多语言科技语料库建设研究*

曾文[1]，张均胜[1]，徐红姣[1]，李颖[1]，刘敏[1]，屈鹏[1]，刘丹[2]
(1. 中国科学技术信息研究所，北京 100038；2. 北京大学图书馆，北京 100871)

摘要：多语言科技语料库建设的重要意义在于它能够服务于多语言科技文献信息的组织、科技文献的自动翻译，以及科技文献的情报分析等。科技语料库的建设采用的主要技术方法是运用自然语言处理和计算机处理技术实现语料的采集、自动加工和处理。本文介绍多语言科技语料库建设方面的相关研究工作，主要涉及多语言词表、平行语料的获取与处理，及多语言语法资源的建设等方面的工作成果。研究工作的不足之处在于语料库的数据资源和语法资源的质量和规模有待于提高和完善。

关键词：多语言；科技；语料库
中图分类号：G35
DOI：10.3772/j.issn.1673-2286.2015.08.008

科技文献数据规模大，来源分布广泛，质量良莠不齐，内容深度千差万别，用户对数字化科技文献数据资源的服务需求日益递增，需要科技文献的服务质量和原文传递服务模式进行相应的改变，开发和探索多元服务模式和技术方法。特别是，面对日益增长的海量外文科技文献，建立高效的数据信息组织和分析处理方法，对于科技文献服务质量的完善和提高具有重要的现实意义，而实现这一目的重要基础是具有高质量的、可应用的多语言科技语料资源，而且多语言科技语料的建设对于图书情报领域的信息组织和情报分析研究具有重要的价值。多语言科技语料库与传统通用语料库的内容不同[1],多语言科技语料库更注重科学技术性和多语言性的特点，多语言科技语料库的内容是以科技文献数据库作为科技语料的主要数据来源之一。多语言科技语料库主要包括：多语言科学技术词表，多语言科技平行语料数据资源，多语言的语法资源库等。其中词表是科技文献资源信息的有效组织工具之一；平行语料是实现自然语言处理技术，进行文本挖掘，实现智能信息处理和机器翻译的数据基础[2]；多语言的语法资源则是自然语言处理中非常重要的环节，语法资源的丰富和完善程度都是保证自然语言处理效果的基础。本文将重点从词表、平行语料和语法资源三个方面来介绍我们目前在多语言科技语料库建设方面的研究工作及所取得的工作成果。

1 多语言科技语料库建设的现状和主要问题

目前，国内外对于语料库建设基本是通用的语料库，也包含一定数量比例的科技文本，如美国国家语料库，英国国家语料库，国内有北京大学计算语言所的汉语新闻语料库等。国内语料库的应用更多偏重于语言本身，特别是英语语法的使用方法[3,4]，国外对于科技数据资源的建设研究则偏重于数据的存储和发布服务，其中以英语为母语的欧美国家，研究偏重于多语言科技文本数据资源有效组织和利用的研究工作[5-7]。

（1）词表研究：欧洲共同体已经集成构建名为Eurovoc的叙词表（Eurovocabulary thesaurus），它可以支持欧盟22种官方语言。AGROVOC是由联合国粮农组织和欧盟委员会于上世纪80年代开发的农

* 本研究得到国家社会科学基金项目“基于事实型科技大数据的情报分析方法及集成分析平台研究”（编号：14BTQ038）和中国科学技术信息研究所预研资金项目“多语言科技语料库建设与应用研究”（编号：YY2015-08）资助。

业多语言主题词表，其领域涵盖农业、食品、渔业、林业、环境等领域。美国DIALOG的DIALINDEX、BRS的CROSS、SDC的Database Index以及ESA的Quest Index等是词表集成的典型应用。在国内，我国学者借鉴国外的研究经验相继开展了一些汉语词表的研究工作，主要有中国医科院医学信息研究所研制的“医学分类主题一体化系统建设”和“统一的中国医学语言系统”、中国中医研究院中医药信息研究所研制的“中医药一体化语言系统”等，但国内对于其它语言词表的汉化及研究方面存在实践成果不多、缺乏方法的深入探讨等问题。

（2）科技平行语料的建设：平行语料是进行多语言数据分析和研究的数据基础。从服务的角度看，国内外科技用户对于非母语科技文献的阅读和研究需求都是相当大的，中国国内用户对外文科技文献翻译服务方面的需求较大，尤以英文科技文献的翻译服务需求为主。因此，平行语料的建设显得尤为重要，国内外建设平行语料的方法基本相同，主要是采集和处理多语言数据集，在国内更为重视中英文平行语料的建设，并服务于中英机器翻译的研究工作。

（3）语法资源建设：目前被广泛使用的是宾州大学的英文和中文树库语法资源，最著名的语言资源管理系统是语言资源联盟（LDC），它由宾州大学建设并负责维护。该平台上提供各种语法资源，其不足之处在于：语法资源的内容还不够丰富，用户的研究和使用成本较高。在国内，中文语言资源联盟与LDC一样，语言资源具有局限性且需要花钱购买，而北京大学计算语言所以及清华大学的语法及树库则只提供简单的介绍，对用户不提供服务。此外，现有的语法资源对科技文献存在的语法现象并未涉及。

2 多语言词表研究

实现高效的科技信息服务的基础是实现对科技文献数据的有效组织。词表作为代表文献资源内容特征的、起关键作用的词集，是实现对文献资源有效、快速检索的基础。数据资源的有效组织是提供数据快速检索和保证数据分析的重要手段。例如，叙词表通过建立术语集，以及术语之间的用、代、属、分、参等关系，实现对数据信息资源的标引和组织。因此，词表是图书文献情报学领域中重要的一种数据资源组织工具。但是现有的词表已经无法满足图书馆工作人员对不同学科、不同语言的科技文献数据信息进行有效组织的需求，客观上需要加强对多语言词表的研究和应用，以更好的实现对科技文献数据的标引和组织。

针对前文所指出的国内对多语言词表研究不足问题，中国科学技术信息研究所语言技术与知识技术研究组采用机器辅助翻译和人工翻译校正相结合的方法，已经完成对英语EI（Engineering Index工程索引）叙词表和日本JST（Japan Science and Technology Agency）叙词表的汉化研究工作，其中EI叙词表，总收词量19296条，其中叙词9926个，非叙词9370个，族首词85个。日本JST叙词表，包含37163个日语叙词，2635个日语非叙词，词间的相互关系为20多万对。

对于多语言词表的研究，主要是词表的汉化翻译方法研究，其主要工作流程包括：词典匹配，机器翻译系统自动翻译，以及人工汉化三个部分。多语言词表汉化翻译的主要的研究方法包括：借助专业领域翻译词典，通过与词典词条的匹配进行英语、日语的自动汉化翻译；借助机器翻译软件进行词表的翻译汉化；通过人工汉化，完成自动翻译叙词表中未实现的术语和词条的翻译工作。为了节省人力、物力和财力，我们研究和开发了叙词表辅助汉化平台辅助人工进行术语汉化，以呈现叙词表中包含的概念和语义信息，保证数据的安全性、完整性和汉化结果的可恢复性。叙词表辅助汉化平台主要功能是实现术语汉化，它借助于叙词表本身的信息及各种外部资源，为汉化工作人员提供多来源的辅助汉化信息。其中，辅助汉化信息包括两类：一类是叙词表本身提供的有助于确定叙词准确含义的信息，这部分信息主要通过叙词表展示模块中提示的信息来获取；另一类是各种翻译资源提供的翻译参考信息，包括翻译词典匹配的翻译结果，机器翻译系统翻译结果等。此外，平台提供词表的浏览、搜索和修改等管理功能。图1和图2分别是EI和JST叙词表辅助汉化平台的界面。

对于汉化翻译结果的评价，采用按类别选词验证和随机选词验证相结合的方法，参与评价的人员是没有介入汉化工作的领域专家。以EI叙词表的汉化翻译结果评价为例：分别选取类别码为400series（Bridges and Tunnels，桥梁与隧道）、700series（ELECTRICAL ENGINEERING，电器工程）和800series（CHEMICAL ENGINEERING，化学工程）的术语，其中抽取叙词和非叙词共计1772个，翻译结果的准确率为98.8%。

图1 EI叙词表辅助汉化平台用户界面

图2 JST叙词表辅助汉化平台用户界面

3 平行语料的获取与处理

平行语料（Parallel corpora）长期以来在机器翻译研究领域中被广泛用于获取词对知识，构建翻译词典，以提高机器翻译的翻译质量[8-9]。此外，平行语料库对科技文献信息的组织和分析同样具有重要的研究意义。基于平行语料，可以辅助建立有效的科技文献组织方法和模型，实现多语言科技文献的有效组织和检索，为科技文献的深层次分析服务。平行语料的获取有两种方式：一种是自动处理对齐质量良好的多语言文本来获取平行语料；二是购买多语言数据，进行平行句对的二次加工。平行语料的获取效果主要依赖于翻译质量的高低，特别是对于前者更是如此，其采用的技术方法主要是利用统计计算模型来自动实现两种或两种以上语言的句子级自动对齐处理，利用机器翻译技术实现两种或两种以上语言的句子和术语的自动匹配和抽取处理。

中国科学技术信息研究所语言技术与知识技术研究组基于已经拥有的美国工程索引EI英汉主题词表和日本科技振兴机构JST的日汉科学技术主题词表，获取英汉双语词1万以上，日汉双语词4万以上。英汉平行句对150万以上，日汉平行句对60万以上，日英的文献标题对2000多万。基于这些已有资源，通过采用计算机自动处理技术，我们在平行语料方面取得了一定的成果积累。例如，我们基于国家知识产权专利局网站（http://www.sipo. gov.cn）上公布的汉英国际专利分类表（IPC）第8版（高级版）进行平行语料的建设，如图3所示，通过人工抽样评测，平行语料的获取准确率约为90%。平行术语的获取是通过在已经完成分词和词性标注的中外文语料中分别形成各自的词语单元集合，并组成对应的汉外词语等价单元，最后通过计算、分析和处理它们之间的关联概率值来解决汉外的词语对齐问题。关联概率值的计算是通过统计汉外词语单元在双语句对中的共现概率来确定的。分别将每个汉语词汇对应的所有外语词语单元的关联概率值，以及每个外语词汇对应的所有汉语词语单元的关联概率值均进行降序排列，则可以生成双语词语集合，通过对生成的词语集合进行匹配分析过滤，最终得到双语术语的匹配生成结果，如图4所示。

采用随机抽样的方法，人工对生成的术语对进行正确率的判断，正确率的计算方法如下：

P_{right}%=（正确的词对数目+0.5×部分正确的词对数目）/词语集合中词对的总数×100%

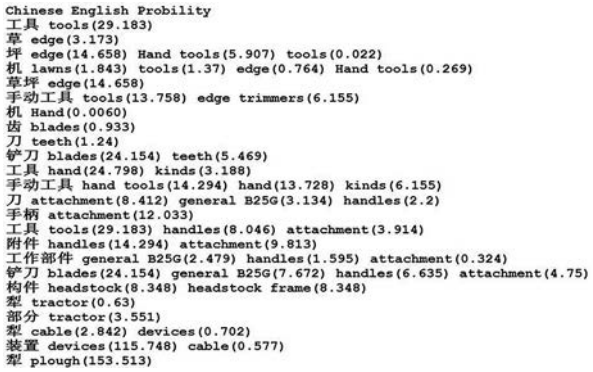
经过计算可知：构建的双语术语集合的正确率为78.8%。

```
en_content                                                    ch_content
Hand tools (edge trimmers for lawns A01G ) ","手动工具(草坪修整机入A01G )"
Spades; Shovels ","锹; 铲"
with teeth ","带齿的"
Hoes; Hand cultivators ","锄; 手动中耕机"
with a single blade ","带单铲刀的"
with two or more blades ","带双铲刀或多铲刀的"
with blades provided with teeth ","带齿铲刀的"
with teeth only ","只带齿的"
Tools for uprooting weeds ","拔草工具"
"Tong-like tools ","钳式工具"
"Combinations of different kinds of hand tools ","不同类型手动工具的组合"
"Attaching the blades or the like to handles (handles for tools, or their attachment, in ge
"for treating meadows or lawns  ","处理草地或草坪用的 "
"Ploughs with fixed plough-shares ","装有固定式犁铧的犁"
"Man-driven ploughs ","人力犁"
"Animal-drawn ploughs ","畜力牵引犁"
"without alternating possibility, i.e. incapable of making an adjacent furrow on return jou
"Swing ploughs ","摆杆步犁"
"Trussed-beam ploughs; Single-wheel ploughs ","桁架式犁梁; 单轮犁"
"Two-wheel beam ploughs ","双轮犁梁"
"Frame ploughs ","架式多铧犁"
"Alternating ploughs, i.e. capable of making an adjacent furrow on return journey ","双向犁
"Turn-wrest ploughs ","翻转犁壁双向犁"
"Balance ploughs ","平衡犁"
"with parallel plough units used alternately ","装有双向使用的平行犁机构的"
```

图3 IPC 中英文平行语料库的局部内容示例1

4 多语言语法资源的建设

语法资源的建设是开展自然语言处理研究的基础[10-11]，语法资源的丰富和完善程度影响到自然语言处理的效果，建设良好的语法资源可以支持许多领域的



```
Chinese English Probility
工具 tools(29.183)
草 edge(3.173)
坪 edge(14.658) Hand tools(5.907) tools(0.022)
机 lawns(1.843) tools(1.37) edge(0.764) Hand tools(0.269)
草坪 edge(14.658)
手动工具 tools(13.758) edge trimmers(6.155)
机 Hand(0.0060)
齿 blades(0.933)
刀 teeth(1.24)
铲刀 blades(24.154) teeth(5.469)
工具 hand(24.798) kinds(3.188)
手动工具 hand tools(14.294) hand(13.728) kinds(6.155)
刀 attachment(8.412) general B25G(3.134) handles(2.2)
手柄 attachment(12.033)
工具 tools(29.183) handles(8.046) attachment(3.914)
附件 handles(14.294) attachment(9.813)
工作部件 general B25G(2.479) handles(1.595) attachment(0.324)
铲刀 blades(24.154) general B25G(7.672) handles(6.635) attachment(4.75)
构件 headstock(8.348) headstock frame(8.348)
犁 tractor(0.63)
部分 tractor(3.551)
犁 cable(2.842) devices(0.702)
装置 devices(115.748) cable(0.577)
犁 plough(153.513)
```

图4 IPC 中英文平行语料库的局部内容示例2

汉语	英语	日语
物流配送	distribution	分布型
填料	filler	充てん剤
锁环	retaining ring	止輪
移行	migration	マイグレーション【ソフトウエア】
编队	formation	構造
抽出	pick-up	ピックアップ【溶接】
急腹症	acute abdomen	急性腹症
分光器	spectrometer	分光計
浸透液	penetrant	浸透剤
声音	sound	音
计划	arrangement	配列
孤残峰	outlier	アウトライア
冷却装置	cooling system	冷却用装置
惯例	usage	用法
外套	barrel	樽
端盖	end plate	端部プレート
视像	view	景色

图5 中英日平行语料示例（平行术语）

图6 中日文平行语料示例（平行句对）

研究工作，特别是对科技文献内容的语义分析具有重要的价值，例如，基于不同语言的语法资源，实现对科技文献内容的自动语法解读和分析，将可以较好的提高科技文献信息的挖掘和分析效果。

目前，我们的研究仅涉及汉语和英语的语法，研究工作尚处于初级阶段，已初步构建了基于宾州树库的中文树库ISTIC-Tree和词汇化树邻接语法树（Lexicalized Tree Adjoining Grammar, LTAG）LTAG-Tree树库的管理平台，初步实现对宾州大学的中文树库和词汇化树邻接语法树（Lexicalized Tree Adjoining Grammar, LTAG）树库的管理，具体包括：

（1）语言资源的存储：宾州大学中文树库的文件形式是文本文件，只有一个文件，每一行为一颗中文树库的语法表示形式， LTAG树资源也是以文本文件形式存储，有多个文件，分类存放，平台采用XML格式和关系数据库同步进行资源的存储和管理，将这些信息组织成容易读取的形式，并梳理这些资源相互之间的关系；（2）语言资源的管理：目前的语言资源还是相对分散，分散存储，分散下载，分散利用。没有形成统一的管理平台，本平台利用MVC（Model模型，View视图，Controller控制）架构的思想将资源的模型、控制和显示相互分离。模型是该平台的业务逻辑，作为类的方法封装在每一个对象类的内部；控制部分是模型与视图之间沟通的桥梁，用于分派用户的请求并选择恰当的视图以用于显示，同时也解释用户的输入并将它们映射为模型层可执行的操作；视图部分用于与用户的交互，通过网页编程语言JSP来实现。例如，ISTIC-Tree可以分页浏览树库，查看树的详细信息，对树进行编辑、修改、删除和添加符合语法规则的树。LTAG-Tree可以分页浏览LTAG树族（Family）库、树库、词汇库，检索词法信息等，分别如图7和图8所示。

图7 查看ISTIC-Tree树详细信息示例

图8 词汇化树邻接语法（LTAG）句法词典的检索结果

该平台为语言研究和信息处理提供一个有利的数据分析平台：（1）语法资源中的句法标注作为分词、词性标注和语义标注的中间环节，将为下一步的语义标注工作打下良好的基础；（2）语法资源蕴涵着丰富的句法信息，它为研究者提供了带有句法标记的汉语句法知识，我们能够从中获得有关句法的各种信息。例如，从词类入手，可以考察某一特定类别词语的句法功能；从短语功能类型入手，可以考察某一特定类型短语的内部构造模式等等；（3）语法资源可以进行数据统计，信息抽取等工作，为情报分析、计算语言学等领域的研究提供便利条件；（4）标注各种句法和语义关系的语法资源对于语言研究具有重要意义，像谓词论元的标注将大大提高机器翻译、信息检索、信息抽取等技术的发展，使信息处理更加准确，有益于情报分析与预测。

5 结语

多语言科技语料库建设是开展数字化科技文献有效服务和进行数据分析的基础，不仅可以服务于数字图书馆及情报学研究人员的科技文献信息组织和分析工作，同时可以向用户提供更高质量的、更具价值的国内外科技文献信息翻译服务。目前，我们的研究工作还处于发展阶段，例如，在科技语料加工工具和语法分析处理工具等问题的研究工作尚有欠缺，仍需改进。此外，还需注重语料库资源和功能多样化的研究工作，特别是在词表的互操作技术方面、科技语料的多语言性以及不同语言的语法资源等方面仍需完善，以丰富和完善现有的科技语料库资源，使其更具有实用性。

参考文献

[1] 张东,王惠临. 关于建立中国国家科学技术语料库的思考[J].图书情报工作，2010,54（6）: 102-106.

[2] Berna Altinel,Murat Can Ganiz, Banu Diri. A corpus-based semantic kernel for text classification by using meaning values of terms[J]. Engineering Applications of Artificial Intelligence,2015,43(8):54-66.

[3] 冷雪莲. 基于COCA语料库辨析英语同义词Capable和Competent[J]. 成都师范学院学报,2015,31(2): 54-58.

[4] 赵勇,施应凤,罗瑞等. 基于语料库和数据驱动的英语同义词的构式语法研究[J].文山学院学报,2015,28（1）: 75-77,98.

[5] ISO5964:Guidelines for the establishment and development of multilingual thesauri [S/OL].[2011-09-05].http://www.iso.org/iso/catalogue_detail.htm?csnumber-12159.

[6] Reinhard Rapp1.The automatic generation of thesauri of related words for English, French, German, and Russian[J].International Journal of Speech Technology, 2009(11):147-156.

[7] Technical Committee ISO/TC46. ISO CD 25964-1 Information and Documentation Thesauri and Interoperability with other Vocabularies-part1 Thesauri for Information Retrieval [EB/OL].(2012-12-10)[2013-02-24].http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ic-s.htm?ics1=01&Ics2=140&ics3=20&csnumber=53657.

[8] 宗成庆. 统计自然语言处理[M].北京:清华大学出版社, 2008:10-15.

[9] 刘洋. 树到串统计翻译模型研究[D]. 北京:中国科学院研究生院, 2007.

[10] 苏劲松. 基于形式化句法的统计机器翻译若干问题研究[D]. 北京:中国科学院研究生院, 2011.

[11] 熊德意. 基于括号转录语法和依存语法的统计机器翻译研究[D]. 北京:中国科学院研究生院, 2007.

作者简介

曾文，1973年生，博士，副研究员，研究方向：智能信息处理、数字图书馆，E-mail：zengw@istic.ac.cn。

Multilingual Science and Technology Corpus Construction

ZENG Wen[1], ZHANG JunSheng[1], XU HongJiao[1], LI Ying[1], LIU Min[1], QU Peng[1], LIU Dan[2]
（1. Institute of Scientific and Technical Information of China, Beijing 100038, China; 2. Peking University Library, Beijing 100871, China）

Abstract: The important significance of the construction of multilingual corpus is that it can serve the organization of the information of multilingual, automatic translation of scientific documents, and information analysis of scientific literature. The main technology methods of science and technology corpus construction are that use natural language processing and computer technology to realize the automatic collection, processing and processing of data. This paper introduced the research work about multilingual science and technology corpus construction， it included multilingual vocabulary, parallel corpus acquisition and processing, and multilingual grammar resources construction, etc. And the deficiencies of the research work are that the quality and size of corpus data resources and grammatical need to be improved.

Keywords: Multilingual; Science and Technology; Corpu

（收稿日期：2015-07-24；编辑：雷雪）