical record (EMR) standard datasetisused, the similarity algorithms and pre－trained models are fused and applied to the candidate entity generation and entity disambiguation stages of entity mapping, and the performance of different similarity algorithms and pre－trained models is compared and analyzed. A method is proposed to improve the mapping effect of drug class entities based on alias similarity, and the Jaccard similarity algorithm and BERT pre－trained model are combined to efficiently realize the task of mapping the entities of massive Chinese EMRs.

〔**Keywords**〕 entity mapping; entity standardization; similarity algorithm; electronic medical record (EMR); BERT model

1 引言

随着计算机技术和生物技术的飞速发展，电子病历（electronic medical record，EMR）、医学报告、医学文献等生物医学文本的数量迅速增长，积累了海量有价值的医学数据[1]，如 EMR 记录患者全部诊疗过程，包括所患疾病、药物、检查和治疗结果等[2]。随着命名实体识别技术成熟，可实现从中抽取疾病、药物、手术操作等特定实体[3]，但实体存在表述口语化、多样化等问题，如“狼疮脑炎”可能被表述为“狼疮脑病”，“氯硝柳胺”可能被写作“灭绦灵”等。如果未经处理就加以利用或者储存入库可能导致各医疗信息系统标准不一，难以实现医院间资源互联互通[2]。因而需要根据实体在文中

〔修回日期〕 2023－02－20

〔作者简介〕 冯凤翔，硕士研究生，发表论文 1 篇；通信作者：任慧玲，研究馆员，硕士生导师。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大专项课题“中文医学术语体系构建”（项目编号：2020AAA0104901）。

语义表达将其映射到知识库中对应的标准实体上，以解决概念内涵不清、语义表达和逻辑不一致等问题，促进独立医疗信息系统间的互操作，实现医疗信息和数据共享[4]。

目前中文电子病历实体映射研究主要由中国健康信息处理会议（China Health Information Processing Conference，CHIP）中的临床术语标准化评测任务推动，该任务使用来自真实医疗数据中的手术原词，由专业人员依据《ICD 9—2017 协和临床版》手术词表进行训练数据标注。目前已有多支队伍完成手术实体映射任务，CHIP 2019 任务最佳结果 *F*1 值为 94.83%，由 Dolphin - ICI 团队取得[5]。实体映射作为一项重要的生物医学技术，近年来取得较大进步。但由于生物医学领域本身的复杂性以及真实医疗应用场景的高标准、严要求，该领域研究还存在一些提升空间，如现有研究大都以预训练语言模型 BERT 为基础进行任务构造，相关研究主要围绕 CHIP 任务中标注的手术实体进行[6]，还需增加基于多类型实体的映射方法研究、改进方法模型以提升实体映射效果。

本文依靠专家知识与标准词表，标注涵盖多种类型实体的自标注标准数据集。结合传统文本匹配方法与深度学习的优势，提出融合传统相似度算法与深度学习模型的联合模型改进实体映射效果。并在此基础上提出基于别名间相似性映射标准实体的方法，为提升较短字符类实体映射效果提供思路。本文提出的方法在自标注数据集中 *F*1 值达到 94.87%，可帮助抽取和规范化医疗数据以扩充中文临床医学术语体系的实体覆盖度，更好地支持临床辅助诊疗系统、精准医学研究和疾病监控等应用。

2 研究内容

2.1 数据来源

从“爱爱医”“众意好医师”“病历网”中抽取 20 000 余条中文电子病历，采用命名实体识别模型对疾病诊断、解剖部位、临床检查、手术操作和药物 5 类实体进行识别[7]。每类实体分别抽取约 500 个样例，邀请 3 位临床专家根据《国际疾病分类第 10 次修订本》（International Classification of Disease V 10，ICD 10）和《常用临床医学名词（2019 年版）》，人工标注标准实体，数据清洗后，形成共包含 2 390 对数据的自标注标准数据集，训练集与测试集数据按照 7∶3 的比例进行分配，见表 1。

表 1 自标注标准数据集内容统计

实体类型	数据量	训练集	测试集
疾病诊断	496	348	148
解剖部位	483	338	145
临床检查	431	302	129
手术操作	505	353	152
药物名称	475	332	143
合计	2 390	1 673	717

2.2 实验设计

实体映射任务可分成候选实体生成（candidate entity generation，CEG）和实体消歧（entity disambiguation，ED）两个阶段，为此设计两个实验。

实验 1 为候选实体生成实验，即为待标准实体找到标准词表中所有可能与之对应的实体，生成候选标准实体集合，高效简便的相似度算法可以较好地完成这一任务。具体而言，针对每个待标准实体，计算其与标准词表中每个标准实体的文本相似度并排序，选取结果中前 15 位作为候选标准实体集合。采用目前应用较广泛的中文短文本相似度算法进行实验，对召回率进行比较以选取最佳算法，包括 Cosine 余弦相似度、BM 25（best match，BM）相关性评分、Jaccard 相似系数、编辑距离（minimum edit distance，MED）与 Dice 系数。

实验 2 为实体消歧实验，生成候选标准实体集后，需要从中预测最可能与待标准实体相对应的实体作为映射结果，这一过程便是实体消歧。将该任务视为一个二分类任务，采用深度学习模型进行预训练，学习实体间语义关系并输出预测结果。谷歌推出的 BERT 模型能结合文本上下文信息并从中提取丰富全面的语义特征，出色完成二分类任务，在其基础上还衍生出如 RoBERTa、ALBERT、RoBERTa - wwm 等在各项自然语言处理任务中取得较好效果的

模型。采用以上 4 种模型进行实体消歧，并比较模型性能，得到表现最好的算法模型组合，见图 1。

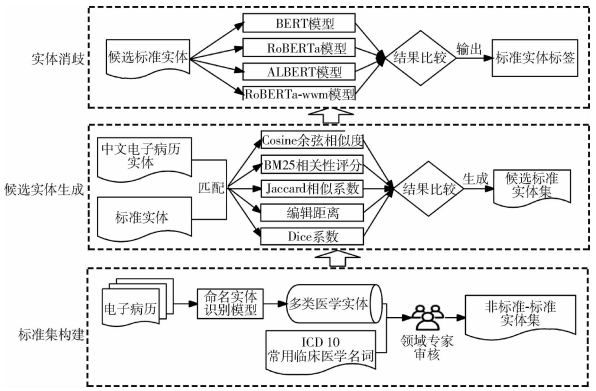

图 1 本文技术路线

2.3 评价指标

采用准确率（accuracy，A）、精确率（precision，P）、召回率（recall，R）和 *F*1 值（*F*1 score）作为评价指标。准确率指所有预测正确的结果所占比例，精确率指预测正确的正例占所有被预测为正例的比例，召回率指所有正例样本中预测正确的样本所占比例。*F*1 值用来衡量二分类模型精确度，可以看作是模型精准率和召回率的调和平均，其最大值是 1，最小值是 0。预测结果有真阳性（true positive，TP），假阳性（false positive，FP），真阴性（true negative，TN），假阴性（false negative，FN）4 种类型。其中，真阳性指预测结果为正，实际亦为正，其他同理。各指标计算方式如下：

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2P \times R}{P + R} \tag{4}$$

3 相关算法与模型

3.1 相似度算法

3.1.1 Cosine 余弦相似度 通过将文本矢量化，计算同个向量空间中两个向量夹角间的余弦值 Cos（θ）来衡量相似度大小，夹角越小则相似度越高。在生成每个文本矢量时根据词频－逆向文件频率（term frequency－inverse document frequency，TF－IDF）原理，用词频向量 V 代替文本矢量每个维度的值[8]：

$$\mathrm{Cos}(\theta) = \frac{V_1 \cdot V_2}{V_1 \times V_2} \tag{5}$$

3.1.2 BM 25 相关性评分 主要原理是先对句子 q 分词，生成若干特征词 q_i。W_i 为每个 q_i 的权重。之后对要与句子 q 进行比较的句子 D 计算每个 q_i 与 D 的相关性得分，最后将 q_i 相对于 D 的相关性得分进行加权求和，得到 q 与 D 的相关性得分[9]：

$$\mathrm{score}(q,D) = \sum_i W_i \cdot \mathrm{R}(q_i,D) \tag{6}$$

3.1.3 Jaccard 相似系数 用于比较两个有限样本集之间的相似性，两个集合 A 和 B 交集元素的个数在 A、B 并集中所占比例称为这两个集合的 Jaccard 系数，Jaccard 系数值越大则相似度越高[10]：

$$\mathrm{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

3.1.4 编辑距离 是对两个字符串差异程度的测量，其原理是计算一个字符串 S_1 经过多少次处理才能变成另一个字符串 S_2，允许的编辑操作有替换一个字符、插入一个字符、删除一个字符 3 种[11]：

$$\mathrm{Sim}(S_1,S_2) = 1 - \frac{d(S_1,S_2)}{\max(\mathrm{len}(S_1),\mathrm{len}(S_2))} \tag{8}$$

其中 len 代表字符串的字符个数，d（S_1，S_2）为字符串 S_1 和 S_2 的编辑距离。

3.1.5 Dice 系数 是一种集合相似度度量指标，可将字符串理解为集合，因此 Dice 系数也可用于计算两个字符串的相似度，其范围为 0～1，值越大表示相似度越高[12]。对于给定字符串 S_1 和 S_2，其 Dice 系数算法如下：

$$\mathrm{Dice}(S_1,S_2) = \frac{2|S_1 \cap S_2|}{|S_1| + |S_2|} \tag{9}$$

3.2 深度学习模型

3.2.1 BERT 是一个无监督双向模型，可以使用纯文本语料库进行训练，并预测受左右上下文制约的单词。BERT 能够在各种自然语言处理任务中表现出优异的性能[13]。经过调整，BERT 可用于进行二分类任务，其模型原理，见图 2。其中［CLS］

位于开头，用于预测文本类别，［SEP］用于分割两个句子，ClassLabel 为模型输出的类别标签。

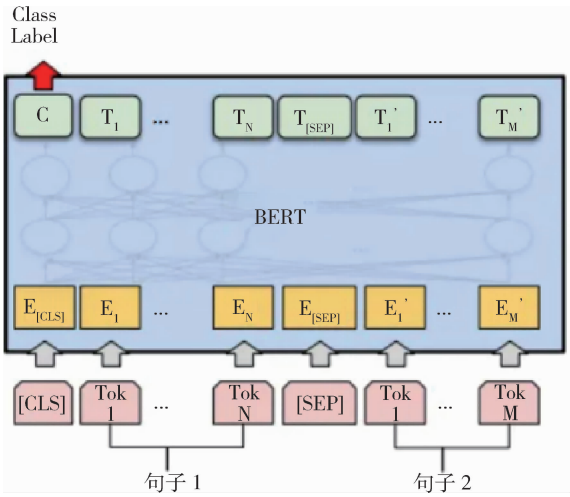

图 2 BERT 模型二分类任务原理

3.2.2 RoBERTa 是在 BERT 预训练模型基础上进行改进的复制研究，其中包括对超参数调整和对训练集大小的影响进行仔细评估。同时该任务使用新数据集 CCNEWS，并使用更多数据进行预训练，进一步提高下游任务性能[14]。

3.2.3 ALBERT 采用因子嵌入参数化和跨层参数共享两种参数简化技术解决 BERT 内存消耗高和训练速度慢的问题。此外，模型还提出句子顺序预测（sentence order prediction，SOP）任务来代替传统的下一句预测（next sentence prediction，NSP）预训练任务，从而获得更好的性能[15]。

3.2.4 RoBERTa－wwm 由哈工大讯飞联合实验室发布，结合中文全词掩码（whole word masking，wwm）技术与 RoBERTa 模型的优势，在 RoBERTa 模型基础上采用全词掩码来屏蔽汉语单词进行预训练。由于这些模型将应用于中文文本分类，全词掩码将允许这些模型结合中文场景，以提供更多有关中文语义的信息[16]。

4 实验结果及分析

4.1 候选实体生成实验

为对比分析不同相似度算法对候选实体生成效果的影响，基于 5 种相似度算法进行候选实体生成，选取相似度概率值前 15 位作为候选实体，并计算其召回率，见表 2。为方便理解，下文待标准实体称为原始词，标准实体称为标准词。从结果中可以看出，基于 Jaccard 相似度算法匹配的效果优于其他 4 种算法，因此选择 Jaccard 算法进行候选实体生成能较好地保证候选标准实体集质量，有利于进行后续实体消歧工作。Cosine 余弦相似度算法和 BM25 算法的生成效果明显低于其他相似度算法，是因为余弦相似度算法依赖高质量的词向量，BM25 算法需要精准分词，本文并未人工进行特征构建，因此二者效果较差。

表 2 候选实体生成实验召回率统计结果

相似度算法	召回率
Cosine	0.774 5
BM 25	0.804 3
Med	0.890 8
Jaccard	0.913 5
Dice	0.907 8

从本实验中 Jaccard 算法未召回准确标准词的样例中可以看出，对字数过短或者原始词与标准词之间差异较大的实体，单纯使用相似度算法无法获取词中的语义信息，会导致召回失败，见表 3。

表 3 Jaccard 算法未召回标准词样例

分类	原始词	标准词	是否召回	第 1 候选标准词
疾病诊断	冯雷克林霍增氏病	多发性神经纤维瘤病	FALSE	克山病
解剖部位	配偶子	生殖细胞	FALSE	精子
临床检查	天疱疮抗体检测	抗表皮细胞间质抗体	FALSE	抗透明带抗体检测
手术操作	颅内硬膜下冲洗术	脑膜切开术	FALSE	胸腔镜下肺叶切除术
药物	安宝	盐酸利托君	FALSE	安乃近

4.2 实体消歧实验

4.2.1 模型训练集构建 基于预训练模型的实体消歧需要构建包含正负例的训练集用于训练模型。根据相似度算法性能对比实验，基于 Jaccard 相似度算法，计算原词 i 与标准词词表中的每个标准词的 Jaccard 相似度系数，取相似度值前 15 位作为候选词表，将标准词 i 从候选词表中去除后剩下的候选标准词 j 用于构建负例[17]。训练集正例为：<原词 i，标准词 i，1>。负例为：<原词 i，标准词 j，0>。由于正样本数较少，与负样本数相差较大，为了扩充正样本，构建 10 *（标准 - 原始 + 原始 - 标准）共 33 460 条正样本。

4.2.2 实验环境及参数 本实验代码使用 Python 3.6 和 Tensorflow 1.8 编写，模型详细训练参数包括学习率（learning_rate）、每个样本处理成的长度（pad_size）、隐藏层中节点个数（hidden_size）、单次训练选取样本数（batch_size）、样本训练轮次（num_epochs），见表 4。

表 4 预训练模型参数设置

参数名	learning_rate	pad_size	hidden_size	batch_size	num_epochs
BERT	0.000 05	64	768	64	50
ALBERT	0.000 05	64	312	64	50
RoBERTa	0.000 05	64	312	64	50
RoBERTa - wwm	0.000 05	16	1 024	64	50

4.2.3 实体消歧实验结果 在基于 Jaccard 算法生成候选实体集的基础上，选取 4 种在二分类任务中表现较好的深度学习预训练模型作为基准进行实验，以对比分析不同模型的性能，见表 5。从结果可知，采用 RoBERTa - wwm 模型的准确率最高，ALBERT 模型的精确率和 $F1$ 值最高，BERT 模型的召回率最高，RoBERTa 模型的各项指标效果都接近理想，因此难以确定最佳模型。同时还注意到，4 个模型整体的准确率、精确率和 $F1$ 值仍不太理想。对输出结果文档进行分析后发现除药物分类之外的其他分类各项指标均已大于 86%，但是药物分类的原始词和标准词之间差异性过大，导致召回率较低，还需对此问题进行解决。

表 5 实体消歧实验结果指标统计

相似度算法 + 预训练模型	准确率	精确率	召回率	$F1$
Jaccard + BERT	0.851 1	0.645 9	0.990 1	0.781 8
Jaccard + ALBERT	0.845 4	0.770 0	0.975 5	0.860 7
Jaccard + RoBERTa	0.848 2	0.710 2	0.981 9	0.824 2
Jaccard + RoBERTa - wwm	0.852 5	0.487 8	0.955 5	0.645 9

4.2.4 优化后实体消歧实验结果 针对药物原始词和标准词差异过大问题，提出通过别名间相似性来进行知识补全从而提高实体映射效果的方法。标准词的别名间大多存在字符相似关系，如“甲硝唑”的别名有“灭滴灵”和“灭滴唑”，“布地奈德”的别名有“普米克”和“普米克令舒”。因此，本文针对药物实体采集第 2 批共 9 481 条标准实体 - 别名语料，构建药品别名映射库，增加药品原始词与药品标准词别名的匹配，若匹配到药品别名，将其链接至标准词。从结果可知，方法优化后，4 个模型组合的实验结果均有较大提升，召回率均超过 98%。采用 Jaccard + BERT 方法进行实体映射的各项指标均优于其他模型，$F1$ 值达到 94.87%，见表 6。

表 6 优化后实体消歧实验结果指标统计

相似度算法 + 预训练模型	准确率	精确率	召回率	$F1$
Jaccard + BERT	0.905 0	0.902 4	0.999 6	0.948 7
Jaccard + ALBERT	0.870 9	0.880 9	0.998 4	0.936 0
Jaccard + RoBERTa	0.887 9	0.886 7	0.998 4	0.939 2
Jaccard + RoBERTa - wwm	0.869 5	0.853 8	0.988 7	0.916 3

5 结语

本文标注了一个中文电子病历实体映射数据

集，结合相似度算法与深度学习预训练模型，探究进行海量实体映射的最佳算法与模型组合。采用相似度算法进行候选实体生成，采用预训练模型进行实体消歧，并比较不同算法模型效果。结果显示采用 Jaccard 相似度算法与 BERT 模型的组合能够达到最优效果。同时本文提出通过别名间相似性改进实体映射效果的方法，各组合模型较改进前的 *F*1 值平均提高 15.69%，达到较理想的实体映射效果，为未来相关研究提供思路。但是本文标注样本量较少，缺乏对不同类型实体映射效果的分别比较，未详细探究术语构成特点，未来可增加对不同类型实体的对照研究，针对性探究改进方案，以提升中文电子病历实体映射效果。

参考文献

1 徐国海. 面向中文医疗文本的命名实体识别研究［D］. 上海：华东师范大学，2019.

2 吴思竹，钱庆. 医学概念标准化工作研究［J］. 医学信息学杂志，2012，33（3）：2－9.

3 WANG Q，ZHOU Y，RUAN T，et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition［J］. Journal of biomedical informatics，2019，92：103133.

4 DOGAN R I，LEAMAN R，LU Z. NCBI disease corpus：a resource for disease name recognition and concept normalization［J］. Journal of biomedical informatics，2014，47：1－10.

5 黄源航，焦晓康，汤步洲，等. CHIP 2019 评测任务 1 概述：临床术语标准化任务［J］. 中文信息学报，2021，35（3）：94－99.

6 胡佳慧，方安，赵琬清，等. 面向知识发现的中文电子病历标注方法研究［J］. 数据分析与知识发现，2019，3（7）：123－132.

7 马诗语，黄润才. 基于 ALBERT 与 BILSTM 的糖尿病命名实体识别［J］. 中国医学物理学杂志，2021，38（11）：1438－1443.

8 陈仕鸿，刘晓庆. 基于余弦距离的中文问答系统中问句相似度计算［J］. 福建电脑，2017，33（2）：31－32.

9 ZHANG Z. An improved BM25 algorithm for clinical decision support in precision medicine based on co－word analysis and cuckoo search［J］. BMC medical informatics and decision making，2021，21（1）：81.

10 于鹏. 逻辑公式间的 Jaccard 距离及其应用［J］. 计算机科学与探索，2020，14（11）：1975－1980.

11 邵清，叶琨. 基于编辑距离和相似度改进的汉字字符串匹配［J］. 电子科技，2016，29（9）：7－11.

12 王立印，张辉，陈勇. 一种基于 Dice－Euclidean 相似度计算的协同过滤算法［J］. 计算机应用研究，2015，32（10）：2891－2895.

13 AlAMMARY A S. BERTmodels for arabic text classification：asystematic review［J］. Applied sciences，2022，12（11）：5720.

14 GAO L，ZHANG L，ZHANG L，et al. RSVN：a RoBERTa sentence vector normalization scheme for short texts to extract semantic information［J］. Applied sciences，2022，12（21）：11278.

15 CHOI B，LEE Y，KYUNG Y，et al. AlBERT with knowledge graph encoder utilizing semantic similarity for commonsense question answering［J］. Intelligent automation & soft computing，2023，36（1）：71－82.

16 CUI Y，CHE W，LIU T，et al. Pre－training with whole word masking for Chinese BERT［J］. IEEE/ACM transactions on audio，speech，and language processing，2021，29：3504－3514.

17 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.