引用格式：方思怡.标准数字化转型下面向“一带一路”倡议的中医药领域国家标准术语语料库构建研究 [J]. 标准科学，2026 (2):94-100.
FANG Siyi. Research on the Construction of a National Standard Terminology Corpus for the Field of Traditional Chinese Medicine Oriented towards the Belt and Road Initiative in the Context of Standard Digital Transformation [J]. Standard Science, 2026 (2):94-100.

标准数字化转型下面向“一带一路”倡议的中医药领域国家标准术语语料库构建研究

方思怡
（上海市质量和标准化研究院）

摘　要：【目的】助推标准数字化和智能化转型，为我国“一带一路”倡议下的国际标准化发展提供技术支撑。【方法】采用大模型技术，聚焦中医药领域，基于上海标准文献馆的馆藏数据资源提出大模型赋能中医药国家标准术语语料库的构建思路并明确构建流程。【结果】完成面向“一带一路”倡议的中医药国家标准术语语料库构建并提出应用场景。【结论】以大模型为代表的人工智能技术能有效赋能“一带一路”倡议背景下的国家标准术语语料库建设。

关键词：术语语料库；中医药标准化；标准数字化；大模型；人工智能
DOI编码：10.3969/j.issn.1674-5698.2026.02.012

Research on the Construction of a National Standard Terminology Corpus for the Field of Traditional Chinese Medicine Oriented towards the Belt and Road Initiative in the Context of Standard Digital Transformation

FANG Siyi
(Shanghai Institute of Quality and Standardization)

Abstract: [Objective] The study aims to promote the digital and intelligent transformation of standards and provide technical support for the international standardization development under China's Belt and Road initiative. [Methods] By adopting large language model technology and focusing on the field of traditional Chinese medicine, based on the data resources of the Shanghai Standard Literature Library, the construction ideas and processes of the large-language-model-empowered national standard terminology corpus of traditional Chinese medicine are proposed. [Results] The construction of the national standard terminology corpus of traditional Chinese medicine for the Belt and Road initiative is completed and application scenarios are proposed. [Conclusion] Artificial intelligence technology represented by large models could effectively empower the construction of national standard terminology corpora in the context of the Belt and Road initiative.

Keywords: terminology corpus; standardization of traditional Chinese medicine; digitalization of standards; large language model; artificial intelligence

---

基金项目：本文受上海市质量和标准化研究院院立项目“基于大模型的国家标准术语语料库构建及应用研究”（项目编号：YRY202506）资助。

作者简介：方思怡，硕士，工程师，知识产权师，研究方向为标准数字化、标准数据挖掘、标准舆情分析、标准知识服务。

0 引言

“一带一路”（Belt and Road）源自建设“丝绸之路经济带”与“21世纪海上丝绸之路”的重大倡议，自2013年该倡议提出以来已日益成为横跨欧亚大陆、连接全球多个地区的国际公共物品与国际合作平台，对推动构建人类命运共同体有重大意义。作为质量基础设施的重要组成部分，标准在“一带一路”建设中持续发挥“软联通”的作用[1]，能够助推我国重点产业和优秀传统文化“走出去”。在共建“一带一路”国家的相关产业中，中医药产业蕴含大量的文化遗产[2]，在全球健康产业中愈发受到重视，亟须在“一带一路”倡议背景下提升中医药国际标准化的影响力。在标准文献的诸多要素中，标准术语通常来自特定领域的权威性规范用语，是特定领域标准制修订的重要依据。国家标准在支撑高质量发展中发挥基础性的作用，中医药领域国家标准术语是建立中医药领域贸易秩序规则体系的重要基石。当前我国的“一带一路”共建国家标准信息平台主要依托权威标准化英汉语料库和标准题录信息为国内外标准化工作者提供标准检索与快速翻译服务，尚未涉及面向共建“一带一路”国家重要语种的标准术语语料库。涵盖共建“一带一路”国家重要语种的中医药领域国家标准术语语料库能够有效提升“一带一路”共建国家的标准化发展水平。在标准数字化转型的时代下，中医药领域国家标准术语语料库也能为共建“一带一路”国家所需的标准智能检索、标准智能问答、标准智能编写等标准数智化公共服务提供必要的数据支撑。为此，本文根据前期研究中制定的大模型赋能国家标准术语语料库的构建路径[3-4]，聚焦中医药领域，以国家标准为对象，结合标准数字化和智能化发展的趋势，提出面向“一带一路”倡议的中医药领域国家标准构建方法，以期能够助推“一带一路”倡议下的国际标准化发展。

1 中医药领域国家标准术语语料库的构建方法

1.1 中医药领域国家标准术语语料库的构建思路

在标准数字化和智能化发展的背景下，以大语言模型（Large language models, LLMs）为代表的人工智能技术正在重塑原有的标准业务模式，标准语料库建设也不例外。为了满足语料库建设的普适性需求以及标准语料数据加工和应用的个性化特点，本文参考了建立语料库的一般原则与通用技术规范[5-6]，细化了标准语料库的通用性构建流程，依次包括标准语料库设计、标准语料收集、标准语料处理、标准语料标注、标准语料库生成、标准语料库管理与维护这6个阶段。作为标准语料库的子集，中医药领域国家标准术语语料库的构建需要在遵循上述流程的基础上进行合理调整。在人工智能时代，上述流程的实现方式将得到大模型的赋能，具体成效与大模型的应用深度相关，主要取决于标准数据底座和技术底座的建设水平。

在标准数据底座层面，当前上海标准文献馆的国家标准馆藏数据多为纸质文本形式，大部分纸质国家标准文献已转化为机器可读取和操作的数字化文本，在一定程度上已具备构建国家标准术语语料库的数据基础；而在标准技术底座层面，上海标准文献馆在近年来已陆续积累了标准语料挖掘与加工[7-8]、标准数字资源库建设[9]、以标准知识图谱为代表的标准知识库建设[10-11]等方面的标准数智化经验，形成了大模型赋能标准数字化应用的路径以及基于大模型的国家标准术语语料库构建方法。在前期研究基础上，本文从上海标准文献馆的标准数智化建设现状出发，结合“一带一路”的国际标准化发展需求，进一步提出大模型赋能中医药领域国家标准术语语料库的构建思路（详见图1）。

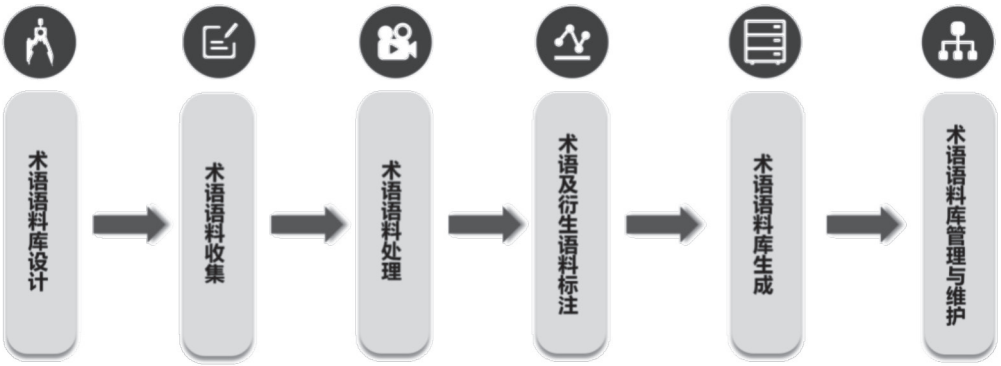

图1 大模型赋能中医药国家标准术语语料库的构建思路

与标准语料库的构建流程类似，大模型赋能中医药国家标准术语语料库构建流程同样涵盖了语料库设计、语料收集等环节，区别在于术语及衍生语料标注部分。中医药国家标准术语语料库高度重视标准首选术语中文名称及其共建“一带一路”国家重点语种对应词的获取。在当前的“一带一路”共建国家和地区中，阿拉伯地区与巴西的人口数量相对较多。其中，阿拉伯地区曾是中国古代丝绸之路的重要节点；巴西地处南美洲，是南半球最大的发展中国家[12]。在全球经济局势日益复杂的当下，中阿关系以及中拉关系下的中巴关系已逐渐成为国际经贸发展与全球治理的焦点[13-14]。本文基于“一带一路”倡议的发展现状与实际需求，将阿拉伯语和葡萄牙语作为本次中医药国家标准术语语料库衍生语料的语种，其中前者为中亚和北非部分地区的官方语言，后者为巴西的官方语言。

在图1所示的构建思路中，大模型主要在中医药领域国家标准术语及衍生语料标注中发挥语义理解和内容生成的作用。在国家标准文献中，标准术语通常位于以标准术语为主题的章节中，不同的标准术语元素在标准术语条目中根据编写规则呈现于相应的位置并具备特定的文本结构特性。例如标准术语编号通常为标准术语条目中出现的第1个标准术语元素。纵观我国的国家标准编写历程可知，国家标准术语编写已历经若干阶段，不同时期、不同领域的国家标准术语条目存在一定的编写规则差异。“前大模型”时代，国家标准术语语料标注大多采用人工或基于规则的方法实现，其中基于规则的方法需要根据不同的标准术语条目编写风格制定相应的实体抽取模板，存在灵活性低、效率不高等问题；语料翻译则通常采用传统的机器翻译工具，无法满足专业词汇的精准翻译需求[15]。与传统的实体抽取和机器翻译技术相比，通用型大模型具有更强的学习迁移能力和跨语言处理能力[16]，能够在无须制定不同模板的前提下兼容不同编写风格的标准术语条目，也更能胜任标准首选术语中文名称兼顾速度和准确度的高质量翻译要求[17]。

1.2 中医药领域国家标准术语语料库的构建流程

1.2.1 中医药领域国家标准术语语料库设计

本文参考GB/T 20001.1—2024[18]中的标准术语条目说明，在前期研究基础上，结合“一带一路”倡议背景下的中医药国家标准术语语料库建设需求，制定了该语料库的整体框架，其示意图见图2。国家标准术语语料库的数据项除了需尽可能涵盖GB/T 20001.1—2024提及的标准首选术语中文名称、标准术语定义等常规标准术语元素外，也要包括标准首选术语中文名称的阿拉伯语和葡萄牙语对应词。

1.2.2 中医药领域国家标准术语语料收集

中医药国家标准术语语料库的源数据均来自上海标准文献馆的馆藏数字化资源。考虑到术语标准通常在所面向的领域中具有较强的影响力，本文从国家标准馆藏数字化资源中选取与中医药密切相关的术语标准，形成中医药国家标准原始数据集。经统计可知，当前该数据集共有13篇现行

标准术语编号	标准首选术语中文名称（简）	标准首选术语中文名称（繁）	标准首选术语英语对应词	标准许用术语中文名称	标准许用术语英语对应词	标准拒用术语中文名称	标准拒用术语英语对应词	标准术语符号	标准术语缩写形式	标准术语领域	标准术语定义	标准术语示例	标准术语的注	标准术语源文献	所属标准	“一带一路”阿拉伯语对应词	“一带一路”葡萄牙语对应词
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……

图2 中医药领域国家标准术语语料库的整体框架示意图

状态的国家标准文献，合计约80.08万字符，其中与标准术语相关的字符数约为78.31万。

1.2.3 中医药领域国家标准术语语料处理

鉴于大模型具有输入数据量的限制，为了确保大模型的语义理解准确性，本文采用自然语言处理技术，设计了面向国家标准数字化文本的标准术语数据单元切分算法，以标准术语条目为数据单元，对中医药国家标准原始数据集进行语料处理，得到中医药国家标准切分数据集。

1.2.4 中医药领域国家标准术语及衍生语料标注

语料标注是中医药国家标准术语语料库构建的核心环节。本文针对若干国产大模型进行选型测试，综合考量经济、硬件等多个条件后分别选取星火SparkMax进行标准术语元素智能抽取，并采用DeepSeek-R1-32B获取标准首选中文术语名称的阿拉伯语、葡萄牙语对应词，共同实现标准术语及衍生语料的自动标注。SparkMax和DeepSeek-R1-32B的基本信息及功能如表1所示，其中大模型的基本信息均源自公开资料。

标准术语及衍生语料标注均建立在提示词工程的基础之上。为了确保大模型能够准确识别标准术语条目中的术语元素，并将标准首选术语中文名称翻译成对应的阿拉伯语和葡萄牙语表述，两者均遵循提示词自动生成、数据输入、记录大模型回答等核心环节。在提示词自动生成环节，本文经过多轮测试分别制定了适用于SparkMax和DeepSeek-R1-32B的提示词模板，通过拼接提示词模板与具体的国家标准文献信息自动生成提示词。在数据输入环节，标准术语语料标注主要是将提示词与标准术语条目输入给大模型，而标准术语衍生语料标注则需要采用自然语言处理技术对标准术语语料标注的大模型回答结果进行结构化处理，将提示词与结构化处理后得到的标准首选术语中文名称输入给大模型。

1.2.5 中医药领域国家标准术语语料库生成

根据中医药国家标准术语语料库的整体框架设计，本文在知识组织视角下，对中医药国家标准术语及衍生语料的标注结果进行数据融合与组织，在符合整体系统环境的要求下生成与标准数字化应用需求相匹配的中医药国家标准术语语料

表1 中医药国家标准术语及衍生语料标注所采用的大模型基本信息及功能

大模型名称	所属大模型家族	是否为蒸馏模型	参数量	应用场景	在中医药国家标准术语语料库构建中的作用
SparkMax	讯飞星火	否	千亿级	适用于涉及数学计算、逻辑推理、文本生成与理解、多模态处理等方面的业务	智能抽取标准术语条目中的标准元素，实现国家标准术语语料标注
DeepSeek-R1-32B	DeepSeek	是	320亿	适用于涉及文本生成与理解、中等复杂程度的数学计算等方面的业务	对标准首选术语中文名称进行智能翻译，实现国家标准术语衍生语料标注

库，并确保中医药国家标准术语语料库能够根据具体的标准数字化应用场景要求转化为不同的交付形式。

1.2.6 中医药领域国家标准术语语料库管理与维护

中医药领域国家标准术语语料库的建设需要长期持续投入。语料库管理与维护主要是为了满足中医药国家标准术语语料库的数据更新与管理、服务方式更新等方面的需求，是语料库建设进入成熟阶段后的必经环节。考虑到当前中医药国家标准术语语料库的建设正处于初步完成阶段，尚不具备长期运维经验，本文就语料库的常规管理与维护要求，明确中医药领域国家标准术语语料库管理制度与风险预案机制的重要方向，包括数据更新、服务更新、容灾备份等方面，后续将针对中医药国家标准术语语料库的实际运行情况进一步落实中医药国家标准术语语料库的管理与维护机制。

2　中医药领域国家标准术语语料库的构建结果与分析

2.1 中医药领域国家标准术语语料库的基本概况

本文所构建的中医药国家标准术语语料库涵盖了中医药领域7 200余条标准术语条目，主要包括中医基础理论与临床诊疗2方面，涉及外科学、内科学、耳鼻喉科学、妇科学等若干领域。由于本文选取的大模型在生成内容上具有伦理、安全等方面的限制，个别涉及大模型敏感词的中医药国家标准术语条目不在本次大模型抽取术语元素的范围内。

2.2 中医药领域国家标准术语语料库的词频统计分析

对中医药国家标准术语语料库中的标准首选术语中文名称进行词频统计，得到词频排序靠前的标准首选术语中文名称主要有“旋耳疮”“肌痹”“骨瘤”“雷头风”“痨病”等，有88.69%的标准首选术语中文名称在中医药国家标准术语语料库中的词频为1。

对中医药国家标准术语语料库的标准首选术语中文名称进行分词，对分词结果进行词频统计并绘制词云图，详见图3。

3　中医药领域国家标准术语语料库的应用前景

中医药领域国家标准术语语料库在“一带一路”倡议相关的国际标准化发展过程中具有广泛的应用前景。本文重点关注标准智能检索、标准智能问答、标准智能编写等标准数智化应用中的核

图3 中医药国家标准术语语料库的标准首选术语中文名称分词后的词云图

心方向，具体如下。

3.1 “一带一路”倡议下的标准智能检索

在“一带一路”倡议背景下的标准智能检索中，中医药国家标准术语语料库主要有2方面的应用：一是从知识融合的层面，通过建立中医药国家标准术语语料库与其他标准知识的关联关系，形成支撑标准智能检索的标准知识图谱；二是从检索增强生成的层面，可将中医药国家标准术语语料库转化为与大模型输入数据相匹配的外部知识库，打造基于大模型的标准智能检索功能。

3.2 “一带一路”倡议下的标准智能问答

与“一带一路”倡议下的标准智能检索类似，面向“一带一路”倡议的标准智能问答同样遵循2方面的应用思路，即构建支撑标准智能问答的标准知识图谱和打造服务于大模型的外部知识库。其与标准智能检索的差异在于标准智能问答更应关注生成内容的合规性，为此应参考相关标准和政策要求，制定适用于中医药国家标准术语语料库的智能问答应用规范。

3.3 “一带一路”倡议下的标准智能编写

对于通用型的标准智能编写，标准术语语料库的作用通常以提供术语编写参考依据为主。为了便于“一带一路”共建国家和地区参与者实现标准编写协作与交流，中医药国家标准术语语料库可在“一带一路”倡议下标准智能编写基础上进一步增加跨语种标准术语对齐的功能，为中医药国家标准术语的传播和使用提供支撑。

3.4 “一带一路”倡议下的标准化相关决策辅助

截至2025年9月，已有百余个国家和多个国际组织正式参与我国提出的“一带一路”倡议，参与者横跨五大洲，通用语言和文化习俗迥异。标准文献是多个利益方协商一致后制定的技术性成果，对沟通质量要求较高。在人工智能时代，与标准化决策相关的AI智能体有望解决“一带一路”倡议参与者的背景差异和地理距离为标准化决策带来的挑战，而中医药国家标准术语语料库可作为该AI智能体知识底座的一部分提升其性能表现。

4 总结与展望

4.1 总结

在标准数字化和智能化转型背景下[19]，本文基于前期研究成果，面向“一带一路”倡议的国际标准化建设需求，提出了基于大模型的国家标准术语语料库构建方法，并在中医药领域构建了中医药国家标准术语语料库，验证了大模型赋能国家标准术语语料库构建路径的可行性。

4.2 展望

中医药国家标准术语语料库的构建与运维是一项长期工程，在后续工作中将继续完善中医药国家标准术语语料库的建设，开展中医药国家标准术语语料库的应用，主要推进以下方面的研究。

4.2.1 制定中医药国家标准术语语料库的质量控制方法

尽管当前许多大模型在知识抽取、智能翻译和语义理解上已达到了较高水平，大模型的输出内容依然无法实现零错误。为了进一步提升中医药国家标准术语语料库的数据质量，后续工作将制定针对大模型生成结果的质量控制方法，重点关注标准术语元素的抽取精确性与共建“一带一路”国家相关语种对应词的翻译准确性，持续完善标准术语语料库的数据治理与知识管理路径。

4.2.2 打造融入中医药国家标准术语语料库的AI智能体

近年来，AI智能体已成为大模型在垂直领域开展应用的重要载体。后续工作将尝试以AI智能体的形式实现中医药国家标准术语语料库的应用场景，包括但不限于标准智能检索、标准智能问答、标准智能编写等。

4.2.3 形成更为广泛的标准知识语义网络

标准知识语义网络不仅是人工智能时代下知识组织与管理的重要形式，也是机器学习环境下标准知识自动化发展的必然方向[20]。当前中医药国家标准术语语料库所能实现的标准知识语义网

络仅局限于标准条目信息范畴。后续工作将建立中医药国家标准术语语料库与其他数据资源的关联关系，形成更为广泛的标准知识语义网络范式，为标准知识服务提供更为坚实的数智底座。

参考文献

[1] 孙莹,张熠,文雁兵,等.标准“软联通”何以促进贸易增长：来自中国与“一带一路”贸易伙伴间的经验证据[J].对外经贸实务,2025,43(4):83–92.

[2] 阚湘苓,朱丹丹,郑蓉,等.文化自信视域下中医药文化遗产保护策略与建议[J].中医药管理杂志,2025,33(1):259–263.

[3] 方思怡.基于大模型的标准术语语料库构建路径与应用场景[J].标准科学,2025(12):138–145.

[4] 方思怡.大模型赋能标准数字化应用的路径思考与发展建议[J].标准科学,2025(6):29–36.

[5] 中国标准研究中心.建立术语语料库的一般原则与方法:GB/T 19101—2003[S].北京：中国标准出版社,2003.

[6] 中国翻译协会.语料库通用技术规范:ZYF 001—2018[S].北京：中国标准出版社,2018.

[7] 方思怡,夏磊.集成电路国家标准语料库ICNSC的构建与分析[J].标准科学,2022(11):38–43.

[8] 夏磊,方思怡,解凌,等.基于BiLSTM模型的冶金领域国家标准指标识别研究[J].中国标准化,2023(3):87–93.

[9] 许平,胡千乔,董建立,等.数据时代标准文献数字资源库的构建策略研究[J].中国标准化,2025(3):51–57.

[10] 方思怡.标准知识图谱的技术路径与应用场景探讨[J].中国标准化,2023(11):49–55.

[11] 方思怡.ISO国际标准知识图谱的构建方法研究[J].标准科学,2024(12):73–77.

[12] 王文,徐天启.中国—巴西“发展中大国关系”典范及其全球治理意义[J].拉丁美洲研究,2025,47(4):102–133.

[13] 王丽影,崔泽宁.全面客观看待中阿贸易投资[J].中国外资,2025(9):34–37.

[14] 闫映宇.中国—巴西贸易关系的发展机遇与挑战[J].对外经贸实务,2024,42(8):41–46.

[15] 李林翰.探究机器翻译的发展现状[J].信息与电脑,2025,37(9):43–45.

[16] 胡新雨,宋博川,仝杰,等.基于大模型的大规模电力数据零样本实体关系抽取方法[J].电力信息与通信技术,2025,23(5):61–67.

[17] 赵衍,张慧,杨祎辰.大语言模型在文本翻译中的质量比较研究：以《繁花》翻译为例[J].外语电化教学,2024(4):60–66.

[18] 全国语言与术语标准化技术委员会(SAC/TC 62).标准起草规则　第1部分：术语:GB/T 20001.1—2024[S].北京：中国标准出版社,2024.

[19] 《国家标准化发展纲要》[J].大众标准化,2024(4):200.

[20] 吕鹏辉,邵建芳,杨善林.基于机标关键词的学科语义知识网络构建研究[J].图书情报知识,2017(2):120–128.