情报工程 ISSN: 2095-915X
情报工程 第 1 卷 第 3 期 2015 年 6 月
第1卷 第3期 64-75 2015 年 6 月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.3 64-75 Jun 2015

doi:10.3772/j.issn.2095-915x.2015.03.009

汉语科技词汇构词研究初探

周雷 1,2，李颖 1，石崇德 1（1. 中国科学技术信息研究所 北京

摘要：基于机器学习的分词模型可以借助科技词汇构词特征分析提升其在科技领域的适应性，本文对传统语言学的句法构词、韵律构词、语义构词几个方面理论进行总结归纳，融合术语学研究理论，围绕提升分词准确率的目的，提出了适用于科技词汇的构词特征标注系统，并对标注系统的结构进行了规划。这为科技词汇构词特征标注工作完成了前期的探索，为后期批量标注，辅助分词等环节提供了基础依据。

关键词：汉语科技词汇，构词法，词标注

中图分类号：TP391.2，H031

An Exploration on Chinese Word Formation in Science and Technology

ZHOU Lei1,2 , LI Ying1, SHI Chongde1 (1. Institute of Scientific and Technical Information of

Abstract: To improve the adaptability of word segmentation model in S&T domain, more features of S&T terms are needed. Based on the exploration on syntactic, rhetoric and semantic method of word formation, as well as terminology, tags are extracted and a labeling system is roughly designed aiming at improving the accuracy of word-parsing system. The research work on S&T word formation is not only the preliminary exploration of S&T terms tagging, but also the foundation of large size tagging and word segmentation.

Keywords: Chinese science and technique terms, word formation, word tagging

基金项目：本研究得到国家自然科学基金项目“面向科技监测的实体识别与关系抽取研究”（编号：71403257）的资助。**作者简介：周雷（1987-），助理翻译，研究方向：自然语言处理，email：zhoulei@wanfangdata.com.cn，联系电话：010-58882726；李颖（1964-），博士，副研究员，研究方向：知识组织与知识工程、语言技术，email：liying@istic.ac.cn，联系电话：010-58882470；石崇德（通讯作者）（1979-），博士，助理研究员，研究方向：自然语言处理，email：shicd@istic.ac.cn，联系电话：010-58882447。**

1 引言

随着机器学习技术引入中文分词领域，构建分词模型、利用人工标注语料对分词模型进行训练成为现阶段中文分词的有效方法之一。但机器学习分词模型主要依赖大量的人工标注语料进行训练，语料情况对于分词准确率影响较大。将基于词典的特征抽取算法与机器学习方法相结合，是进一步提升分词准确率的一种有益尝试，以往的研究主要进行了两方面的探索：一是面向常用语篇的构词规则的深入研究。如 Meishan Zhang Chinese Parsing Exploiting Charaters 一文在树库的基础上，在词内部进行序列标注，利用字的前后特征辅助边界判断，同时增加了“ Direction of head character ”（中心词位置）信息，将构词结构融入了以往的标注体系中 [1]。再如赵亮的《基于构词法的中文分词方法研究》一文，通过对传统构词法的归纳总结，在《人民日报》语料库标注体系的基础上增加了句法构词规则标签，据此对语料进行标注后对分词及句法分析系统进行训练和改进 [2]。构词规则的研究可以提供更多特征信息，但由于这类研究通常借助新闻语料库进行标注方法的改进，应用于特定学科领域时会受到相应学科语料库情况的影响。二是面向特定学科语篇的构词规则研究。如《化工专业词典结构设计及中文分词系统的开发》[3]，《航空术语的构词分析》[4]，通过某一学科领域的深入研究，总结其术语构词规律，为学科内文献分词提供更多特征信息。此类研究对于特定学科的分词算法改进有着积极意义，但由于需要深厚的学科背景，这类研究方法很难快速复制到多个学科领域。科技文献分词是中文科技文献信息处理的基础，由于科技领域至今仍缺乏大量的人工标注语料，机器学习分词算法的优势难以充分发挥，在机器学习分词算法基础上结合科技词汇构词特征，可以一定程度上解决分词模型的科技领域适应性问题。论及构词特征，科技词汇内部构造既受到汉语常用词汇、短语构造规律制约，也受到学科概念体系、命名规则的影响。中文科技文献处理领域应用术语学理论亦有较多先例，如《广义后缀树及其在汉语科技词系统中的应用研究》[5]，即是从构建汉语科技词汇词系统的角度对于中文术语概念耦合和具备属性标识的特点进行了论证和利用。将传统语言学的构词研究和术语学相关研究应用于汉语科技词汇分词，是本研究的主要思路。本文对科技词汇构词研究进行了初步探索，通过研读传统语言学构词、造词、短语构成等方面的研究，对其理论成果进行归纳和抽象，融合术语学相关研究，构建了易于实现人工标注的科技词汇构词标注体系，并对标注系统进行了规划，提出借用系统简化人工标注工作的设想。

2 构词理论研究

2.1 构词研究的现状

张科的《汉语构词法研究综述》[6], 和鲁晓娟的《汉语构词法研究综述》[7] 两篇文章皆认为从共时的角度（区别于历史的角度），构词法可以分为传统型和反传统型，传统型以应用句法规律对合成词进行构词分析和应用语音规律对单纯词进行分析为多，而反传统型则在语义构词、韵律构词等方面有更多探索。2.1.1 传统构词法研究传统构词研究受到结构主义语言学影响，主要利用句法理论对合成词的构词进行成分分析。句法构词体系相对稳定，近年大多研究都集中于对原有体系的继承性讨论。《现代汉语》（黄伯荣、廖旭东版）[8] 对词结构的分类可以说是传统构词法的综合总结，其分类体系见图 1：

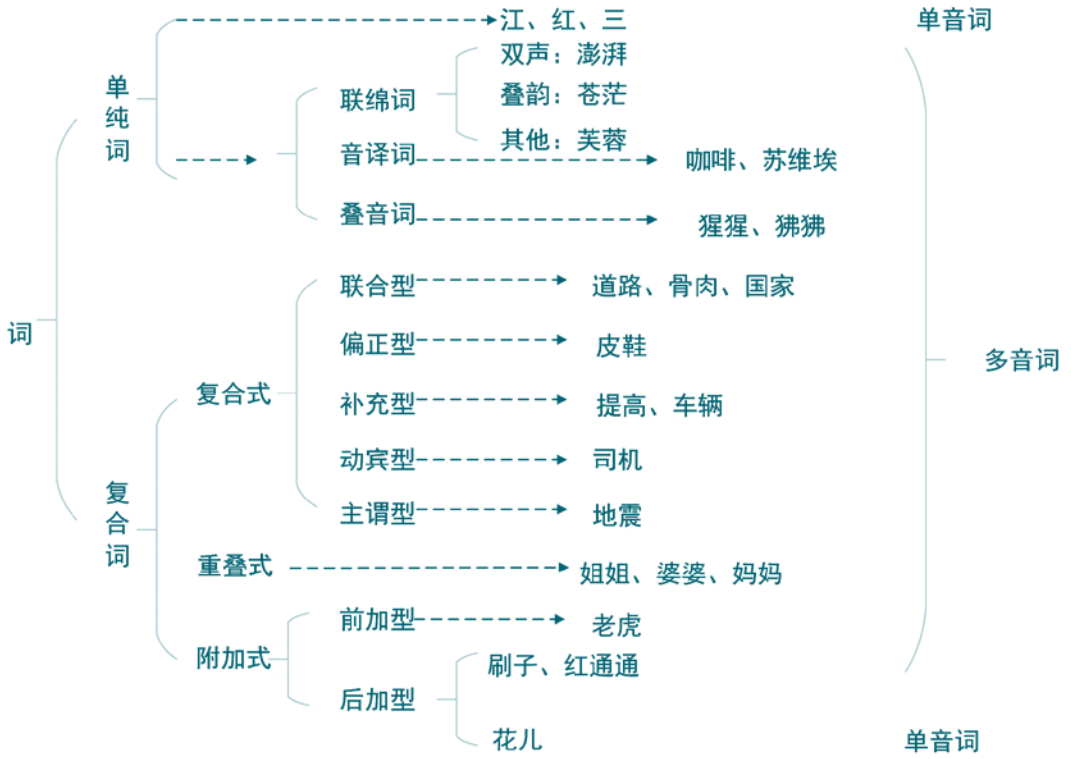

图 1 《现代汉语》词的结构类型表

2.1.2 韵律构词研究韵律构词属于反传统型构词研究。冯胜利所著的《汉语的韵律、词法与句法》[9] 一书提出：音步遵循“二分枝”原则，音步是一个节奏片段，最少要两个音节才能体现“轻重抑扬”的节奏，因此标准音步包含两个音节。由于韵律词至少包含一个音步，因此韵律词至少包含两个音节。韵律词在汉语中普遍相对稳定是由于标准音步具有绝对优先实现权。单音成分附着在标准音步上，形成超音步，超音步构成超韵律词，其使用受到一定的语境限制，但相对于退化音步要更灵活一些。而单音节通过停顿或拉长元音形成的蜕化音步，则不属于韵律词，其使用有着相对严格的语境限制。由于汉语音节与语素具有对应关系，因此不止语言过程，复合词构词过程也会受到韵律的影响。比如三音复合词大多有双音词和单音成分组合而来，相对少见三个语素分别作为主谓宾成分的情况。再如语义上变化不大的凑补，如“咸盐”、“玉石”，又如长词的缩略时优选二到三字的作为缩略语等情况。这种在语法、语义差别不大的情况下的构词方式选择通常是受到韵律规则驱动而产生的。2.1.3 语义构词研究句法构词研究的兴起与结构主义语言学的发展是密不可分的，而现代汉语作为分析语缺乏形态变化，缺乏明显的词性标志，因而用句法理论进行分析存在较多有争议的情况。比如“运动是绝对的”这一语句中，“运动”做主语时是转变为动名词，还是体现了动词的指代功能。后来随着系统功能语言学和认知语言学的发展，逐渐有学者提出应从认知和思维的角度探寻汉语语义组合的驱动力，还原汉语在思维中的组织过程，而将语义分析理论和方法带入构词研究，形成了语

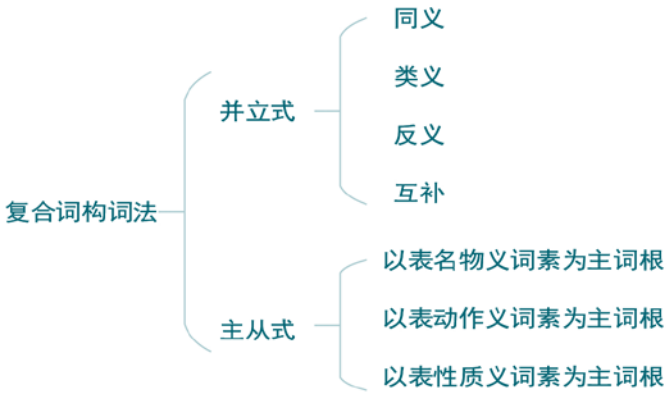

图 2 《汉语构词法和造词法研究》复合词构词法体系

义构词研究方法。语义构词主要研究复合词构词，早期的研究受句法构词影响较多，与句法构词体系仍有交叉，直到李仕春的《汉语构词法和造词法研究》[10]，提出了相对独立的构词体系，见图2。朱彦的《汉语复合词语义构词法研究》[11] 通过述谓结构分析的方法对复合词深层语义结构进行了描述，并对语义框架进行了形式化的表达，是对语义构词研究理论和方法的极大丰富。

2.2 科技词汇构词特征分析

2.2.1 字符特征汉语科技词汇字符构成相较于汉语常用词汇的字符构成有其自身特性。首先，科技词汇内部具有更多字符。以万方科技词表为例，其中又四个字符组成的词最多，占总词数的 30.78%，其次是三个字符和五个字符组成的词，分别占总词数的 17.67% 和 17.36%，而六个字符的词也要占到总词数的11.86%。而《现代汉语常用词表》中则是两个字符组成的词即双音词最多，占到总词数的 71.51%，三音词和四音词分别是 12.01% 和 10.61%。对比可以看出，科技词汇总体字符数要比汉语常用词汇字符数更多，科技词汇当中既有词汇结构的特点，同时又有短语结构的特点，相对汉语常用词要更为复杂。第二，科技词汇包含更多数字、字母、符号等特殊字符。这部分字符通常会出现格式不一致的情况，因此需要对这部分字符进行预处理。第三，用字特征。科技词汇中存在一部分在科技领域应用率高而常用词汇少用或几乎不用的一部分汉字，包括化学、化工、药学领域中有关化学物质的一类字，如醇、肽、氨、钾等；有关疾病的字，如癌、疹、癣等；有关动物体组织的字，如胰、膀胱等。这类字由于在常用语料中出现的几率相对较少，在处理过程中需要这类语素的语义判断其在构词过程中的角色，因此在标注过程中有可能需要查询相关学科资料。2.2.2 词性和构词特征科技词汇的词性及构词方面也与常用词有一定差异。科技词汇的词性类别相对比较集中，以名词、动词、形容词为主，鲜见单独的介词、连词，基本不包含感叹词、拟声词；从句法构词角度来说以合成词居多，鲜见连绵词、叠音词。科技词汇造词过程受到相关学科理论的影响较多，这使得其构词与常用词有一定差别，同时学科内的术语有其特有的命名原则。其中比较典型的是化学介词，如“氢氧化钾”一词，“化”作为化学介词，将“氢”“氧”“钾”三个元素名词连缀形成化合物名称，因此在对化合物名称

进行标注时需要注意对其内在关系的理解。《现代术语学引论》[12] 中提出“在现代汉语术语中，语缀比在一般汉语词汇中要丰富的多”“这些语缀与主要的语素结合，明显表示整个术语的意义内涵。这些语缀也可以与单词或者词组组合起来，成为其中的一个辅助成分”。科技词汇中的词缀与常用语中完全虚化的词缀略有不同，科技词汇词缀语义丰富，且通常具有明显的词义特征和学科特征，因此，在对科技词汇进行分析时词缀进行特殊标注，并将词缀作为单独的分析对象是非常有必要的。科技词汇具有系统性，这是术语规范化的结果，相同类别的概念通常有着相似结构，这一点即体现在词缀上，也体现在具备“限定成分 + 语义核心词”这一结构的科技词汇当中。例如“神经系统”、“列表假脱机系统”、“临时插入系统”，再如“连接装置”、“连续给纸装置”“料斗锁定装置”，这类词虽然分属不同学科，但由于他们概念相近使得其对应词汇具有相同的语义核心词。组成科技词汇的单词数量较小，其构造过程符合“术语形成的经济率”[13]，既用少量的单词构造大量术语的现象。这种特性使得未登录词中包含经过标注的成分的概率相对较大，若能提取成分及其位置和序列信息，总结规律，则可以为未登录词的自动识别提供更多依据。

3 构建标注体系

3.1 整体思路

构建体系的出发点在于依据标注体系对科技词汇进行标注，从标注结果中提取构词特征，用于机器学习分词系统，因此标注体系需要具备一下几个特点：(1) 简单。基于机器学习的分词系统需要大训练语料，过于复杂的标注体系影响标注进度、增加错误几率、致使具有统一特征的样本量偏少，不利于分词系统的训练。因此标注体系宜简单明了。(2) 易于判断。由于需要快速生成大量训练语料，应减少标注系统对操作人员学科知识及语言学基础的依赖，力求做到可以利用基本常识进行简单的查询完成标注。(3) 允许模糊。科技词汇是各个学科概念在语言环境中的映射，体系相当复杂，很难从规则描述的角度对其构词特征进行完整概括，因此设计系统时应当允许模糊的分类存在，可以通过边标注边修正对其进行完善。

3.2 需要解决的问题

3.2.1 科技词汇复杂结构的拆分问题科技词汇中存在大量短语型科技词汇，因此有必要对其进行切分切分单位。根据韵律构词法分析的观点，韵律词包括标准韵律词（双音节）和超韵律词（三音节）两种类型，四音及以上词是由韵律词组成的。因此切分时将韵律词作为基本单位。切分标识。切分时用“/”这一符号将不同片段分隔开。词缀处理。由于词缀与其所附着的词根关系密切，因此有词缀的情况下优先将词缀与其所附着的词根视为一个切分片段。而后再对词根和词缀进行分离。例如“正射投影仪”切分为“正射/投影//仪”，其中“//”即可看作是 2 次切分，也可表明“投影”和“仪”之间关系相较“正射”和“投影”直接关系要更为紧密。

3.2.2 单音词构词问题单音词的构词研究相对较少，这里采用了《汉语构词法和造词研究》[10] 中的单音词构词分析观点，从语义的角度将单音词分为基本范畴词和下位范畴词，如“羊”和“羯”。

3.2.3　几种构词理论的融合问题句法、韵律、语义几种构词研究各有其优势的同时也各有其不适于科技词汇构词分析，或不适于科技分词目标的地方。句法分析，理论体系明确且稳定，但应用于科技词汇分析时，其中连绵词、叠音词在科技词汇中极少见，同时由于科技词汇往往结构复杂，则很难用句法构词理论分析短语型科技词汇的结构。因此主要利用了其中句法角色的判定方法，用于分析具有述谓结构的词汇中不同组分的句法角色。韵律分析，相关研究不容易抽象成标注标签，主要在切分词内结构时作为依据。语义分析，科技词汇与概念之间的关系密切，语义分析方法是理清科技词汇结构，还原其构造过程的有效方法，理论体系非常复杂，需要通过简化抽取才能形成可标注的标注体系。

3.3　标注体系及其界定

本研究融合了句法构词理论和语义构词理论，提出了一套简单易行的科技词汇

3.3.1　单音词构词界定单音词分为基本范畴词、下位范畴词和词缀三种类型，其余不好判定其分类的归入其他类别。分类约定如下：表示物种名称的单音词中，可以作为“目”、“科”、“属”、“种”名的单音词认定为基本范畴词（如狮，狮种，鼠，鼠科），与“目”、“科”、“属”、“种”同义且一定语境下可以互换的单音词认定为基本范畴词（如艾，艾草），其余物种名称认定为一般范畴词。生命体的器官、组织、结构名称属于下位范畴词。化学物质名词的单音词中，元素名称认定为下位范畴词，同位素名称认定为下位范畴词（如，氕氘氚），其余化合物、有机物名称认定为上位范畴词（如酸、羟、烃、笨、酚等）。若单音词对应的英文翻译为词缀形式，如“pros-”、“milli-”，则标注为词缀。虽然这部分词在汉语中语义可能并未完全虚化，但由于受到翻译的影响，这部分词在运用时很可能充当词缀的角色，因此需统一标注为词缀，也方便日后再做细分。剩余词汇的范畴判定往往需要深厚的学科知

表1　构词标注体系

种类			举例
单音词	基本范畴		羊、马
	下位范畴		羯、骥
	词缀		半 侧
	其他		胆、尾、孵
多音词	语义核心词	名物义	正态概率分布函数
		动作义	单点经纬仪观测
		性质义	单雌群的
	构词特性	并立	氢氧化钠、环绕
		限定	代数微分方程
		述谓	峰值功率测量

识，会给标注带来一定困难。因此需要将这部分单音词先归入其他类别，若标注过程中发现其他明确的范畴关系可以补充到以上约定中，再对这类词进行重新分类。

3.3.2 多音词构词界定多音词的标注通过判定语义核心词和判定构词特性实现，这两个判定间是相互联系的而不是对立关系。汉语科技词汇中表现为“限定成分 + 语义核心词”的限定式构词较多，因此可以对是否属于这一类别进行判定，汉语具有将语义核心词放于右侧的习惯，可以观察右侧第一个切分单元是否是整个词的语义核心，其他切分单元是否起到修饰、限定和区别的作用。而后判定语义核心词，属于名物义、动作义、还是性质义。这里判断的是词义而不是词性，名物义相对比较容易判定，如“仪器”、“设备”、“装置”等。判断动作义是需要注意，如“超声波测量”不许要考虑测量在这一语境中是动词还是名词，从语义角度上看“测量”是一种行为，标注为动作义即可。其余表明性状、颜色、性质的归属于性质义，如“A 型”、“靛蓝色”，以及以“的”结尾的形容词。限定式需录入语义核心词，录入时优选双音词或三音词，如不符语义结构再选择词缀。并立式，前后成分词性相同，且不具备修饰关系的一类标注为并立式关系，如“输入输出”，前后成分之间没有修饰限定关系，并且词性相同，地位相同。并立式可以两个成分并立也可以是多个成分并立，比如“碳酸氢钠”是三个名词成分的并立。并立式中每一个并立的成分都是核心词，由于其性质相同，对其中一个进行判断即可。词中包含具有并列以为的连接词的，如连接词两侧成分符合并立式判定标准，也归入并列式，如化学介词中的“化”、“合”等。并列式不录入语义核心词。述谓式，述谓式即具备述谓结构的一类词，等同于句法构词分析中的主谓、述宾、补充式，在李仕春的《汉语构词法和造词法研究》[10] 中，这一类词被称为互补式，并入到并立式当中。但由于在进行完构词判定后还须录入述谓结构，因此将这一个分类独立了出来。述谓结构不录入核心词。述谓结构的录入。由于科技词汇很多是从具有完整句法结构片段转变而来，通过成分缺省（如分析生长数据——生长分析）、顺序调换（如宾语前置，分析生长数据——生长数据分析）完成了词汇化的过程，但这种倒装或是缺省具有一定的任意性，比如导致倒装前后的词汇都存在，有无成分有无缺省情况的都存在，分析述谓结构有助于增强分词系统对这一类变化的识别程度。另一方面在相同的述谓结构框架下，某一成分是可以在同类词之间进行替换的，比如生长分析和发育分析。若可以对分词系统进行述谓框架和同义词的训练，则可以一定程度上提高分词器对于具有述谓结构的未登录词的切分准确度。这里的述谓结构并不是指语法层面的动词，而是指语义层面的谓词。述谓结构分析并不限于述谓式，限定式当中包含谓词的也需要进行录入。对于谓词成分缺省的一类词可以依据尝试进行补充，但不作为硬性要求，补充式需要加注圆括号。

3.4 其他标注内容

词性序列标注，在现有基于序列标注分词方法序列标注标签的基础上增加了两种标签。由于本研究需要对科技词汇进行词内切分，并对其词内边界进行标注，故设定将一级切分单位结尾标为 E’，二级切分单位结尾标为 E’’。增加后的序列标注标签见表 2：

词中表2 序列标注体系

标注	含义
S	单字成词
B	词首
E’	非切分单位结尾
E’’	非切分单位结尾
E	词尾
M	

4 构建系统设计方案

4.1 标注流程规划

标注流程见图 3，其中菱形表示判定环节：

4.2 标注系统设计

对目标词表进行标注，需要有表层进行标注

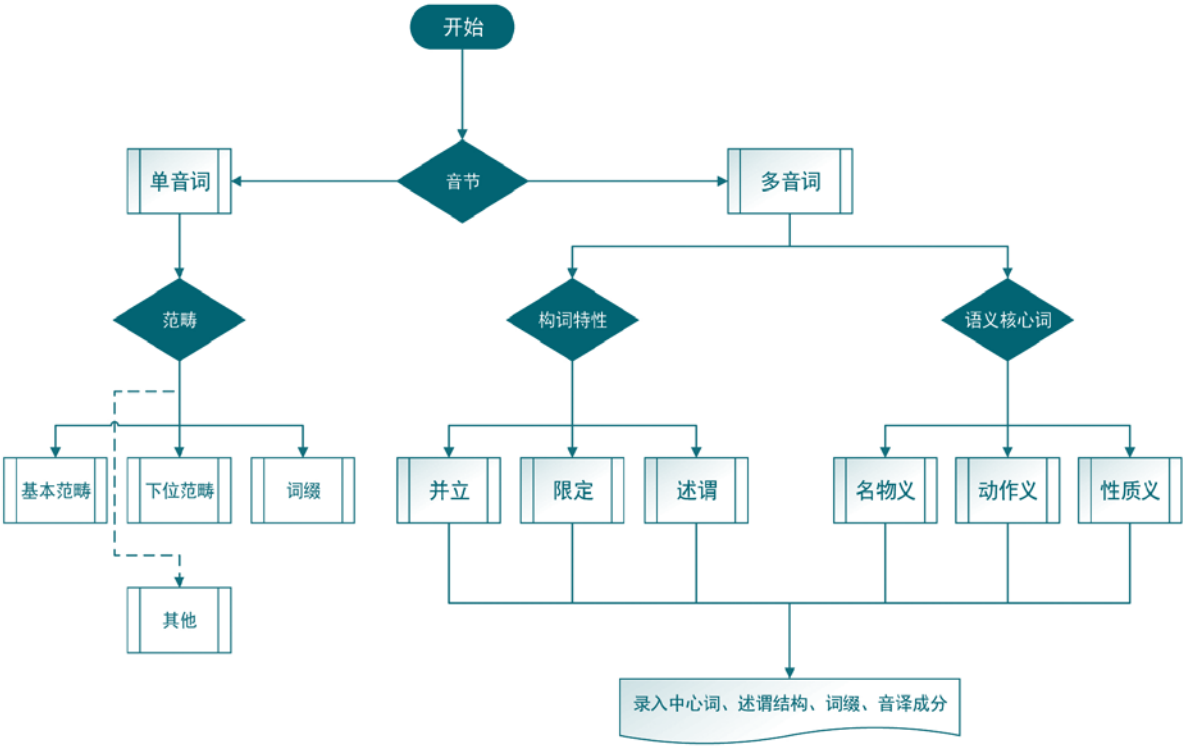

图 3 标注流程图

因此标注系统首先是要分为两层的，即标注层和数据层。4.2.1 数据层首先要对数据进行清理。将其中的符号进行半角和全角的统一。中文词重复部分将其英文名称、学科进行合并，以完成数据排重。数据层总体分为三个表，包括原始表，分割表和标注表。原始表是在原有数据清理结果基上增加一个字符数的字段，样例如表 3：分割表，是对标注层对于科技词汇的切分进行记录，并自动生成带有序列标注符号的结果。样例如表 4：标注表，根据标注层的选择和录入情况将信息录入到标注表中，样例如表 5 和表 6，由于字段较多，分为两个表展示，在数据层当中应为一个

序号	词	英文	学科	字符数
	环炔烃基	cycloalkynyl group	有化	
	环染纤维	ring dyed fiber	纺	
	环绕的	ambient	科技	
	环绕复形	mixed complex	统计	
	环绕结扎术	circumferential wiring	口腔	
	环绕颧骨结扎术	circumzygomatic wiring	口腔	
	环绕颧骨悬吊上颌术	circumzygomatic wire suspension of maxilla	口腔	
	微波鉴相器	microwave phase discriminator	电子	
	微波接力通信电路	microwave relay link	通信	
	微波接力通信链路	microwave relay link	通信	
	微波接力通信系统	microwave relay communication system	通信	
	微波接收机	microwave receiver	电讯	
	微波晶体管	microwave transistor	电子	
	微波晶体管放大器	microwave transistor amplifier	电子	
	微处理机输入输出	microprocessor I/O	计	
	微处理机数据记录	microprocessor data logging	计	

序号	词	分割	序列标注
	环炔烃基	环炔烃 / 基	环B炔M烃E'/基E
	环染纤维	环染 / 纤维	环B染E'/纤M维E
	环绕的	环绕 / 的	环B绕E'/的E
	环绕复形	环绕 / 复形	环B绕E'/复M形E
	环绕结扎术	环绕 / 结扎 // 术	环B绕E'/结M扎E''//术E
	环绕颧骨结扎术	环绕 / 颧骨 / 结扎 // 术	

环B绕E'/颧M骨E'/结M扎E''//术E
环绕颧骨悬吊上颌术 环绕/颧骨/悬吊/上颌/术 环B绕E'/颧M骨E'/悬M吊E'/上M颌E'/术E
微波鉴相器 微波/鉴相//器 微B波E'/鉴M相E''//器E
微波接力通信电路 微波/接力/通信/电路 微B波E'/接M力E'/通M信E'/电M路E
微波接力通信链路 微波/接力/通信/链路 微B波E'/接M力E'/通M信E'/链M路E
微波接力通信系统 微波/接力/通信/系统 微B波E'/接M力E'/通M信E'/系M统E
微波接收机 微波/接收//机 微B波E'/接M收E''//机E
微波晶体管 微波/晶体管 微B波E'/晶M体M管E
微波晶体管放大器 微波/晶体管/放大//器 微B波E'/晶M体M管E'/放M大E''//器E
微处理机输入输出 微//处理//机/输入/输出 微BE''//处M理E''//机E'/输M入E'/输M出E
微处理机数据记录 微//处理//机/数据/记录 微BE''//处M理E''//机E/数M据E'/记M录E

表4 数据层分割表

表3 数据层原始表

表 5　数据层标注表

序号	词	基本范畴	下位范畴	名物	动作		性状	并立	限定	中心词
176238	环炔烃基				YES				YES	基
176239	环染纤维				YES				YES	纤维
176240	环绕的						YES		YES	绕
176242	环绕复形				YES				YES	复形
176243	环绕结扎术				YES				YES	结扎术
176244	环绕颧骨结扎术				YES				YES	结扎术
176245	环绕颧骨悬吊上颌术				YES				YES	术
383748	微波鉴相器				YES				YES	鉴相器
383749	微波接力通信电路				YES				YES	电路
383750	微波接力通信链路				YES				YES	链路
383751	微波接力通信系统				YES				YES	系统
383752	微波接收机				YES				YES	接收机
383753	微波晶体管				YES				YES	晶体管
383754	微波晶体管放大器				YES				YES	放大器
384071	微处理机输入输出					YES				输入输出
384072	微处理机数据记录									

表 6　数据层标注表

序号	词	述谓	主语	谓语	宾语	by	词缀	音译成分
	环炔烃基							
	环染纤维			环染	纤维			
	环绕的						的	
	环绕复形							
				结扎		环绕	术	
	环绕颧骨结扎术			结扎		环绕颧骨	术	
	环绕颧骨悬吊上颌术			悬吊	上颌	环绕颧骨	术	
	微波鉴相器						器	
	微波接力通信电							
	路微波接力通信							
	链路微波接力通							
	信系统微波接收							
				接收	微波		机	
	微波晶体管微							
				放大		晶体管	器	
	微处理机输入输出	YES		输入输出				
	微处理机数据记录	YES		记录	数据			

4.2.2 标注层标注层的目的在于逐个提取待标注词汇，嵌入简单的计算程序，对规则化的标注体系进行提示，方便快速选择，减少手工录入。目前计划用html做标注页面，标注页面设计如图 4：其中蓝色填充部分为按钮，可以直接点击选择，蓝色边框部分为文本框，需要录入。上方“环绕式存储器”部分文本框默认提示待分割文本，支持输入 /。标注结果自动保存到后台数据表当中，其中序列的标注是通过程序自动添加的，故标注层不需要进行这部分操作。

5 结论

通过对传统语言学构词理论的总结和科技词汇自身特征的分析，本研究提出了一套简单易行的科技词汇构词特征标注体系，并对科技词汇结构的界定进行了描述。后期研究将运用标注系统对万方科技词表进行构词标注，并不断修正标注体系。对标注结果进行不同维度的定量分析总结归纳科技词汇结构特征分布情况，并将标注结果及分析结果应用于分词、词性一体化标注系统，对科技文献进行分词，以验证此种方法的有效性。

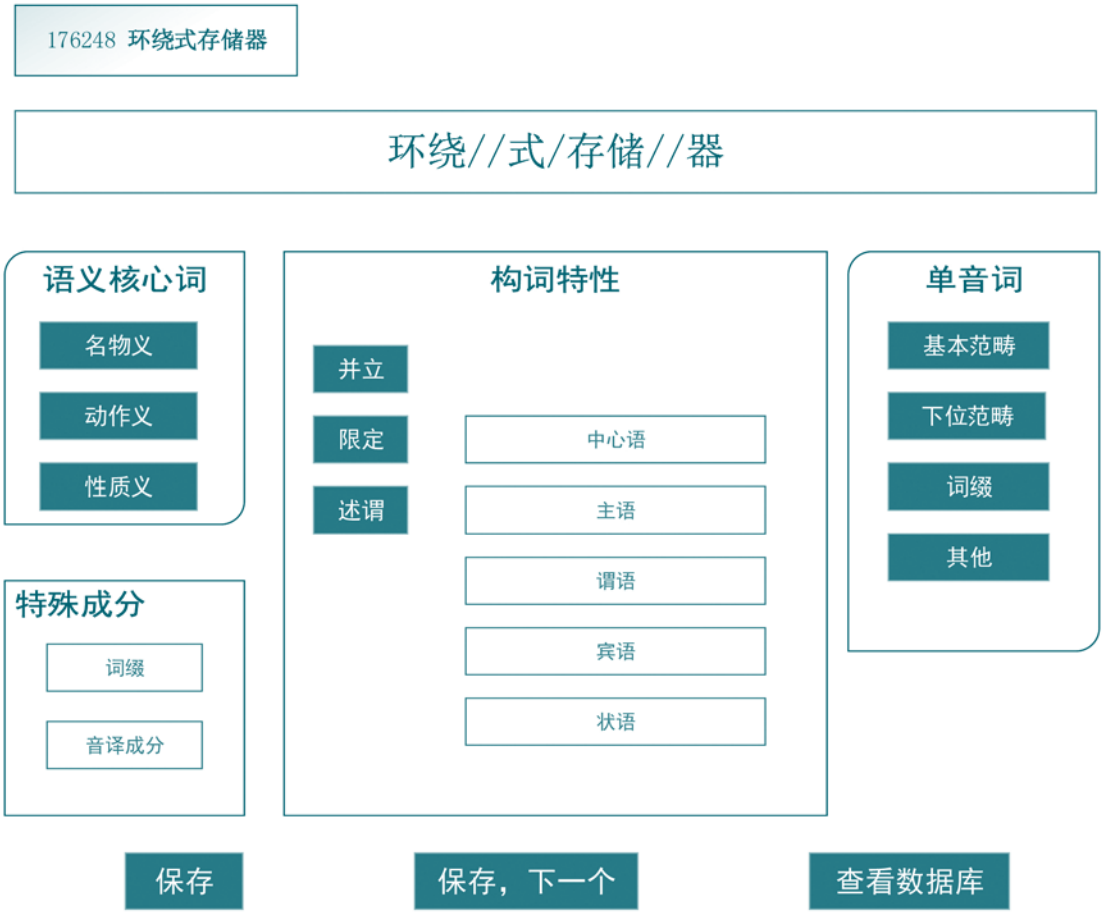

图 4 标注系统表层设计图

参考文献

[1] ZHANG Meishan, ZHANG Yue, CHE Wanxiang, et al. Chinese Parsing Exploiting Characters [C] // 51st Annual Meeting of the Association for Computational Linguistics,

[2] 赵亮 . 基于构词法的中文自动分词方法研究 [D]. 北京 :

[3] 齐皓爽 . 化工专业词典结构设计及中文分词系统的开发

[4] 周其焕 . 航空术语的构词分析 [J]. 中国民航大学学报

[5] 徐硕 , 乔晓东 , 朱礼军 , 等 . 广义后缀树及其在汉语科技词系统中的应用研究 [J]. 数字图书馆论坛

[6] 张科 . 汉语构词法研究综述 [J]. 语文学刊

[7] 鲁小娟 . 汉语构词法研究综述 [J]. 社会科学论坛

[8] 黄伯荣 , 廖旭东 . 现代汉语 [M]. 第四版 . 北京 : 高等教

[9] 冯胜利 . 汉语的韵律、词法与句法 [M]. 北京 : 北京大

[10] 李仕春 . 汉语构词法和造词法研究 [M]. 北京 : 语文出版

[11] 朱彦 . 汉语复合词语义构词法研究 [D]. 上海 : 华东师

[12] 冯志伟 . 现代术语学引论 [M]. 北京 : 商务印书馆 ,2011.

[13] 冯志伟 . 术语形成的经济率——FEL 公式 [J]. 中国科技术