科技大数据背景下的中英双语语料库的构建及其特点研究

苏晓娟[1] 张英杰[2] 白 晨[2] 吴 思[2]
（1.北京石油化工学院，北京 102617；2.中国科学技术信息研究所，北京 100038）

摘要：首先通过对双语语料库全过程构建的描述，提出基于专业领域词库快速构建双语语料库的方式，并用于快速发现科技大数据基础语料的多属性，完成语料的标注，这对于科技大数据知识检索、知识图谱方面的应用具有基础性支撑作用。然后通过分析新时期科技大数据对语料库构建的要求，从期刊、专利中选择“分布式能源”主题数据集，结合“神经网络机器翻译 + 统计机器翻译”的机器翻译技术，构建形成20834个双语词对初试语料集，利用中国科学技术信息研究所专利数据库、德温特专利数据库形成6428条专利数据对双语词对初试语料集进行测试应用。最后从忠实度、流畅度、可理解度3个方面进行人工评测。

关键词：科技大数据；双语语料库；机器学习；语料库构建；机器翻译引擎
中图分类号：G354.4 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2019.06.013

Research of Bilingual Corpus Construction and Its Characteristics in Big Data

SU XiaoJuan[1], ZHANG YingJie[2], BAI Chen[2], WU Si[2]
(1. Beijing Institute of Petrochemical Technology, Beijing 102617; 2. Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Firstly, based on the description of the whole process of constructing bilingual corpus, this paper puts forward a fast way to construct bilingual corpus based on specialized field lexicon. It can be used to quickly discover the multiple attributes of basic corpus of science and technology big data and complete the marking of corpus, which plays a fundamental supporting role in knowledge retrieval and knowledge mapping of science and technology big data. And then, we construct a corpus of 20834 bilingual word pairs for the preliminary test by analyzing the requirement of large scientific and technological data for corpus construction in the new era, selecting the subject data set of “distributed energy” from journals and patents, and combining the machine translation technology of “neural network machine translation + statistical machine translation”, 6428 patent data are generated from ISTIC Patent Database and Derwent Patent Database to test the bilingual corpus. Finally, the whole process of building bilingual corpus is described through manual evaluation in three aspects: adequacy, fluency and intelligibility.

Keywords: big data, bilingual corpus, machine learning, corpus construction, machine translation engine

作者简介：苏晓娟（1978—），女，北京石油化工学院讲师，研究方向：语言学；**张英杰**（1979—），男，中国科学技术信息研究所副研究馆员，研究方向：数据治理（通信作者）；**白晨**（1980—），女，中国科学技术信息研究所副研究员，研究方向：资源共享；**吴思**（1992—），女，中国科学技术信息研究所研究实习员，研究方向：信息资源管理。

基金项目：中国科学技术信息研究所重点工作“面向中信所资源大数据建设的多源异构数据库内容获取与融合平台建设（二期）”（ZD2019-04）。
收稿日期：2019 年 6 月 21 日。

1　概述

语料库特指能够被计算机存储的数字化语料库，广泛应用于编撰字典、语言教学、自然语言处理、人工智能等方面。随着各学科、各领域科学研究的融合发展，语料既能够记录各类科技活动，反映特定时期的科技发展特征，也能够支撑科技大数据的丰富应用场景。

国外的语料库比较著名的有欧盟的术语数据库（IATE）、美国当代语料库（COCA）等。欧盟术语数据库（IATE）开始于1999年，收录有1017288条实体，7961980个术语，旨在为所有欧盟术语资源提供基于网络的基础设施，提高信息的可用性和标准化[1]。美国当代语料库（COCA）是美国唯一一个大型的、类型均衡的语料库，也是最广泛使用的英语语料库，包含超过5.6亿字的文本，涉及口语、小说、流行杂志、报纸和学术文本等内容[2]。在商业领域，图灵机器人目前拥有1300多亿条对话语料库。

国内的语料库比较知名的有国家语委现代汉语通用平衡语料库[3]和中国科学技术信息研究所编制的《汉语主题词表》。国家语委现代汉语通用平衡语料库立项于1991年，全库约有1亿字符，语料选材类别广泛，时间跨度大。标注语料库为国家语委现代汉语通用平衡语料库全库的子集，约5000万字符，准确率大于98%[4]。中国科学技术信息研究所编制的《汉语主题词表》，多年来持续建设和维护的中文基础词库收录词汇总量达到500万条，包括中文叙词表、全国科学技术名词审订委员会审定公布的规范名词、文献关键词、专业词典、术语标准、百科等多种来源的词汇，词汇信息丰富，包括词间关系、词汇分类、英文、注释等属性。此外，部分有特色的语料库包括中国传媒大学的新词语研究资源库[5]、哈尔滨工业大学信息检索研究室开发的对齐双语句对的语料库[6]、清华大学的汉语均衡语料库TH−ACorpus[7]、中国科学院计算技术研究所的跨语言语料库[8]。国内商业领域的搜狗、网易围绕中文新闻语料库也积极进行了探索实践。

从科技语料库的构建研究来看，MatildeT−revisani依托科技文献语料，探讨了通过词生命周期聚类的方式从科学语料库中进行知识动态发现[9]。英国曼彻斯特大学的Nhung T.H.利用生物多样性文献，建立了一个COPIOUS语料库，提出了服务于生物命名实体的金标准[10]。国内语料库的研究主题主要涉及语言文学、教育学、计算机科学、临床医学、图书馆和情报学。其中，语言文学的研究主题主要聚焦于平行语料库、语料库语言学等内容，如谢家成[11]自建了60万对的平行语料库；王克飞[12]开展了中国英汉平行语料库的设计实践；教育学主要涉及翻译教学、英语写作、外语教学、自主学习等主题，如秦洪武等[13]、方秀才[14]、张宇[15]围绕语料库与教学实践，就翻译教学中的理论依据和实施原则，中国英语教学与语料库结合的成就与不足等问题开展研究；计算机科学涉及语料库的自然语言处理、词性标注、语音合成等研究主题，尤其是在大规模语料库的词性标注方面，张虎等[16]提出了基于主题聚类和分类的语料库词性标注一致性检查新方法，保证大规模语料库标注的正确性；图书馆、情报学主要利用大型科技文献数据库、搜索引擎，抽取其中的关键词构建知识元词库，进行基于语料库的对比研究以及围绕主题词语料库的研究，如李淑平[17]提出了基于语料库的主题图式构建，李佳[18]以科技论文中英文关键词、主题词作为语料库开展了跨语言检索平台研究。

在新时期科技大数据蓬勃发展的背景下，科技资源建设的重点已经不限于单一来源、单一维度数据的开发、应用，更多的是通过整合不同的数据，揭示新规律，发现新关系，支撑新决策，形成新的情报服务模式。本文试图探讨在科技大数据日益复杂、多样的情况下，以现有科技大数据中已有的自标注中英文语料为基础，通过机器学习的方式，形成双语语料库的构建流程，并对最终的双语语料库进行忠实度、流畅度和可接受度评测基础上，总结上述构建方法的优势与特点。

2　机器学习双语语料库构建

2.1　双语语料库的构建

以“分布式能源”为主题，进行机器学习双语语料库的构建实验。对数据的获取及处理方法是以“分布式能源”为检索词，首先在中国科学技术信息研究所科技大数据仓储 1.2 亿条的数据中检索相关期刊论文，随后提取其中有中英文摘要和关键词的论文，最终形成中英文关键词词对的对齐、匹配，累积形成 20834 个双语词对初试语料集。同时，以“分布式能源”和“Distributed Energy”为检索词，在中国科学技术信息研究所专利数据库、德温特专利数据库中进行检索，下载、查重后形成 6428 条专利数据，然后分别按照语种提取其中的专利形成中文专利数据集和英文专利数据集，供后续实验应用。

在实验中，利用了新译科技公司的机器翻译引擎进行训练，基本过程是将 20834 个双语词对初试语料集导入机器翻译引擎，经过机器翻译引擎自我学习、深度学习后，生成一个机器翻译模型，使用“神经网络机器翻译+统计机器翻译”，图 1 为机器翻译引擎训练示意图。

相比于传统的统计机器翻译（Statistical Machine Translation, SMT），神经网络机器翻译（Neural Machine Translation, NMT）已经在翻译、对话和文本概要总结方面获得非常好的成绩。领域术语语料库在整个过程中保证专业词汇在翻译过程中的专业性和一致性。在训练模块，系统挂载自有术语库，以确保翻译结果精确度更高，翻译结果更加符合业务场景。

译文评估主要从忠实度、流畅度和可接受度 3 个方面开展。忠实度是评测译文是否忠实地表达了原文的内容，按 0–5 分打分，打分可含一位小数，最后的得分是所有打分的算术平均值。流畅度是评测译文是否流畅和正宗，按 0–5 分打分，打分可含一位小数，最后的得分是所有打分的算术平均值。可理解度则是从用户的角度对最终的翻译结果进行评测，如表 1 所示。

以下是一个轮次的实验步骤。

第一步：利用没有经过训练的机器翻译引擎，对中外文专利进行互译；

第二步：各选取 100 条数据，对机器翻译结果进行人工校对，按照表 1 的标准对忠实度和流畅度进行打分；

第三步：根据检索词（略），抽取相关期刊论文，将其中的中英文词对进行抽取，梳理成对应的词对；

第四步：将期刊的中英词对用于机器学习，学习后的语料对先前提供的数据集进行二次翻译；

第五步：选取第二步已经校对过的数据，进行二次人工校对判别，按照表 1 的标准分别对忠

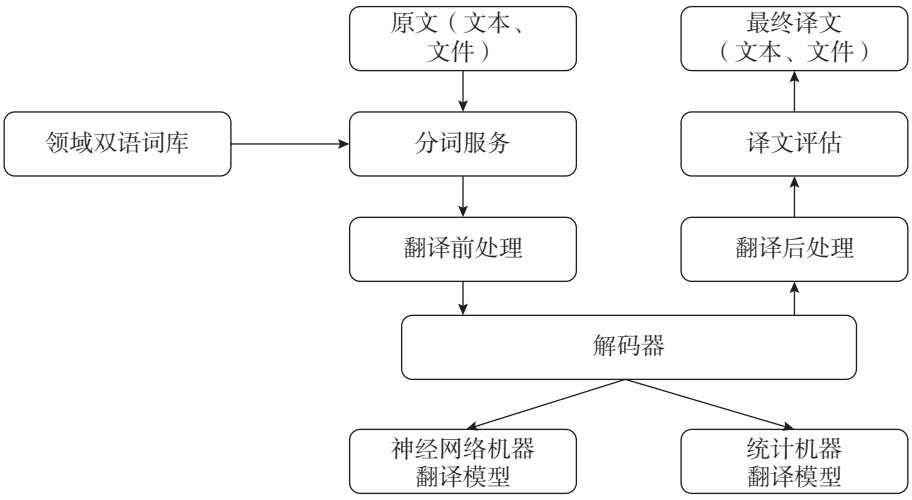

图 1　机器翻译引擎训练示意图

实度和流畅度进行打分；

第六步：选取中英文同族的专利，将机器翻译的结果和原始提供的译文进行比对，按照表2的标准对可理解度进行打分。

第七步：总结语料构建的流程和效果。

2.2　双语语料库的评测及结果分析

2.2.1　实验结果分析

在整个实验过程中，邀请了10名北京石油化工学院外语系英语专业2015级学生参与评测，共进行两轮打分，第一轮对机器翻译前的数据打分，第二轮对机器翻译后的数据打分。

从表3可以看到，在“神经网络机器翻译+统计机器翻译”的机器翻译模型中，虽然模型自身有一定的迭代优化，但双语专业领域词库在机器翻译中扮演着重要的角色，故在双语语料库的构建中要加大双语专业语料的建设工作，同时，也需要根据多主题科技领域的复杂性、时效性特色，进行神经网络模型参数的优化。

此外，从用户接受的角度，最终语料的可理解性依赖于通用语料和专业语料的结合，通过机器学习迭代后，从统计结果来看，“分布式能源”最终的双语语料基本能达到80%的级别。

为了检验译文在专业双语词库介入前后翻译效果，对介入前后忠诚度和流畅度的人工评价结果进行检验，来分析专业双语词库的效果，结果如表4所示。首先检验数据是否服从正态性，当正态性检验的W值对应的概率小于0.05，则认为数据不服从正态分布，四组数据的$Pr<W$均小于0.05，说明这四组数据都不服从正态分布；其次，选择检验统计量，对于不服从正态分布的数

表1　忠实度、流畅度打分标准

分数	忠实度得分标准	流畅度得分标准
0	完全没有翻译出来	完全不可理解
1	译文中只有个别单词和原文相符	译文晦涩难懂
2	译文中有少数内容与原文相符	译文很不流畅
3	译文基本表达了原文的信息	译文基本流畅
4	译文表达了原文的绝大部分信息	译文流畅但不够正宗
5	译文准确完整地表达了原文信息	译文流畅而且正宗

表2　可理解度打分标准

分数	得分标准	可理解程度
0	完全没有译出来	0%
1	看了译文不知所云或者意思完全不对，只有小部分词语翻译正确	20%
2	译文有一部分与原文的部分意思相符；或者全句没有翻译对，但是关键的词都孤立地翻译出来了，对人工编辑有点用处	40%
3	译文大致表达了原文的意思，只与原文有局部的出入，一般情况下需要参照原文才能改正译文的错误。有时即使无需参照原文也能猜到译文的意思，但译文的不妥明显是由于翻译程序的缺陷造成的	60%
4	译文传达了原文的信息，不用参照原文，就能明白译文的意思；但是部分译文在词形变化、词序、多义词选择、得体性等方面存在问题，需要进行修改。不过这种修改无需参照原文也能有把握地进行，修改起来比较容易	80%
5	译文准确流畅地传达了原文的信息，语法结构正确，除个别错别字、小品词、单复数、正宗性等小问题外，不存在很大的问题，这些问题只需进行很小的修改；或者译文完全正确，无需修改	100%

表3　双语翻译结果打分统计表

	第一轮	第二轮
忠实度	2.6	3.4
流畅度	2.3	3.8

据可以采用非参数的Wilcoxon秩和检验，主要看*Pr*>|*Z*|对应的概率，若小于0.05，那么就有95%的把握认为两组数据存在差异。从结果来看，忠诚度和流畅度均存在显著性差异，同时结合均值，可以认为通过搭配专业双语词库进行的翻译与原来的方法相比，在忠诚度和流畅度方面都有非常杰出的表现。

2.2.2　案例实证分析

表5分别选择一条中文专利原文和英文专利原文，比较应用“领域双语词库”后的译文效果，以对最终的中英双语语料进行评估。

从第一条的中译英情况来看，第二次的译文与第一次的译文相比，在忠实度和流畅度方面都有所提高，特别是第二次译文中，热电联产系统（CHP）这样的专业词得以体现，专业领域方面表现良好，且在译文样式方面跟原文相比，具有较好的吻合度。

从第二条的英译中情况来看，第二次的译文相较于第一次的译文，在用词方面更为紧凑、准确，对用户表现出较好的可理解度，但在整体的流畅、完整性方面还有待进一步优化。

2.3　双语语料库构建方法的优势分析

经测评和实证分析表明，本文构建的中—英双语语料库通过结合神经网络机器翻译和基于统计的机器翻译方法，对于后续双语语料库的构建具有以下优势。

（1）提高语料库选择与处理效率。采用神经网络机器翻译和基于统计的机器翻译方法，可大幅度地减少语料选择和处理的人工工作量，并在专业主题数据库基础上，快速形成专业领域的双语语料库，相应构建语料库的时间可从按年规划缩短到按月响应，同时由于大量训练样本和算法的介入，处理所耗费的人工也大幅度得以减少。

（2）快速发现基础语料多属性。基于神经网络的自学习机制，可从期刊、专利等规范化文本中，快速发现多属性的基础语料，从而丰富基础语料的多属性值，完成语料的标注，支撑科技大数据知识检索、知识图谱方面的应用。

表4　忠诚度和流畅度的假设检验结果

维度		均值	正态性检验		Wilcoxon秩和检验	
			W	*Pr*<*W*	*Z*	*Pr*>|*Z*|
忠诚度	介入前	2.6	0.620152	<0.0001	1.2734	0.0100
	介入后	3.4	0.872957	0.0713		
流畅度	介入前	2.3	0.839089	0.0095	5.2838	<.0001
	介入后	3.8	0.873825	0.0731		

表5　双语语料效果案例

序号	原文	第一次译文	优化后译文	忠实度		流畅度		可理解度
1	一种热电联产系统控制机房机柜改良结构	Improved structure of room cabinet of cogeneration system control room	One Combined heat and power（CHP）system control Room cabinet improvement structure	2.3	3.5	2.8	3.6	译文传达了原文的信息，专业词选用得体；样式符合要求
2	System for controlling battery energy management system for smart grid, has battery energy storage system for charging or discharging electricity, and power conversion device whose side is connected with power meter	用于智能电网的电池能量管理系统的控制系统,具有用于充电或放电的电池能量存储系统,和与功率计连接的功率转换装置	智能电网蓄电池能源管理系统，具有充电或放电储能装置，一侧与功率计连接的功率转换装置	2.2	3.1	2.6	3.2	译文传达了原文的信息；专业术语选用的当；行文样式有待优化

（3）提升双语语料库的工程化构建水平。语料库是人工智能时代的基础工程，已经从传统的文本语料向图片、声音、视频等全媒体语料库转变，其服务模式也从传统的纸质语料向Web接口、API等多种方式提供对外共享应用服务，本文验证的快速语料库构建方法，可以促进各专业领域语料库的工程化水平。

3　启示与结论

本文通过分析新时期科技大数据对语料库构建的要求，从期刊、专利中选择“分布式能源”主题数据集，结合“神经网络机器翻译+统计机器翻译”的机器翻译技术，最后通过人工评测的方式，描述了进行中英双语语料库构建的全过程。我们发现，在人工智能技术发展的大背景下，通过综合利用人工智能技术、大数据技术，新型的语料库构建模式不仅满足了语言学自身的发展，而且通过工程化的语料库构建开发专业领域语料库和服务标杆语料库，在诸如生命科学、人种语音等新兴前沿领域，都处于专业领域语料库的建设期，这为本文构建的双语语料库的实施方法提供了丰富的应用场景。

参考文献

[1] Europe union terminology [EB/OL].[2019−05−30]. https://iate.europa.eu/home.

[2] Corpus of Contemporary American English (COCA) [EB/OL].[2019−05−30].https://www.english−corpora.org/coca/.

[3] 现代汉语语料库[EB/OL].[2019−05−30].http://www.cncorpus.org/.

[4] 国家语委现代汉语语料库[EB/OL].[2019−05−30]. http://corpus.zhonghuayuwen.org/CnCindex.aspx.

[5] 中国传媒大学文本语料库检索系统[EB/OL].[2019−05−30].http://ling.cuc.edu.cn/RawPub/.

[6] 哈工大信息检索研究室语料库[EB/OL].[2019−05−30].http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm.

[7] 汉语均衡语料库TH−ACorpus[EB/OL].[2019−05−30].http://www.lits.tsinghua.edu.cn/ainlp/source.htm.

[8] 中国科学院计算技术研究所跨语言语料库[EB/OL].[2019−05−30].http://mtgroup.ict.ac.cn/new/resource/index.php.

[9] MATILDE T, ARJUNA T. Chronological corpora curve clustering: From scientific corpora construction to knowledge dynamics discovery through word life−cycles clustering[J]. MethodsX, 2018(5): 1576−1587.

[10] NGUYEN N, GABUD R, ANANIADOU S . COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature[J/OL]. [2019−06−15].https://doi.org/10.3897/BDJ.7.e29626.

[11] 谢家成. 小型英汉平行语料库的建立与运用[J]. 解放军外国语学院学报, 2004, 27(3):45−48.

[12] 王克非. 中国英汉平行语料库的设计与研制[J]. 中国外语, 2012, 9(6):23.

[13] 秦洪武, 王克非. 对应语料库在翻译教学中的应用：理论依据和实施原则[J]. 中国翻译, 2007(5):49−52.

[14] 方秀才.基于语料库的英语教学与研究综述:成就与不足: 根据22种语言学类CSSCI来源期刊近30年的统计分析[J].外语电化教学,2012(3):19−24.

[15] 张宇.语料库在英语教学中的应用研究[D]. 杨凌:西北农林科技大学, 2016.

[16] 张虎, 郑家恒, 刘江. 语料库词性标注一致性检查方法研究[J]. 中文信息学报, 2004, 18(5): 11−16.

[17] 李淑平. 基于语料库的主题图式构建[J]. 情报科学, 2009(3):402−405.

[18] 李佳. 基于词共现的跨语言检索平台研究[J]. 情报杂志, 2015(8): 195−198.

（上接第86页）

[5] 国家国际科技合作与交流专项2015年度报告[R/OL]. http: //www.istcp.org.cn/2015科技年报/index.html.

[6] 邵丽珍, 季庆庆.基于项目的地方国际科技合作能力提升路径研究: 以江苏省常州市为例[J].科技管理研究, 2017年(23): 119−123.

[7] 吴建南, 郑长旭, 姬晴晴. “一带一路”战略实施与国际科技合作创新[J].情报杂志, 2016, 35(4): 32−36.

[8] 付岩.发达国家科研创新机构科技成果转移转化的特点及启示: 以德国弗劳恩霍夫应用研究院和日本科学技术振兴机构为例[J].中国科技资源导刊, 2017, 49(3): 97−103.