中医典籍汉英平行语料库建构研究

赵　莹　Zhao Ying
(郑州升达经贸管理学院　基础部　郑州　451191)

摘要：本文通过梳理汉英语料库及中医典籍汉英平行语料库的研究成果，针对《黄帝内经》经典英译文本，采用 Python 语言作为建构平行语料库的工具，通过文本预处理、文档识别、段落对齐、句子对齐等环节，最终提出一套具有高度灵活性和实用性的中医典籍汉英平行语料库建构方法体系。为世界各地传播中医典籍的医学价值、文化价值和学术价值，推动中西方医学和翻译学提供了重要的尝试。
关键词：中医典籍；平行语料库；建构研究

党的十八大以来，以习近平同志为核心的党中央站在实现中华民族伟大复兴的战略高度，对传承和弘扬中华优秀传统文化作出一系列重大决策部署。中医典籍作为中华优秀传统文化的一个分支，凝结了先贤医家的毕生经验与感悟，中医精神存在于华夏历史的肌肤之中，浸润于亿万百姓的日常生活之中。2019 年，国务院总理李克强指出，中医药学是中华民族的伟大创造。新中国成立以来，我国中医药事业取得显著成就，为增进人民健康做出了重要贡献。尤其是 2019 年新冠病毒爆发以来，中医将其归为疫病范畴，各种对症的中药组方在治疗方面起到了重要作用[i]。在全国中医药大会上，习近平总书记指出，中医药学包含着中华民族几千年的健康养生理念及其生活实践经验，是中华文明的瑰宝，凝聚着中国人民和中华民族的博大智慧。

《黄帝内经》作为中国传统医药学典籍，其英译对中医文化走向世界有着举足轻重的作用。然而，目前国内中医英语缺乏统一标准，没有有效信息搜索工具，检索效率低下[ii]。同时，我国传统历史典籍术语库的建设也存在不足，其重点主要集中在翻译方法和策略等理论研究上[iii][iii][iii][iv]，对翻译实践与术语库检索关注较少。随着机器学习的发展，同计算机技术结合而兴起的语料库建设为语言研究、翻译研究等提供了最好的平台[vii]。20 世纪 90 年代以来，我国在医学英语语料库的建设取得了一定的进步，包括支持向量机 (support vector machine，SVM) 和 BERT (bidirectional encoder representations from transformers) 等算法的实体识别技术成为医学语料库构建的重要组成部分[viii][ix][x]。

基于此，为了更好地实现中医典籍传承与发展，本文采用 Python 语言，建构了以《黄帝内经》中英版本为基础的中医典籍汉英平行语料库，并为术语库的构建打下基础。中医典籍不仅是医学传播的工具，更是构建中国话语体系与彰显文化身份的有效路径。中医典籍汉英平行语料库建构可以作为中华医学研究领域的重要建设资源，可以为中医领域研究者及中医典籍爱好者提供数据支撑，最终实现借助互联网平台和平行语料库资源向世界传播中华优秀传统文化的目标。

一、中医典籍汉英平行语料库研究近况

近年来互联网技术飞速发展，各行各业都开始和互联网产生了密切联系，这对整个社会的经济发展起到了变革性作用。中医典籍汉英平行语料库的建设可以和互联网融合为“互联网+中医典籍语料库”全新研究模式。通过互联网 web 端、移动端，打造成为中医典籍英译基础性数据资源。中医典籍汉英平行语料库建构将给互联网中医典籍的传播提供数据资料，为人工智能、大数据处理提供数据库支撑。

近 20 年间，基于语料库对中医典籍英译进行研究的论文有 15 篇，其中核心期刊 8 篇。其研究内容主要涉及中医典籍、中医术语、四大名著。唐国顺[xi]自建双语对应语料库，走机器翻译的捷径将其应用到中医文献的实际翻译中，其最大的特点就是建设周期快、专业性强、成本廉价、便于推广；叶晓等[xii]创建了多个《黄帝内经》英译本双语平行语料库，对中医“脉象”名称的英译进行了探讨；朱剑飞[xiii]也从语料库的视角介绍了建构《黄帝内经》双语语料库的方法；姚丽娟[xiv]基于语料库的《黄帝内经》两个英译本的规范化翻译进行了研究。由此可见，由于《黄帝内经》英译本丰富，为中医典籍语料库研究提供了资源保障。除此之外，海霞等[xv]自建微型语料库，对《伤寒论》中方剂名译法进行对比分析研究。在中医汉英语料库研究做出贡献的王小芳等[xvi]基于语料库对中医术语中的治湿诸法术语、脉诊术语翻译做出了研究，探讨了较为合理的译法。

二、中医典籍汉英平行语料库的研究思路、原则及方法

建构中医典籍汉英平行语料库具有创新性及应用价值，既是语料库应用研究的新拓展，又是对中医教学、中医英语教学、中医学术英语研究的新贡献。本文以《黄帝内经》为研究对象，通过文本转换、段落对齐等手段，设计了一套完整的平行语料库建构方法。

（一）中医典籍汉英平行语料库建构研究思路

为满足中医药传承发展需求，顺应中医药与大数据技术结合的潮流，按照语料库建构的一般要求。首先在深入分析语言学及相关学科研究成果基础上，重点梳理汉英平行语料库领域中医典籍英译研究成果，然后对《黄帝内经》等经典英译文本进行甄别和筛选，通过专业软件进行转换和对齐，最后构建出完整的英汉平行语料库，为传播中医典籍医学价值、文化价值和学术价值，推动中西方医学和翻译学发展提供智力支持。

（二）中医典籍汉英平行语料库建构研究原则

本文建构的中医典籍汉英平行语料库为专门用途的语料库，遵循三个关键原则。一是真实性原则，采集的语料具有真实性，方能保证语料库的应用价值。因此收集语料时，需保证原文和译文出自相关领域正式出版物或文献，而非网络版本以及机器翻译后未经人工校对的译文。二是代表性原则[xvii]，《黄帝内经》成书于春秋战国时期，是我国医学宝库中最早的一部医学典籍。它奠定了中华传统医学的基础，被后人称之为“医之始祖”。本语料库的语料皆来源于《大中华文库》之《黄帝内经·素问》汉英对照本（全三册，李照国 英译，刘希茹 今译）。三是规范性原则，即要符合一定的学术规范。语料库的建构既符合语言学的学术规范，也符合人工智能的技术规范，从文本选择、格式转换、词句对齐等方面有一套完整的学术规范和技术体系，便于同行和使用者进行复制和检验。

（三）中医典籍汉英平行语料库建构研究方法

首先对不同版本的《黄帝内经》进行比对和筛选，最终选取清晰度较高的三卷本的 PDF 版本进行处理。采用 Python 语言作为构建平行语料库的工具，通过文本预处理、文档识别、段落对齐、句子对齐等环节，最终建立了一套具有高度灵活性和实用性的中医典籍汉英平行语料库的方法体系。

三、中医典籍汉英平行语料库建构技术路径

以《黄帝内经》三卷本作为语料信息来源，经扫描生成图片格式的 PDF 文档，分别对应 I、II、III 卷，其内容是中英对照（古文无英译对照，白话文有英译对照）。针对上述处理对象，本文提出的汉英平行语料库的构建方法如图 1 所示。

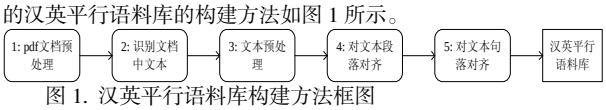

图 1. 汉英平行语料库构建方法框图

（一）PDF 文档预处理

图 1 中圆角矩形内左上角数字表示操作步骤。步骤 1 是将 PDF 文档中与建构平行语料库无关的页面去掉。这些页面或与《黄帝内经》内容无直接联系，或无中英文对照。这些页面包括封面、扉页、总序、前言、译文括符号使用说明、目录、插图以及相关的英文注解。其定位信息由人工操作完成，之后交给程序做筛选处理。步骤 1 处理后，得到的 PDF 文档共有偶数个页面，中文页面在前，英文页面在后，交替进行，包含了绝大部分构建平行语料所需内容。

（二）输入识别命令

步骤 1 得到的 PDF 文档仍是扫描图片，无法利用程序自动化处理，还需利用文本识别程序识别出其中的文本信息，即图 1 中步骤 2。文本识别采用的 OCR (Optical Character Recognition)库是 Tesseract V4.0.0。Tesseract V4.0.0 是开源的 OCR 库，源码存在于 GitHub 上。基于其开源的特性，可以添加接口，丰富其功能，并且其包含的 Leptonica 组件有着优越的图像分析性能，能够保证文字识别的精度。Tesseract V4.0.0 进行识别时需预先告知所识别的文本语言，因此在实现识别时，采用偶数页识别中文、奇数页识别英文的规则。

步骤 2 处理后的文本信息仍需相关的预处理，因为其中含有以文言文形式存在的中文，及由边缘灰色像素产生的乱码信息。上述这些文本信息应删除，采用先构造这些信息正则表达式，再将这些信息替换为空的方法，从而达到删除的效果。这些预处理即图 1 中步骤 3。

（三）语料段落对齐

步骤 4 是将步骤 3 生成的双语语料进行段落对齐，即将汉语与英语语义等价的段落映射在一起。根据步骤 3 生成的语料文本特征，我们制定以下规则进行段落对齐：

汉英按章节序号匹配对应。该规则比较简单，例如，将中文 1.1 节起始的部分与英文 1.1 起始的部分对应，其余类推。

若相互对应的汉英某个章节序号后跟随了多个段落，则我们利用模式“\n\n”将第 1 段后的属于本序号的其他段落进行对齐。

上述段落对齐过程中，通过编写相应的 Python 程序以实现。本步骤利用语料源固有的章节序号信息及已识别的文本中分割段落的模式信息，具有较好的段落对齐效果。

（四）语料句对齐

双语语料库句对齐是构建平行语料库的一个关键步骤，目的是将原文与译文建立句子级别的对应关系，保证双语检索的提取效果[3]。语料句对齐的软件工具选择较多。如计算语言学研究者经常使用的 Tmxmall，但是由于我们在构建语料库时大量的任务是基于 Python 语言完成的，所以本文图 1 中步骤 5 句对齐任务，我们选用开源的句对齐 Python 工具包 Bilingual-Sentence-Aligner (https://github.com/aswinpradeep/Bilingual-Sentence-Aligner)。该工具包采用 Google 推出的语言无关 BERT 句子嵌入向量（Language-agnostic BERT Sentence Embedding，简称 LaBSE）模型。LaBSE 模型能为 109 种语言生成语言无关的跨语言句子嵌入，并且该模型对没有数据的低资源语言也有效。因此，Bilingual-Sentence-Aligner 保证了本文语料库建构过程中句对齐的高精度效果。

Bilingual-Sentence-Aligner 的接口要求输入两个 txt 文件，分别对应平行语料库的两种语言，并且每个文件需做分句处理。因此以存放汉语的 ch_ss.txt 文件作为输入 1，以存放英语的 en_ss.txt 文件作为输入 2，且对这两个 txt 文档做了分句处理，然后进行句对齐。根据汉英语言特性的不同，汉语分句采用基于 Python 语言的 SnowNLP 包，英语分句采用基于 Python 语言的 NLTK 包，其比较经典且常用的。经上述处理后，ch_ss.txt 和 en_ss.txt 就可以交给 Bilingual-Sentence-Aligner 做句对齐操作了。图 2 是 Bilingual-Sentence-Aligner 的运算过程，表 1 是句对齐结果示例。

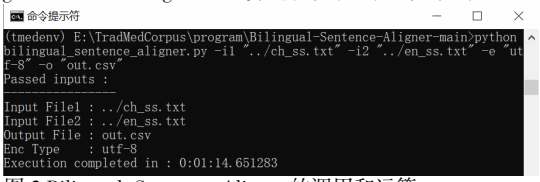

图 2 Bilingual-Sentence-Aligner 的调用和运算

表 1 Bilingual-Sentence-Aligner 句对齐结果示例（部分）

COLUMN-1	COLUMN-2	SIMILARITY
到了 35 岁，阳明经脉衰弱，面部开始憔悴，头发开始脱落；	At the age of thirty-five, Yangming Channel starts to decline, her face begins to wither and her hair starts to lose.	0.7676883020433408
1.4 岐伯回答说：女子 7 岁，肾气就开始旺盛，牙齿开始更摘，毛发生长；	1.4 Qibo answered, For a woman, her Shenqi (Kidney-Qi) becomes prosperous and her teeth begin to change at the age of seven.	0.7653792181404953
到了 28 岁，筋骨强健，头发最为旺感，身体最为强壮；	At the age of twenty-eight, her musculature and bone become strong, her hair grows long enough. Her body has reached the summit of development.	0.7138830630064855
1.3 黄帝问道：人老了就不能生育子女，是精力衰竭了呢，还是自然规律所限呢?	1.3 Huangdi asked, Old people cannot give birth to any children.	0.7026549027782316
正因为如此，不当嗜好不能扰乱他们的视听，淫邪之举不能惑乱他们的心境，无论是愚笨的人、聪明的人、有才能的人还是无才能的人，都不会因外物而动其心，所以符合养生之道。	That is why improper addiction and avarice could not distract their eyes and ears, obscenity and fallacy could not tempt their mind.	0.7025656574610918

表 1 中第 1 列是汉语句子，第 2 列是英语句子，第 3 列是 2 个对齐句子的相似度。从表 1 中可看出，Bilingual-Sentence-Aligner 是按相似度由大到小顺序输出结果的，并且能做到一对一、一对多、多对一情况的句子对齐，绝大部分做到了精准对齐，个别地方还需微调。

（五）语料标注

由于本研究只选用了一种语料信息来源，即世界图书出版公司出版的《黄帝内经·素问》三卷本，无过多的元信息，因此这一步骤被省略。另外，《黄帝内经·素问》中含有大量的中医术语，它们多以文言文的形式存在，这就涉及到这些术语的信息标注。这将是我们未来深入研究的选题。

四、结语

中医典籍汉英平行语料库对中医英语教学、语言学、中医药文化传播学、纸质和网络辞书编撰等具有重要的参考价值。中医典籍汉英平行语料库的建设涉及机辅翻译、领域术语自动抽取等相关语言信息处理技术，可以借助这一技术方法，开发中医术语自动识别、提取技术以及机器辅助术语自动翻译技术。借助语言信息处理技术方法减少译者的中医术语翻译工作量，进而有效促进中医典籍翻译质量的提高。

目前，国内中医典籍汉英平行语料库的建构尚处于起始阶段，有大量的问题需要解决，还需与中医理论紧密结合。本文以世界图书出版公司出版的《黄帝内经·素问》三卷本为例，创新性地提出了建构中医典籍汉英平行语料库的方法，本文方法的实现全部基于 Python 语言代码级实现，具有高度的灵活性和个性化。在此基础上，我们将扩大语料信息来源，进行中医术语语料标注，以使中医典籍汉英语料库具有更大的社会应用价值。

参考文献：

（下转第 313 页）

（上接第 311 页）

[1]张倩.中医药科技期刊新型冠状病毒肺炎相关论文出版概况分析及建议[J]. 学报编辑论丛, 2021(00):596-601.

[2]窦川川，余静，李婷玉 . 中医汉英双语平行语料库的研制与应用研究 [J]. 当代教育实践与教学研究，2017（11）：25+13.

[3]吴丽萍，王岩 .《墨子》中哲学术语的翻译研究 [J]. 海外英语，2016（3）：124-125.

[4]宋依然. 中医典籍翻译报告：以《中医诊断学》翻译为例 [D]. 沈阳：沈阳师范大学，2017.

[5]牛海燕，刘凯，蒋辰雪，等. 生态翻译学关照下中医典籍中养生术语英译研究 [J]. 环球中医药，2018（10）：1618-1620.

[6]郭惠琴. 中华典籍翻译方法与策略研究：以《庄子》核心思想术语的英译为例 [J]. 汉字文化，2019（3）：101-102.

[7]王克非. 中国英汉平行语料库的设计与研制 [J]. 中国外语，2012（6）：23-27.

[8]杨锦锋，关毅，何彬，等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报，2016，27(11):2725-2746. YANG J F， GUAN Y， HE B， et al. Corpus construction for named entities and entity relations on Chinese electronic medical records [J]. Journal of Software，2016， 27(11):2725-2746.

[9] GAO Y， GU L， WANG Y， et al. Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes [J]. BMC Medical Informatics and Decision Making，2019， 19: 67-78.

[10]崔博文，金涛，王建民. 自由文本电子病历信息抽取综述[J/OL]. 计算机应用[2020-12-05]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210105.1351.004.html.

[11]唐国顺.以双语对应语料库快译中医文献的研究[J].中国科技翻译，2014，27（4）：24-27.

[12]叶晓，董敏华.中医“脉象”名称英译探讨——基于两个《黄帝内经》英译本中的脉象英译比较[J].中国中医基础医学杂志，2015，21（1）：94-96.

[13]朱剑飞.《黄帝内经》英译研究的语料库视角[J].中国中医基础医学杂志，2015，21（9）：1161-1164.

[14]姚丽娟.基于语料库的《黄帝内经》两个英译本的对比研究启示[J]. 湖北函授大学学报，2017，30（3）：180-181.

[15]海霞，丁东.基于《伤寒论》双语平行语料库的中医方剂名称翻译方法探析[J].南阳理工学院学报，2018，10（5）：102-107.

[16]王小芳，刘成.基于语料库的中医术语英译研究[J].科技视界，2019（9）：215-216.

[17]江莉，王茎，张四红，等.基于语料库的中医术语翻译中essence 和 spirit 的差异性研究[J].科教文汇（上旬刊），2013（3）：109-111.

[18]叶晓. 基于多个中医术语英译标准的治湿诸法英译辨析[J]. 中国中医基础医学杂志，2019，25（9）：1307-1310.

[19]刘成，董益敏，王小芳. 基于语料库的中医脉诊术语英译规范探讨[J].中华中医药杂志，2019，34（11）：5064-5068.

[20]王克非. 英汉/汉英语句对应的语料库考察 [J]. 外语教学与研究， 2003(6):410－416，481.

作者简介：赵莹（1983-），女，河南安阳人，文学硕士，郑州升达经贸管理学院副教授，主要研究方向：外国语言学与应用语言学。

基金项目：本文系河南省哲学社会科学规划项目“多模态医学经典术语英语语料库构建与应用研究”（2020BYY018）的阶段性成果。