融合半监督学习与同义计算的传染病名称自动映射研究

宋培彦[1] 杨青香[12*] 胡博深[1] 杜博雅[1]
[1] 天津师范大学管理学院，天津300387
[2] 华中师范大学人工智能教育学部，武汉430079
spyer2008@126.com

摘要

医学古籍蕴含着丰富的专业知识，然而由于古代疾病名称、术语与现代标准表述不一致等问题，严重影响了公共卫生知识组织和服务质量，现有研究主要采用专家手工映射、词义计算等方式解决，存在着工作效率和准确率偏低等问题，以古籍术语辞典作为语料进行挖掘、建立传统医学术语与现代医学术语的同义关系，并映射到国际规范，形成“古-今-外”三语互通的知识库是可行方法。为此，本文以知识组织和知识发现理论为基础，设计了古今疾病名称跨语言自动映射方法，并以传染性疾病名称为例进行验证。具体过程是：首先，利用snowball算法抽取古今疾病名称同义模式，获取了12个与传染性疾病相关的疾病名称关系模式和134个同义词对。其次，依据桑基图从关联性、成熟度和延展性3个角度分析疾病名称历时演变进行可视化关联分析。同时，结合sapbert词向量和余弦相似度将传统医学疾病名称向ICD-11国际标准映射，经过人工验证映射结果达到0.23的hit@1、0.42的hit@5以及0.61的hit@10。本文发现，通过专业辞典语料，可以抽取疾病名称的语言变异情况，提高同义术语的发现效率，为构建专业知识库提供更多的入口词和语义关联性，缓解信息孤岛问题。研究还表明，以辞典中的现代医学术语名称作为映射起点，关联到ICD-11国际规范，为开展跨语言领域知识工程建设提供参考，对实现专业知识“古为今用“和国际传播也具有重要现实意义。

关键词： 关系抽取；模式匹配；术语映射；机器学习算法；术语计算

A study on automatic mapping of infectious disease names by integrating semi-supervised learning and tautology computation

Song Peiyan[1] Yang Qingxiang[12*] Hu Boshen[1] Du Boya[1]
[1] School of Management,
Tianjin Normal University,Tianjin 300387
[2] Artificial Intelligence Education Department,
Central China Normal University,Wuhan430079S
spyer2008@126.com

Abstract

Ancient medical texts contain a wealth of professional knowledge, yet the organization of public health knowledge and the quality of services are seriously affected by problems such as inconsistencies between ancient disease names and terminology and modern standard expressions. Existing studies have mainly used manual mapping and

*通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
本文为2020年度国家社科基金重大项目“国家重大突发事件信息公开质量研究”（20ZD141）成果

lexical meaning computation by experts to solve the problems of low efficiency and accuracy, etc. It is feasible to use ancient terminology thesaurus as the corpus for mining, establishing synonym relations between traditional and modern medical terms, and mapping to international norms to form a trilingual knowledge base of "ancient-present-foreign". In this paper, we propose a feasible way to mine the corpus of ancient terminology, establish the synonym relationship between traditional medical terms and modern medical terms, and map them to the international norms to form a knowledge base of "ancient-modern-foreign" trilingual interoperability. In this paper, based on the theory of knowledge organisation and knowledge discovery, we design a cross-lingual automatic mapping method for ancient and modern disease names, and validate it with infectious disease names as an example. The specific process is as follows: firstly, the snowball algorithm is used to extract the synonym patterns of ancient and modern disease names, and 12 disease name relationship patterns and 134 synonym pairs related to infectious diseases are obtained. Secondly, the evolution of disease names over time was analysed from three perspectives: relevance, maturity and extensibility for visual association analysis based on Sankey diagram. Meanwhile, the traditional medical disease names were mapped to the ICD-11 international standard by combining the sapbert word vector and cosine similarity, and the mapping results reached 0.23 hit@1, 0.42 hit@5, and 0.61 hit@10 after manual validation.In this paper, it is found that the linguistic variability of disease names can be extracted from professional thesaurus corpora, and the discovery of synonyms can be improved. efficiency, provide more entry words and semantic correlations for building a professional knowledge base, and alleviate the problem of information silos. The study also shows that taking the modern medical terms in the thesaurus as the starting point for mapping and linking them to the ICD-11 international standard can provide a reference for the construction of cross-linguistic knowledge engineering, which is also of great practical significance for the realisation of professional knowledge and the international dissemination of professional knowledge.

Keywords: relational extraction , pattern matching , term mapping , machine learning , algorithm Terminology calculations

1 引言

面对重大公共卫生事件特别是传染性疾病，不同国家和地区使用了知识组织工具和术语来描述，如ICD-11、UMLS、SNOMEDCT等，为实现有效的知识服务和决策支持提供了必要的保障。为了提高数据质量、实现知识互通，2019年5月世界卫生组织审议通过了《国际疾病分类第十一次修订本（ICD-11）》，ICD-11的传统医学章节中所使用的英语术语难以准确匹配到相应的传统医学术语，如太阴"Greater yin"国际上常译为"Tai yin"，很难与中医术语的含义匹配。同时，由于医学知识体系、历史文化的差异性，传统医学（以下简称"中医"）与现代医学（以下简称"西医"）在相同概念的匹配上存在知识特色宝藏，例如诺贝尔奖获得者屠呦呦发明的青蒿素受到东晋葛洪《肘后备急方》启发，证明了古代医学典籍中蕴含着丰富的专业知识可以"古为今用"。因此，通过术语映射手段建立中文术语与国际标准之间的对应关系，有助于将古代公共卫生知识进一步标准化、国际化，促进信息资源共享和知识服务。前人主要围绕中医学与西医学、口语化与国际疾病分类标准、中文与UMLS之间的术语映射展开研究。此外，在PubMed数据库中，生物医学信息检索者常用的一个主要检索类型是疾病名称（杨海锋，2015），疾病名称术语映射本质上体现了不同知识体系之间的融合与互补性。通过映射，可以将来自各体系的相关概念和术语整合在一起，形成一个更加全面和统一的疾病知识框架，从而提高知识的完备性和可共享能力。基于此，本文借助包含中西医疾病名称语句的语言规则构建同义词模式，进而挖掘同义词群。同时，将西医疾病名称作为隐性知识，采用语义计算方法实现中国传统医学与国际标准术语的关联映射，最终提高公共卫生领域术语的高质量建设和知识服务。

(c) 2024 中国中文信息学会计算语言学专业委员会

2 相关研究与理论基础

2.1 术语映射

2.1.1 中西医疾病名称术语映射

现有研究主要是利用疾病名称的字面相似度、词向量训练、模式匹配等3种方法实现疾病名称映射。郭思成（2019）等以《中国中医药学主题词表》和《中文医学主题词表》为例，利用深度学习工具Word2Vec提出了一种在语义层面上进行自动化匹配计算的方法。姚涛等（2017）提出了一种基于映射字典学习的跨模态哈希检索算法，利用映射字典学习降低了算法复杂度。丁泽源等（2021）使用结合注意力机制的双向长短期记忆网络抽取中文生物医学实体间的关系。对于中西医病名之间字面相似度偏低的问题，如“核病”与“注气”、“霍乱”与“紧病”，传统的基于字面相似度的匹配算法难以起到较好的映射效果。尽管利用Word2Vec等机器学习技术训练得到的词向量可以通过计算余弦相似度来获取一定的匹配效果，但这种方法需要较长的训练时间，且对于低频词汇的效果欠佳。模式匹配法是从句法角度分析语料，依赖于自定义的抽取模式，从语料中进行语义关系抽取，属于传统的语义抽取方法之一。有学者提出了基于模式识别的中医药术语同义词提取方法，对中医药行业的基础学科名词术语进行规范化考订（Liao X，2020）。

模式匹配虽能依据事先定义的规则进行快速准确的信息提取，但其局限性在于人工定义规则无法覆盖所有可能的匹配模式。半监督学习是机器学习的基本方法之一，且在标记数据稀缺的情况下，利用未标记的数据可以显著提高模型的性能。王加楠（2017）提出了基于模式的远监督关系抽取算法,使用了基于核的机器学习算法来克服关键信息不突出、数据集线性不可分等问题。杜小坤(2015)提出了一种基于信息元的模式匹配方法,解决了结构化信息不够准确、缺少有效的描述形式、处理耗时等缺点。王丰（2019）提出了一种迭代优化的模式匹配方法IOSMA, 利用已经匹配成功的元素对优化模式匹配算法，使模式匹配算法随着迭代取得了更好的匹配效果。Snowball（Agichtein, Eugene and Luis Gravano，2020）是一种半监督的机器学习方法，提供了一种从文本文档生成模式和提取元组的方法。并通过计算抽取模式的置信度，评估Snowball算法所生成的模式和元组的质量，只有那些被认为“足够可靠”的元组和模式才会被算法保留，用于系统的后续迭代，有利于规避错误规则造成的语义漂移问题，因此，本文将模式匹配法与词向量相互结合，形成覆盖率和准确率较高的新方法。

2.1.2 中英文术语映射

美国国立医学图书馆构建了生物医学领域的UMLS，对具有来源、格式、内容异构特点的医学词表统一术语格式，以解决相同概念在不同系统中名称不统一的问题，被广泛应用于医学健康领域的概念及概念关系识别、语义标注、语义消歧、本体构建等方面。刘娇等(2018)避开了外部词典和知识库,利用余弦相似度对中英韩三个语种的对齐语料库,建立了不同语种词汇间的对应关系。Chen L（2023）等通过系统地探索基于字符串、基于语义，以及字符串-语义组合的映射策略，来研究将中文医学实体映射到UMLS的技术边界。Dong Xin等（2022）提出了一种基于子网络的症状术语映射方法,有效地表示了临床症状术语，尤其是未记录术语的嵌入特征。Kim HK等（2020）结合字符串匹配方法和BioBERT 生成的术语嵌入向量，利用结构和上下文信息来计算源词和目标词之间的相似度量。Ruan T等人（2017）通过呼叫百度翻译，将中文医疗实体翻译成英文，计算平移和UMLS概念之间的Jaccard距离以进行映射。由于医学实体的多样性，翻译质量并不稳定，仅通过比较Jaccard的距离，很难其准确映射到UMLS上。随着深度学习的发展，预训练语言模型（PLMs），如BERT（Devlin J，2019）、GPT（2019）等，已被广泛应用于自然语言处理（NLP）的各个领域。预训练语言模型不仅可以学习医学术语的字符特征，同时可以获得与医学术语上下文相关的语义信息。刘方宇提出了一种对比度量学习方法来学习UMLS概念的自对齐，并训练了一种名为SapBERT的跨语言语言模型，为中英文疾病术语映射提供新的方案。由此可知，统一医学语言系统是一个相对完整的医学术语系统，对于医疗实体的规范化至关重要，但其主要由英语医学术语组成，对于英语以外的语言映射效果有待检验。

2.2 知识组织与知识发现

国内外对医学信息组织的相关研究主要围绕知识表示、知识聚合以及跨词表映射等。孙海霞（2013）等针对当前医学量表资源组织与服务粒度相对粗放问题，提出一种细粒度医学

(c) 2024 中国中文信息学会计算语言学专业委员会

量表文档知识表示框架，于诗睿（2024）等通过集成现有知识体系中的术语与新文献中的主题，使用上下文嵌入聚类主题模型对生物医学领域的专业术语进行层次化组织和映射。Saitwal H（2012）等将电子健康记录中的专有药物编码映射到SNOMEDCT和UMLS元词库中，建立一个多层次的分类系统，用于查询一个i2b2临床数据仓库。知识发现是从大量数据中获得有效、新颖、有潜在应用价值的和最终可理解的模式的高级处理过程（全国科学技术名词审定委员会，2019）。马捷（2024）等基于统计分析与关联规则的方法，挖掘中医用药规律与临床症状以及症—药规律。毛进（2024）等提出一种疾病知识网络表示学习模型及其链路预测算法,以进行疾病知识的关联关系挖掘与预测。Eissa等（2016）设计了粗糙集理论、人工神经网络、遗传算法和粗糙计量理论相结合的混合知识发现模型，通过在冠心病和丙型肝炎病毒数据集上的应用，展示了混合知识发现模型在提炼关键医疗指标和降低误诊率方面的潜力。

本文认为，对于中西医疾病名称映射，或是中英文疾病名称映射，均以数据驱动为核心，从各种辞典、文献资料中以机器学习的方式抽取同义词的过程。且在文本语料中，同义词与其连接词共同出现，可以将其归纳为根据关系规则获得疾病名称同义词对的过程。其目的在于向使用者屏蔽原始数据的繁琐细节，从原始数据中提炼出有效的、新颖的、潜在有用的同义词组。因此，本文以知识组织和知识发现为理论依据，利用半监督学习算法将分散分布的不同医学体系下的疾病名称建立联系，构建疾病名称同义词群，并利用预训练语言模型映射到国际疾病分类标准，实现古代专业知识的规范化和国际化，提高知识服务能力，在理论上是可行的。

3 模型构建

为了实现术语映射，本文设计了疾病名称映射框架图。主要包括3个方面，一是筛选目标数据，以中西医疾病术语映射、中英文术语映射为具体目标，采集相关度和专业程度较高的数据源，减少数据噪声；二是确定术语关联关系，挖掘关联关系和寻找内在规律的过程，针对医学疾病名称术语的字面语义关系较低，选择Snowball半监督学习算法挖掘关联关系，从而规避语义漂移问题；三是通过算法抽取的疾病名称同义词对构成同义词群，降低人工抽取的主观偏差。基于此，本文设计了如图1所示的疾病名称映射的主要框架。（1）采集多源数据：由于本

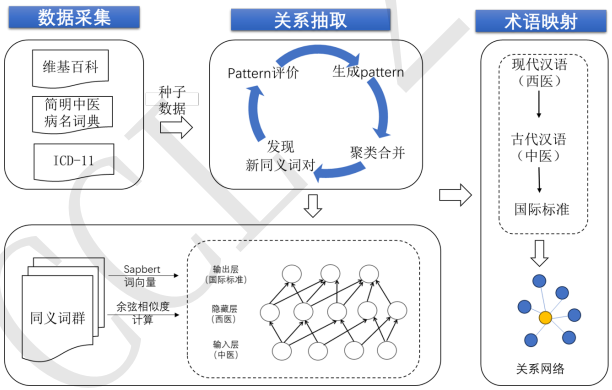


图 1. 疾病名称映射框架图

文涉及古代疾病名称、现代疾病名称、国际标准名称三者之间的相互映射，因此需要根据不同数据存储特点，完成数据采集任务。而《简明中医古病名辞典》参阅了历代有关中医古籍，辑录临床各科的古病名，并进行了中西医病名对照，可以从中抽取术语表达模式；ICD-11是世界卫生组织发布的最新版疾病分类系统，具有较高的国际地位和影响力，是本文国际标准术语映射目标；医学百科存在三种知识关联结构，即网状关联结构、层级关联结构和混合关联结构，包含疾病名称同义词、上位词、缩微词等信息，可以作为映射框架，将术语名称采用同义、近义、上下位等方式进行多种类型的映射。（2）中西医疾病名称映射：Snowball算法是一种从文本数据中抽取结构化信息的技术，其本质是一种半监督学习方法，即利用少量标注数据在未知语料中预测相关关系。《简明中医古病名辞典》内疾病名称之间的同义模式较为稳定，如句子“大瘕泄《难经·五十七难》大瘕泄者，里急后重，数至圊而不能便，茎中痛。即指由湿毒阻滞肠道所致腹痛、腹泻、里急后重、便下脓血者，相当于现代医学之细菌性痢疾”，且关系模式的重复率较高，基于此，本文利用Snowball算法完成疾病名称同义关系抽取，利用Snowball半

（c）2024 中国中文信息学会计算语言学专业委员会

监督学习算法，抽取中西医疾病名称同义关系模式，构建中西医同义词群。（3）中医向国际标准映射：西医疾病名称在国际上标准化程度较为完善，本文借助多层感知机的思想，将每一个中医疾病名称作为网络的输入层，每一个西医疾病名称作为网络的隐藏层，而国际标准疾病名称术语作为输出层，输入层不涉及任何计算，只需融合Sapbert生物医学预训练语言模型生成的Bert词向量，利用余弦相似度实现隐藏层和输出层的同义计算，即可完成中医向国际标准的映射。

3.1 疾病名称术语映射

3.1.1 基于Snowball算法抽取疾病名称同义词对

利用算法Snowball算法抽取疾病名称同义词对主要流程如下：首先，将人工定义的一般性种子数据作为算法的输入部分，以此减少人工参与的程度，实现从文本中自动抽取结构化数据的目标。其次，利用该算法自定义的五元组（L，标签1，M，标签2，R），其中标签1和标签2代表疾病名称实体，L、M、R代表与疾病名称实体相关联的左、中、右的句子上下文。Snowball算法通过将句子上下文信息编码为带权重的向量来学习文本中的疾病名称同义关系，这些向量基于实体周围一定范围内的词语及其出现频率，以此衡量每个词语在特定上下文中的重要性，并据此抽取同义词对。例如，模式（[有时，0.2]，感冒，[又称，0.5]，伤寒），可以匹配“有时感冒又称伤寒”，其中“有时”和“又称”分别带有重要性权重0.2和0.5，从而帮助识别“感冒”和“伤寒”作为同义的疾病名称。该方法使得算法能够在仅有少量标注数据的情况下，从大量未标注的语料中有效地学习和扩展疾病名称的同义关系。然后，利用公式1得到不同五元组向量之间的相似度以及最小相似度的阈值。借助数学领域的内积概念，将五元组内“左”、“中”、“右”三个向量与其他元组向量进行乘法运算，以此衡量不同元组的相似度。其中匹配是指抽取的同义关系模式与文本中的具体实例相符合。例如，若模型抽取的同义关系模式为（[有时，0.2]，感冒，[又称，0.5]，伤寒），并且有一句文本是“鼠疫，又称，黑死病”，则认为该模式匹配这句话的内容。不匹配则指在关系模式匹配过程中，某个模式与文本中的具体实例不相符。例如，在某个关系模式下，如（[有时，0.2]，感冒，[又称，0.5]，伤寒），识别出感冒又称“鼠疫”，而在先验知识下，感冒又称“伤寒”而不是“鼠疫”，则（伤寒，鼠疫）同义词对被认为不匹配。

$$\operatorname{match}\left(t_p, t_s\right)=\begin{cases} l_p l_s+m_p m_s+r_p r_s, (\text{ 匹配}) \\ 0, \quad (\text{ 不匹配}) \end{cases} \tag{1}$$

由此可知，Snowball算法利用文本数据中的重复性和固有结构，通过迭代式的模式扩展过程，能够高效地从大量的文本数据中抽取出目标关系。此外，Snowball算法的优点在于能够自动调整抽取模式置信度，使挖掘的关联规则更加准确，即在每一次挖掘模式和同义词对的迭代过程中，仅让大于阈值的关系模式和实体对进入下一轮的迭代，从而保证最后生成的同义词对的质量。

3.1.2 Bert词向量计算

Snowball算法在关系模型和同义词对抽取过程中，其核心是简单的规则匹配，而对于语义层面的同义词抽取效果欠佳，造成覆盖率不够、类推性不足等语义漂移问题。中医疾病名称向ICD-11映射的关键在于语法结构和表达习惯的差异性。虽然BERT模型在一定程度上能够处理未知词汇，但对于某些古代疾病名称的相关词汇难以生成准确的词向量。本文利用Snowball算法抽取的古今疾病名称同义词对，以现代医学疾病名称为中介桥梁，利用Bert词向量表征语义关系，从而从语义角度提高疾病名称映射准确性，克服单纯依靠规则匹配造成的语义漂移问题。

SapBERT是一种用于生物医学实体链接任务的开源预训练模型。与传统的BERT模型相比，SapBERT在预训练阶段通过自我对齐目标来学习生物医学实体的表示，将模型的同一实体在不同语言或领域中的不同描述或表示对齐到一个共同的表示空间，有利于处理跨语言或跨领域的任务，从而更好地理解实体之间的关系。此外，利用余弦相似度公式量化向量间的距离，如公式2所示，该值的范围为[-1,1]，-1为完全不相似，1为完全相似。因此，本文基于嵌入表示

来衡量中文疾病名称与国际标准之间的语义相似性，以此促进医学术语的规范化和互操作性。

$$similarity(A,B)=\frac{AB}{||A||\times||B||}=\frac{\sum_{n}^{i=1}A_i\times B_i}{\sqrt{\sum_{n}^{i=1}A_i^2}\times\sqrt{\sum_{n}^{i=1}B_i^2}} \tag{2}$$

3.2 术语映射评价指标

3.2.1 置信度

通过评估同义关系模式以及元组的置信度筛选同义词对，利用公式2计算在模式P下识别正确同义词对与错误同义词对的个数统计。若处于“xx又称xx”模式下，可以在4条语句中识别正确，1条语句中识别错误，则该模式的置信度为Conf(P)=4/5=0.8。在对关系模式进行置信度评价后，会得到若干个候选同义词对。对于每个同义词对，存在若干模式集合以及该元组句语中上下文和匹配函数的相似性度量，如图2所示。

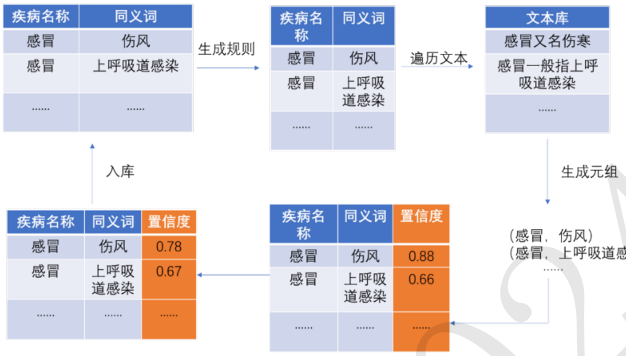


图 2. 疾病名称同义词关系抽取流程

因此，利用公式3和公式4按对应模式和上下文的匹配程度来表示计算元组置信度，且模式和上下文之间的匹配程度越低，产生无效同义词对的机会就越高。若同义词对周围的上下文“足够相似”，即同义词对相似度达到阈值，则根据语料库中出现的已知元组进行聚类，进而形成关系抽取模式。最后，将形成的关系抽取模式运用到语料库里面提取更多的同义词对，不断重复上述过程。同时将置信度较高的预测结果加入到种子数据中，作为新的训练数据，重新训练模型。

$$Conf\,(P)=\frac{P(right)}{P(right)+P(error)} \tag{3}$$

$$Con(T)=1-\prod_{i-0}^{|p|}(1-(conf(pi)\times Match(ci,pi))) \tag{4}$$

3.2.2 hit@k指标测度

由于国际疾病分类（ICD-11）标准是根据病因、病状等因素进行综合分类，而语言风格、文化背景以及医学知识的差异性，中英文疾病名称一对一、一对多或者零对应的映射关系，因此常见的测评指标如精确率、召回率、F1值等很难客观评价映射准确性，更适合采用相关性评估进行评测。hit@K，也称为Top-K准确率，是一个衡量推荐系统性能的指标，它表示在所有搜索查询中，系统在前K个推荐结果中至少正确推荐了一个相关项目的比例，该指标常用于评估推荐系统的准确性。如公式5是hit@K指标的计算方法，其中分子是所有满足条件的查询数量（即前K个推荐中至少有一个是正确推荐的），分母则是所有搜索查询的总数，K值可选取1、3、10等。

$$hit@K=\frac{\sum_{q\in Q}isCorrect(q,K)}{|Q|}\% \tag{5}$$

因此本文通过近似匹配的原则，利用余弦相似度的算法得到中文疾病名称与多个英文疾病名称的若干个余弦相似度，并引入hit@k指标计算其有效性，即在Sapbert模型给出的前k个余弦相似值中，正确映射占所有映射结果的百分比。如果Sapbert预训练语言模型计算余弦相似度的正确映射排名越靠前，说明该方法完成中英文映射效果较好。

CCL 2024

4 实验

4.1 数据采集与预处理

本文选取传染性疾病作为实验语料，首先，从维基百科获取传染性疾病词条信息。然后，将以图片形式存储的《简明中医古病名辞典》利用OCR识别工具转换为文字，在此基础上采集包含“传染”的相关疾病描述，共计125个疾病名称。如图3所示，将上述采集的语料按照HTML标签，以人工方式进行命名实体识别（使用NAM表示疾病名称、BOOK表示疾病名称出处），经过去重整理后得到条165条文本数据。最后，在世界卫生组织官方网站下载《国际疾病分类》第11次修订本（简称ICD-11）作为为跨语言映射的基准，剔除非疾病术语，如药品名称、体征和症状等，得到2478个国际标准术语。

<NAM>小腿慢性湿疹</NAM>。治宜疏风解毒祛湿，方用草薢渗湿汤加味。
<NAM>下痞</NAM> 《华佗神医秘传》卷16: “华佗治下府神方。” 即<NAM>秽疮</NAM>。详见该条。下症疮《景岳全书》卷64: “治痔漏下痈疮。” 即秽疮。详见该条。
<NAM>下格</NAM> 《医钞类编》卷15: “……此为下格。” 即指肠内因秽浊，恶物结滞，阻碍大便传导所致大便数旬不通，时时作呕，饮食不尽，食入即吐等病证。相当于现代医学的便秘、肠梗阻等症。治宜荡涤肠垢，方用小承气汤。
<NAM>下注疮</NAM> 《世医得效方》卷19: “名曰下疳疮。” 即<NAM>烂腿疮</NAM>。详见该条。
<NAM>下部病</NAM> 《心印绀珠经》下卷: “瘙疝之病俗云下部病。” 即<NAM>狐疝</NAM>。详见该条。下部愚 《别录》: “苦参，……(治)恶疮、下部慝。” 即阴疮。详见该条。
<NAM>下消</NAM> 《丹溪心法》: “……下消。” 又称肾消、消肾。指因肾水亏竭，蒸化失常

图 3. 部分语料示例

4.2 传染性疾病中西医疾病名称映射

依托于GitHub内开源Snowball半监督学习模型，将源代码下载到本地，利用PyCharm编译器完成如下实验步骤:

（1）初始化snowball算法，根据“二八定律”经验值，在许多情况下大约80%的结果通常来自于20%的原因。因此，本文随机抽取22个可以反映出待挖掘关系的样本种子，如同义关系（丁疮；疔疮）、上下位关系（食积；九积）作为snowball算法的种子数据，如表1所示。设置迭代轮次10次，规则置信度为0.6，公式1同义词对匹配函数阈值为0.6。

疾病名称	别称	疾病名称	别称
丁疮	疔疮	儿晕	儿痉
七情饥饱款	七情伤感嗽	九子疱	颈淋巴结结核
八脚虫疮	阴虱疮	九道出血	大衄
儿风	妊娠风痉	刀癣	泛发性神经性皮炎

表 1. snowball算法部分种子集

（2）利用自举（bootstrapp）方法初始化迭代。在每次迭代中，根据步骤（1）设置的初始化阈值，Snowball算法尝试在文本语料中寻找与种子数据相似的模式，然后将这些模式和它们对应的关系实例加入到种子集合中。随着迭代的进行，种子集合会逐渐增大，算法能够识别的关系模式也会越来越丰富。

（3）在自举迭代过程结束后，Snowball算法会将所有提取的关系写入到输出文件中，该文件包含了每个关系实例的详细信息，如实体对、它们之间的关系、以及关系的置信度等。本文在迭代轮次为10次，规则置信度及同义词对阈值为0.6的条件下，得到21条规则，如表2所示。

（4）以现代疾病名称为入口，整理与其具有较高相关性的古代疾病名称形成同义词群，如狂犬病为现代疾病名称，与之相对应的古代疾病名称有颠犬伤、狂犬伤、狂犬啮人、疯犬咬伤、犬伤、狗啮疮、风犬伤毒、风狗咬、狗啮。经噪声整理后，去除错误规则，最终提取出12个古今同义词关系模式，如表4所示，并将算法抽取的规则划分为等同关系、等级关系以及相关关系。其中等同关系频次最高，相关关系频次最低。共计64条古代疾病名称可以映射到现

序号	同义词对	置信度	规则
1	（八脚虫疮，阴虱疮）	0.81	(……即……)
2	(干血痨，干血)	0.79	(……称……)
3	(白缠喉，白喉)	0.67	(……相当于现代医学的……)
4	（禽流感病毒，甲型流感病毒）	0.63	(……属……)
5	（百日咳，鸡咳）	0.66	(……俗称……)

表 2. Snowball算法生成的部分规则

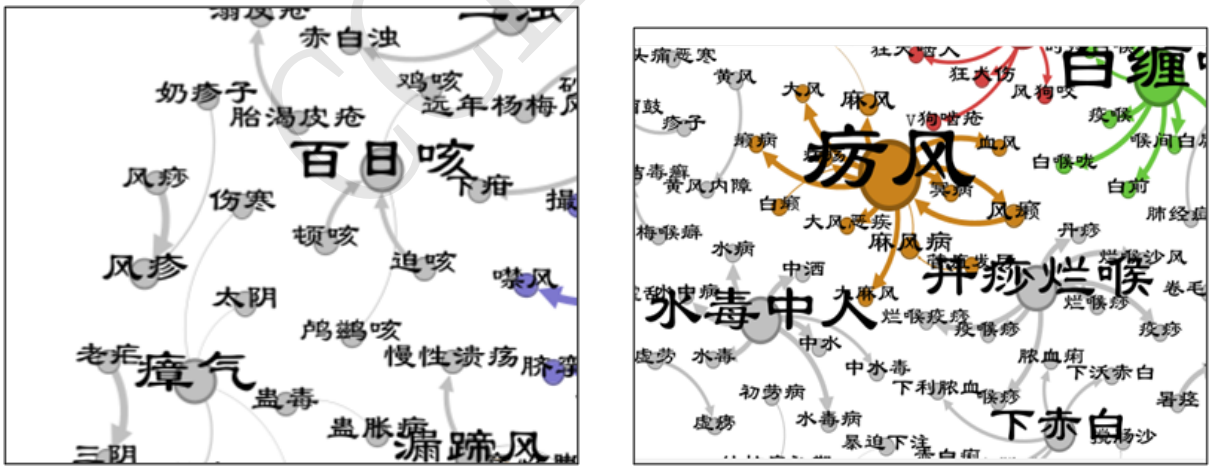
代疾病中，而34条古代名称未能与现代疾病名称相对应，且存在多个古代疾病名称表示一个现代名称的情况，如现代结核病大致相当于于丁吴痞、尸注、初劳病、抱儿痨、注气、虚劳。而伏气、后不利、血风等未找到与现代疾病名称的映射结果。

关系	规则	频次
等同关系	又称、又名、亦指、是一种	64
	即、简称、名为、指、俗称、俗名	43
等级关系	属于、之一、属	15
相关关系	类似、如、参见	9
	病因、故、引起、多由、故又称	6

表 3. 疾病名称按不同关系统计频次

4.2.1 构建中西医疾病名称同义词群

部分疾病名称未能识别或识别错误，情况主要分为以下2类：一是疾病名称以符号连接，表示一般性停顿，以致未能识别其他同义词对。如语句“赤白痢，又称脓血痢，小肠泄，下利脓血，下沃赤白”中，“赤白痢”和“小肠泄”以及“赤白痢”和“下利脓血”之间没有连接词，导致未识别，说明对多个统一术语的抽取能力还有一定局限性。二是语义误差，即存在同义关系模式，但其语义表达不符合抽取要求。如语句“由于风疹的疹子来得快，去得也快，如一阵风似的，“风疹”也因此得名。”描述了疾病命名来源，但错误识别为“疹子…如…一阵风”。



(a) 疾病名称离散分布　　(b) 疾病名称集中分布

图 4. Gephi古代疾病名称可视化效果

为了进一步揭示疾病名称的关联性，本文利用Gephi进行可视化展示中西医疾病名称同义词群，其中节点代表疾病名称，边代表疾病名称的关系之间的关系，并按照其出现的频次作为该关系的权重，以表示关系的重要性。平均加权度是指整个网络中节点的度数以及节点之

间边的权重，反映了各个节点之间直接关联的程度。在该网络中，平均加权度为33，表明疾病名称之间存在强而紧密的联系。利用Gephi内置的Modularity聚类算法展示疾病名称的同义词群，并按颜色划分不同词群，由图4可知疠风、白缠喉、时毒等聚类效果较为明显，表明利用Snowball算法抽取疾病名称同义词对效果显著。此外，由网络分布的离散状态，推测疾病名称相关性较高，且疾病名称分布越集中，表明其涵盖的知识越粗糙，反之细粒度较高。如图4所示。

4.2.2 中西医疾病名称历时演变

由于《简明中医古病名辞典》内涵盖疾病名称的来源信息，如“丁奚《诸病源候论》卷47，小儿丁奚病者”。基于此，本文以“霍乱”、“结核病”、“猩红热”、“疟疾”和“风疹”5个中医疾病名称为例，根据《简明中医古病名辞典》内这5个疾病名称的出处典籍以及4.2.2节构建的中西医疾病名称同义词群，以疾病名称所在典籍成书时间作为节点，反映疾病名称在不同时代的演变过程和知识脉络。并利用镝数图表在线平台，构建的疾病名称历时变化桑基图。其中第一列为西医疾病名称，第二列为中医疾病名称，第三列为中医疾病名称选自的医学典籍，第四列为医学典籍的成书年代。如图5所示。

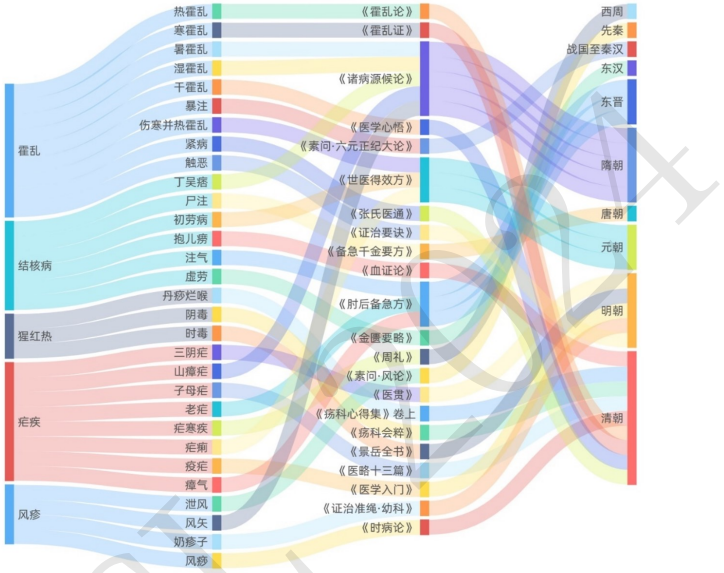


图 5. 疾病名称历时演变

图5从关联性、成熟度和延展性3个角度分析中西医疾病名称历时演变，对知识的关联性和时序进行更直观的揭示和溯源。一是从中西医疾病名称的关联性可知，两种医学体系命名相互融合。如西医称为“霍乱”，从战国至秦汉中医体系下将其依次命名为“暴注、湿霍乱、暑霍乱、触恶、热霍乱、干霍乱”，用户通过模糊检索“霍乱”即可得到大部分与之相关联的疾病名称。二是从医学典籍成书时间终止点明晰疾病发展的成熟度。以西医“结核病”为例，由图4可知自东汉至清代均有对该种疾病的医学记载，与其关联的中医疾病名称先后出现了“注气—丁吴痞—初劳病”的名称变化。本文结合辞典注释进行分析，由于古代对该种疾病症候较为模糊，使用注气这一泛化的名称表达对气滞症状的初步理解。随着对病理机制的不断加深，以丁吴痞命名，不断细化此类疾病，其中丁表示疼痛，吴表示胀满，痞表示阻塞。由此可知，在明朝时期结核病已形成较为完整的治疗体系。三是以节点密集程度衡量疾病名称的延展性。从典籍节点密集程度来看，《诸病源候论》中涉及霍乱、疟疾、风疹三种疾病名称表明该部典籍蕴含综合性知识。因此，典籍节点越密集知识覆盖面越广，典籍节点越稀疏表明其对某种疾病的专业度越强。从时间节点密集程度来看，隋朝至元代时期古代医学较为发达或处在疾病爆发时期。且从西周至清朝时期，节点大小呈递增趋势，表明医学知识逐渐丰富，医学典籍数量逐渐增多。

通过对疾病名称的溯源，一方面便于医学典籍的知识表示和智能应用，为用户提供更多的入口词，另一方面可以反映古代疾病名称的时间演化，进而归纳疾病的历史演变历程，技术的成熟度以及疾病的延伸情况，为用户推送更具时序逻辑关联性的知识内容与服务。

(c) 2024 中国中文信息学会计算语言学专业委员会

4.3 中医传染性疾病向国际标准映射

由4.1和4.2节，本文构建了中西医传染病疾病名称同义词群。基于此，将西医疾病名称和ICD-11国际标准术语转化为词向量的形式，采用余弦相似度计算的方法实现中国传统医学与国际标准术语的关联映射，主要流程如下：首先，加载SapBert预训练语言模型，使用tokenizer对ICD-11疾病分类术语集进行编码，并将其转换为模型的嵌入向量，从而得到2478个疾病术语的向量表示，将其以数组形式存储在numpy文件中。然后，由4.2.2节构建的中西医疾病名称同义词群，本文将西医疾病名称输入到Sapbert预训练语言模型中，利用python编程语言定义query的变量，输入待映射中文疾病名称，且同上所述将其转换为向量表示，作为查询向量。最后，调用利用ICD-11生成的numpy文件，利用公式2计算查询向量与所有ICD术语嵌入向量之间的余弦相似度，并按余弦相似度值降序排列。若余弦相似度排在第一位的结果即为正确映射，则hit@1为100%。而真实情况是在余弦相似度排在第5个甚至第10个预测时才会出现正确映射结果，因此本文依次评估hit@5和hit@15，表示Sapbert模型在前5个和前10个余弦相似度中出现真实值的输出占所有疾病名称的百分比。

序号	现代疾病名称	ICD-11标准	模型预测结果	余弦相似度
1	鼠疫	plague	plague	0.924
			bronchitis	0.925
			kleptomania	0.926
			trance disorder	0.926
			rabies	0.927
			dyspnea disorder	0.928
2	霍乱	Cholera	pulmonary oedema	0.932
			syngamosis	0.932
			pulmonary eosinophilia	0.933
			asthma	0.933
			pica	0.935
			adjustment disorder	0.936

表 4. 部分疾病名称中英文映射

如表4所示，展示了部分中英文疾病名称Sapbert映射的余弦相似度值。以人工核验的方式，按照中文版ICD-11国际标准进行校对。利用公式5计算中英文疾病名称映射实验结果，达到23%的hit@1，42%的hit@5以及61%的hit@10。实验结果表明，在利用Sapbert这一预训练语言模型的强大基础上，并将国际疾病分类第十一版（ICD-11）的标准化术语整合进Sapbert模型中，能够提高中英文疾病术语之间的精准映射。

5 研究总结

本文针对疾病名称古今不对应、国际化标准化不对应这两个关键问题，提出了融合半监督学习与语义计算的疾病名称跨语言映射方法，能够将模式规则和词向量结合，自动挖掘古今疾病名称对应关系，有助于形成可信、可解释的知识表示与知识发现，推动构建人机两用的专业知识库，提高知识发现效率。最终实验结果表明，hit@10映射结果达到0.61，证明该方法切实可行。其优点在于，ICD-11结构简单、规范的特点，保证了语义映射机制和映射过程的规范性，提高了多知识组织系统集成的映射效率。通过映射能够快速提取“传统医学—现代医学—国际标准”的专业知识关联性，分别从语义关联性和时间演变两个维度建立知识的关联网络，有助于提高信息质量和服务效率。本文采用专业词典和维基百科等数据进行“富知识数据”进行抽取，获取疾病的古今疾病名称对应关系，准确性高但覆盖率还有待检验。今后可以重点对大规模古籍文本语料、古代医案等数据进行知识挖掘和建立关联关系，以更好的反映知识全貌。

参考文献

王加楠,鲁强. 2017. 基于模式的远监督关系抽取算法, 中文信息学报,31(02):122–131.

(c) 2024 中国中文信息学会计算语言学专业委员会

丁泽源,杨志豪,罗凌. 2021. 基于深度学习的中文生物医学实体关系抽取系统, 中文信息学报,35(05):70-76.

刘娇,崔荣一,赵亚慧. 2018. 基于共现词映射的中英韩跨语种文档相似度计算, 中文信息学报,32(03):55-63.

赵洁,贾君枝,司莉. 2023. 基于社会化问答的用户健康信息需求概念体系构建——以糖尿病为例, 情报理论与实践,46(07):150-157.

杨海锋. 2019. 用户行为在信息检索中的研究现状及发展动态评述, 图书情报知识,(06):79-88.

郭思成,李纲,周华阳. 2019. 基于*Word2Vec*的医学知识组织系统互操作研究——以词表间语义映射为例, 情报理论与实践,42(09):160-165+176.

杨海锋. 2019. 用户行为在信息检索中的研究现状及发展动态评述, 图书情报知识,(06):79-88.

姚涛,孔祥维,付海燕,等. 2018. 基于映射字典学习的跨模态哈希检索, 自动化学报,44(08):1475-1485.

王丰,王亚沙,赵俊峰,等. 2019. 一种基于迭代的关系模型到本体模型的模式匹配方法, 软件学报,30(05):1510-1521.

Kim HK,Choi SW, Bae YS. 2020. *A Context-Aware Term Mapping with String Matching and Embedding Vectors*, *APPLIED SCIENCES-BASEL*,10(21).

Dong Xin,Zheng Yi,Shu Zixin. 2020. *TCMPR: TCM Prescription Recommendation Based on Subnetwork Term Mapping and Deep Learning.*, *APPLIED SCIENCES-BASEL*,10(21).

Liao X,Bu Y,Jia Q. 2020. *Traditional Chinese medicine as supportive care for the management of liver cancer:past,present,and future*, *Genes&Diseases*,7(3):370-379.

杜小坤,李国徽,王江晴,等. 2015. 基于信息元的模式匹配方法,软件学,26(10):2596-2613.

Agichtein, Eugene and Luis Gravano. 2000. *Traditional Snowball: extracting relations from large plain-text collections.*, *Digital library.*

Chen L, Qi Y, Wu A. 2023. *Mapping Chinese Medical Entities to the Unified Medical Language System.* , *Health Data Science.*

Ruan T,Wang M,Sun J,Wang T, Zeng L, Yin Y, Gao J. 2017. *Interoperability and mapping between knowledge organization systems: Meta thesaurus—unified medical language system of the National Library of Medicine* , *Journal of Biomedical Semantics.*

Devlin J,Chang M-W,Lee K, Toutanova K. 2019. *BERT: Pretraining of deep bidirectional transformers for Language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies*, *Journal of Biomedical Semantics.*

孙海霞,郝洁,郭臻. 2023. 基于知识元的细粒度医学量表文档知识表示框架构建, 数字图书馆论坛*l informatics*, 19(12):86-98.

于诗睿,李爱花,杨雪梅. 2024. 融合知识组织体系的层次化主题挖掘方法研究, 数据分析与知识发现, 1-18.

Saitwal H,Qing D, Jones S, et al. 2012. *Cross-terminology mapping challenges: a demonstration using medication terminological systems*, *Journal of biomedical informatics*, 45(4),613-625.

毛进,侯博文,王依蒙. 2024. 基于知识元引用网络的细分领域演化特征研究, 情报理论与实践,47(02):107-115.

全国科学技术名词审定委员会. 图书馆·情报与文献学名词, *https://www.termonline.cn.*

Mohammed M. Eissa, Mohammed Elmogy, Mohammed Hashem. 2016. *Rough-Granular Computing knowledge discovery models for medical classification*, *Egyptian Informatics Journal*,17(3):265-272.

CCL 2024