领域本体学习语料的自动获取与预处理方法研究*

王思丽 [1,2,3]　祝忠明 [1,2,3]　刘巍 [1,2]　杨恒 [1,2]

[1]（中国科学院西北生态环境资源研究院 文献情报中心，兰州 730000）

[2]（中国科学院兰州文献情报中心，兰州 730000）

[3]（中国科学院大学，北京 100049）

Research on Automatic Acquisition and Preprocessing Methods of Domain Ontology Learning Corpus

Wang Sili[1,2,3]　Zhu Zhongming[1,2,3]　Liu Wei[1,2]　Yang Heng[1,2]

[1](Literature and Information Center of Northwest Institute of Eco-Environment and Resources, CAS, Lan Zhou 730000, China)

[2](Lanzhou literature and information Center of Chinese Academy of Sciences, Lan Zhou 730000, China)

[3](University of Chinese Academy of Sciences, Bei Jing 100049, China)

摘要：[目的/意义]实现领域语料的自动获取与预处理，为机器/深度学习驱动的领域本体自动构建提供数据及数据处理技术基础。**[方法/过程]**首先，对所涉及语料的类型、获取方法及应用研究现状进行分析，提出多源异构领域语料的自动获取方法，包括基于 Web Spider 的网络开放领域语料和基于 Web API 的科学文献领域语料的自动获取等。其次，分析提出领域基础知识词典的自动构建方法，为语料预处理奠定基础。最后，通过对主流分词方法及开源分词工具进行测试与评估，提出基于增量训练 HanLP-SP 领域分词模型的多策略混合的自动分词与新词发现方法，并进行实验研究。**[结果/结论]**方法能够有效获取到领域语料，并实现分词等预处理任务。

关键词：领域语料；本体学习；自动获取；预处理；分词

Abstract： [Purpose/Significance]Realize the automatic acquisition and preprocessing of domain corpus, and provide data and data processing technology basis for machine learning or depth learning driven domain ontology automatic construction. **[Method/Process]**Firstly, the types of corpora, acquisition methods and application research status are analyzed. The automatic acquisition methods of multi-source heterogeneous domain corpus are proposed, including Web Spider-based network open domain corpus automatic acquisition and Web API-based scientific literature domain corpus automatic acquisition, etc. Secondly, an automatic construction method of domain basic knowledge dictionary is proposed, which lays a foundation

*本文系中国科学院兰州文献情报中心 2018 年主任基金项目“基于深度学习的领域本体自动构建方法研究”（项目编号：Y8AJ012005）和中国科学院 2019 年西部之光项目“开放学术资源的情景化组织与服务研究” (项目编号：Y9AX011001)的研究成果之一。

for preprocessing corpus. Finally, through the test and evaluation of the mainstream word segmentation method and the open source word segmentation tool, a multi-strategy hybrid automatic word segmentation and new word discovery method based on the incremental training HanLP-SP domain segmentation model is proposed and experimental research is carried out. **[Result/Conclusion]**The method can effectively acquire the domain corpus and realize the preprocessing tasks such as word segmentation.

Keywords：Domain Corpus; Ontology Learning; Automatic Acquisition; Preprocess; Word Segmentation

1 研究背景、意义与主要研究目标

无论是传统的机器学习还是当前的深度学习，无论是无监督学习还是有监督学习或增强学习，所驱动的任务都是从给定数据（集）开始，挖掘到隐藏特征，预测未知模式（分类/理解/最大化收益等）结束。可见，数据（集）是机器自动化智能化学习任务中相当重要的一环，其质量好坏也将决定着学习结果的好坏。

领域本体学习语料，即领域本体的数据来源，也是进行领域本体构建时必须系统考虑和预先准备好的基础知识资源。随着领域本体构建方法的不断演化发展，领域本体的数据来源及获取方式也在不断发生变化。在过去以领域专家参与为主的手工构建方法体系中，其数据来源一般是专家的“先验知识”和“主观认知”，获取方法也主要依赖专家或本体建设人员的人工判别和整理输出，时间和经济成本过高，严重限制了本体的发展。当智能化技术驱动的以自然语言处理和机器学习为主的自动化构建方法流行起来后，所需支撑的数据来源、类型、格式变得更加复杂，数据量也在不断增大，相应的也为数据精准获取与处理增加了难度，再完全依靠人工已不现实，因而也必须考虑其自动获取与预处理的方法。

调研发现，目前能够支持领域本体自动化学习构建的语料数据按其结构化程度大致可分为三种：结构化的数据[1,2,3]，如机器可读的词典、叙词/主题词表、分类法、存储在关系数据库的数据、已经过语义标注处理的各种模板化、规则化语料等；半结构化的数据[4,5,6]，如 RDF、XML、DTD、嵌入了 RDFa 标记的 HTML 网页等；非结构化的数据[7,8,9]，主要是指以 TXT 格式存储的纯文本。然而，我们经常能接触到的第一手数据大多都为结构化和半结构化的数据，而深度学习方法所需要的数据格式一般为经过预处理的 TXT 纯文本格式，若是其他两种格式，最终也必需转化为纯文本格式备用。

因而，本文的主要研究目标是研究领域本体学习语料自动获取与预处理的核心方法和技术，主要包括多源异构领域语料的自动获取方法、领域基础知识词典的自动构建方法、领域文本的自动分词与新词发现方法等，为机器/深度学习驱动的领域本体自动化构建研究提供数据及数据处理技术基础。

2 领域本体学习语料的自动获取方法研究

2.1 语料分类、获取方法与应用研究现状

语料一般是指借助文献调研、文献计量、统计语言等手段科学采样，经过人工或计算机分析与加工处理而形成的大规模数字化文本库。语料是深度学习、机器学习、自然语言处理等研究中都十分依赖的不可或缺的基础资源，相关模型的训练常需要大量的语料作为输入。根据研究目的和应用场景不同，语料一般可分为以下四种：

（1）通用语料，又常被称为异质语料，一般指没有特定收集原则和目的，没有领域和主题限制，而广泛收集存储的跨学科公共基础知识语料，如各种通用本体库、WordNet、HowNet、Wikipedia、百度百科语料、人民日报标注语料库 PFR 等，可用于各种通用领域文本分析挖掘、分类聚类、搜索引擎技术研究等。

（2）领域语料，又常被称为同质语料，一般指具有领域、主题或内容结构限制，只收集存储相同领域主题或内容结构类型的语料，如各种专业词典、领域主题词表、领域叙词表等，可用于各种专业领域文本采集抽取、分析挖掘、分类聚类、语义标注研究等。

（3）系统语料，是指根据一个预先确立的原则和比例进行语料收集，使其具有一定的系统性，平衡或不平衡性，能够在一定程度上代表某一范围内的事实特征。系统语料可直接通过科学手段原始采集获得，也可以在通用语料或领域语料基础上经过分层抽样、主题抽取等二次处理而获得。如深度学习和机器学习常需要预先准备好的阳性语料（正例样本集）与阴性语料（负例样本集），测试语料（测试集）与训练语料（训练集）等，并需要满足一定的比例。

（4）专用语料，顾名思义，指专门为了某一特定用途而收集组织的语料或数据集等。如哈尔滨工业大学设计完成的具有 5 层编码结构表示的同义词林扩展版语料等，专用于对通用领域的公共基础知识词汇进行同义语义扩展研究。如 2011 年，Baroni 等创建的 BLESS 数据集[10]，主要包含一些上位关系、共同下位

关系（同义关系）、部分-整体关系等一些语义关系词组，专用于对分布式语义学研究的语义相似度进行评估研究等。

此外，语料还可按照语种分为单语语料，双语语料（如中英文双语词典等），多语语料等；按照组织形式可分为平行语料库，又称对齐语料库，主要用于双语词典编纂、机器翻译研究等；比较语料库，主要用于多语言特征对比研究等。如 Lison 等从电影和电视字幕中翻译提取形成的大型多语言平行语料库 OpenSubtitles[11]，目前包含大约 62 种语言，1782 种双语文本，373 万多个文件，22G 的标记和 3G 的句子片段等。如由英国多位语言学家、计算机科学家、软件工程专家、自然处理研究人员等组成的工作组公开发布了 Sketch Engine 工具[12]，该工具包含 500 多个随时可用的基于 90 多种语言的语料库，每个语料库大小可达 300 亿个单词等，还提供了词汇编纂、词汇计算、术语提取等支持文本分析挖掘的应用程序和方法等。

2.2 多源异构领域语料自动获取方法研究

2.2.1 基于 Web Spider 的半结构化网络开放领域语料自动获取方法

半结构化网络开放领域语料一般是指以 HTML 格式公开发布在领域专业/主题网站、权威机构/新闻网站中的领域科技新闻、资讯、政策、观点等信息。这类语料一般具有词汇丰富、新颖、时效性强、覆盖范围广、更新速度快等特点，是领域热词、新词的发祥地，可及时为领域本体概念扩充提供新的词汇来源和帮助探索发现新的领域语义关系。本文通过编写和设计 Web Spider，即网络爬虫来实现对该类领域语料的自动获取，核心流程框架如图 1 所示，具体方法如下：

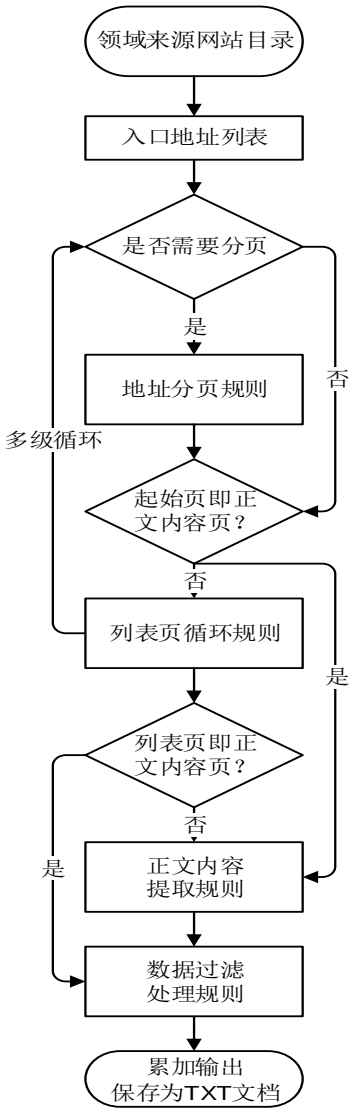

图 1　基于 Web Spider 的网络开放领域语料自动获取框架

（1）调研与梳理待爬取领域来源网站目录，形成一个入口 URL 地址列表。

（2）设计采集规则流程 1：入口地址分页规则。该规则主要采用基于 URL 相似性规律的规则表示方法[13]。前期研究发现，分页地址格式通常由固定前缀部分和可变参数部分组成，其中可变参数部分又常遵循 4 类规则：等差数列规则、等比数列规则、A-Z 或 a-z 字母变化规则、时间及分隔符组合格式规则。针对不同源应用该规则生成可变的入口地址分页规则的逻辑表达式进行分页请求。

（3）设计采集规则流程 2：列表页多级循环规则。该规则主要采用基于迭代循环和逐层链接访问的方法[13]。其核心思想是将入口 URL 地址包括其分页地址视为 0 级网址，研究发现 0 级网址内应该至少包含一个列表页 1，而 1 级列表页内的每一个链接指向的页面内可能同样包含了列表页 2，依次迭代，最终列表

页内的链接必定将指向一个正文内容页。这个迭代循环的次数，就是实际采集的深度，理论上可以无限极迭代。但常见领域新闻资讯类一般都为 2 级采集，即 1 级列表页内包含的链接可以直接指向正文内容页，我们需要的有价值的领域语料一般都包含在正文内容页中。

（4）设计采集规则流程 3：正文内容提取规则。该规则主要采用基于正则表达式的内容标签模板匹配方法[13]和基于启发式规则的文本密度判定方法。前者主要适用于来源领域网站正文内容页结构相对简单固定，标签结构等级鲜明，待采集语料内容一般包含在一个特定的标签对组件之间的情况，如 * 或<article>*</article>此类等。可将标签对作为模板，待提取的信息作为参数，构造正则表达式进行参数搜索与模式匹配提取。后者主要适用于来源领域网站正文内容页结构灵活多变，待采集语料内容并非只固定包含在一个可以事先准确预知的位置的情况。其核心思想是研究发现有效文本密度一般应高出噪音文本部分很多，其所在位置应是节点连续的文字稠密的，基于这一文本特征，采用启发式规则算法去统计分析整个页面的所有节点及文字并计算其密度，自动提取并组合其中包含最多文字的连续节点作为采集结果内容。两者相比，前者采集定位及采集结果可能更准确，但需要介入较多人工预先去获悉每一个来源网站的正文内容页源码并根据具体内容标签组件构造正则表达式；后者采集定位是来源、结构、标签无关的，也不需要人工干预，可全自动执行提取，但可能会在某些特殊场景下存在采集结果不准确的问题，如恰好噪音文本也包含在一个具有较多文字的连续节点内，则会误将噪音文本也归入采集结果中。具体应用时，可根据实际情况，将两者分别使用或结合起来配置使用。

（5）设计采集规则流程 4：数据过滤处理规则。由于最终需要的语料是纯文本的不含任何格式的，因此需要对初始采集到的内容进行过滤处理。首先是对内容中的样式标签进行全过滤处理，包括 form 表单、frame 框架、style 样式、script 脚本等各种 HTML 标签，各种特殊字符及其转义字符等。其次还需要对内容进行编码统一转换，否则会因编码不一致导致各种乱码问题而影响下一步读取使用。常见的网页内容编码类型有 GB2312、Unicode、URL 转义编码、UTF8 编码等，本文将其统一转换为 UTF8 编码格式输出。

（6）将得到的无格式的内容文本累加输出并保存在同一个 TXT 文档中。

此外，还可以利用该方法，将网络开放搜索引擎如百度、必应等作为入口地

址，根据事先遴选的一些领域关键词进行垂直搜索爬取，以扩大领域语料来源。

2.2.2 基于 Web API 的半结构化科学文献领域语料自动获取方法

半结构化科学文献领域语料一般是指存储在大型科学文献数据库、机构知识库等中的论文、专利等的题名、摘要、关键词、主题词等元数据信息。这类语料的词汇受益于论文、专利等文献出版过程中的同行评议和编辑审核机制，与以公众话题为主的网络开放领域语料相比，词汇格式更为标准规范，专业性和学术性更强，也是领域本体概念绝不可忽视的重要来源，并可为规范化领域本体概念提供专业知识依据。这类数据库一般会免费开放或向商业付费用户提供 Web API，允许用户按一定标准协议免费或基于口令授权后使用机器程序自动调用其 API 批量获取文献元数据信息。常见的 Web API 有 OAI-PMH、Web Services、Restful API 及一些自定义 API 等，采集到的原始数据流一般为 XML、JSON 等格式，因此还需要编写数据解析程序，使用如 Dom4j（XML API）、Fastjson（JSON API）等之类的组件库，根据相关节点元素含义从上述数据流中依次解析出所需要的元数据信息。具体流程如图 2 所示：

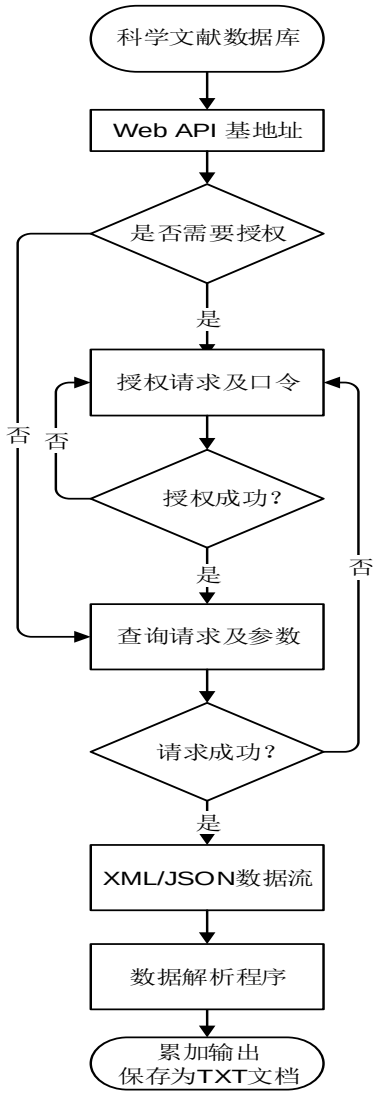

图 2　基于 Web API 的科学文献领域语料自动获取框架

2.2.3 其他所需专用语料及获取方法

除了上述必需的领域语料之外，在研究中可能还需要一些其他语料作参考、对比、测试、评估和数据处理等，这类语料一般是相关同行研究人员或研究机构在以往研究试验中打造或积累的词汇表、样本集、训练集、测试集、评估集、基准集等，很多可能是已经过大量人工或机器特定处理的已标注的或具有特定类型/关系/属性的数据集，因而具有较高的参考应用价值。如 Wikipedia、WordNet、人民日报标注语料 FPR、同义词林扩展、语义关系库 BLESS、微软概念图谱为短文本理解构造的 Probase[14]等。但这类语料一般没有固定的获取位置，可根据相关研究论文中标注的数据来源链接地址去访问下载或根据搜索引擎搜索后下载。此外，还有一些综合开放的在线语料库集成系统，国内的如北京大学中国语

言学研究中心的语料库 CCL PKU、中国科学院自动化研究所的中文语言资源联盟 CLDC、北京语言大学语料库中心 BCC、国家语委教育部语言文字应用研究所的语料库在线 CNCORPUS 等，国外的如英国国家语料库 BNC、美国国家语料库 ANC、英国牛津大学的兰开斯特语料库 LCMC、欧盟委员会第七框架计划资助的 EuroMatrix 项目产出的欧盟议会诉讼平行语料库 EPPPC 等。这些语料库均提供相应语料的在线浏览、检索、下载等功能。

3 领域基础知识词典的自动构建方法研究

如果说领域语料是机器/深度学习研究不可或缺的基础资源，那么领域基础知识词典就是基础中的基础。首先是因为直接从在线开放网络或文献数据库中采集获取到的纯文本领域语料并不能直接用于深度学习模型的训练和构建，还必须经过一系列分词处理，而领域基础知识词典中常包含大量已得到认可的具有一定规范性的领域种子概念和基础词汇，可为领域语料的分词处理提供重要的参考依据和词典工具，促使更多的领域词汇可以被规范化处理和被正确识别出来。其次，目前即便是最先进的分词工具和方法在对领域语料进行分词时，由于领域术语词汇的专业性与独特性，也可能会出现误拆与错分，领域基础知识词典的加入将有利于提高分词工具的对领域语料分词处理的准确率。此外，在领域文本特征词嵌入模型的自动训练与生成阶段，还可以考虑将领域基础知识词典与经过分词的初始领域语料一起构成新的领域语料，作为模型训练的输入语料，以扩充模型的训练样本集和提高模型训练的精度。

由于领域基础知识词典中的词汇应是标准的规范的，且应尽量能够覆盖和表达所研究领域的重要概念和术语，以往的领域基础知识词典主要依靠大量领域专家或相关人员费时费力手工整理创建或长久积累而成，因而其规模、结构、覆盖范围、质量也参差不齐，且并非所有的研究领域都已具备领域基础词典。后来，随着领域主题词表、叙词表、本体的研究众多，它们中的概念术语和语义关系也已成为构建领域基础知识词典的重要来源。目前，由中国科学技术信息研究所从 1980（开始和探索阶段）-2009（正式立项）-2017（完成）耗费 30 余年编制的，被誉为中国第一部大型综合性中英双语叙词表的《汉语主题词表》[15]，已囊括了自然科学、工程技术、生命科学和社会科学领域的文献关键词、专业术语、百科全书、叙词表等丰富的词汇资源约 500 万余条，同时又按照不同具体学科如数学、

物理学、化学、天文学、海洋学等对词汇进行了特征详细分类与组织，使得即可以按词汇语义关系分类获取数据又可以按文献学科分类获取数据，并且还提供了基于 Web 的概念智能检索、自动主题标引、自动关系标注、自动语义关联等在线服务。本研究拟以《汉语主题词表》及其提供的分类数据获取服务为依托，并结合所研究领域的科学文献作者关键词等，快速自动构建领域基础知识词典，以解决领域基础词典缺乏问题和为后续研究奠定基础。主要构建方法如下：

（1）对所研究的领域进行文献调研或专家咨询，遴选出一些核心领域主题词或关键词，构成专业检索式，基于 Web API 从科学文献数据库中自动批量获取文献元数据信息，具体方法流程同上文 2.2.2。

（2）对获取到文献元数据信息，依次提取作者中文关键词和英文关键词字段，一篇文献的多个关键词也按标记拆开存储，并计算每一个关键词出现的频率，将结果去重后首先存储在关系数据库表 Dic 中，行结构格式为<ID，中文关键词名称，英文关键词名称，词频>。同时，对 Dic 进行统计分析，将可能存在的一些过于通用的没有实际领域意义的或解析错误导致乱码的关键词进行预先过滤删除，如“研究”、“方法”、“进展”；数值型词汇如“2018 年”、“四十年”等。

（3）考虑将 Dic 中的初始领域文献关键词与从《汉语主题词表》中对应领域的词汇分类与文献学科分类中获取到的词汇数据进行匹配和余弦相似性计算，将能够完全匹配或余弦相似性比较高且接近某个阈值（如 0.95 及以上）的词汇按词表中的词汇规范及分类关系进行自动机器标注与关系标引；对无法匹配到或相似性比较低的词视为新词，暂时不做其他处理，作为原子词存在。

（4）基于标引结果对 Dic 进行扩充，行结构格式为<ID，规范中文关键词名称，规范英文关键词名称，词频，优选词项 Y，代项 D，上位词项 S，下位词项 F，相关词项 C ，族首词项 Z>，其中新增的标记字段都是来自《汉语主题词表》的标引规范。如优选词项 Y 表示的是该学科领域文献中比较通用的权威的新颖的词，通常是指全称、新称、学名、泛指词等规范名称，而代项 D 与优选词项 Y 相对，指的是简称、旧称、音译名称、俗称、专指词等不太规范的名称变体。族首词项 Z 表示的是优选词项 Y 所属的更广泛的上位概念项，如“黄海海域”的族首词项 Z 一般为“海域”或“中国海”。

（5）其中（3）和（4）的语义扩充步骤若实际情况不允许也可以跳过。最终得到的 Dic 即为一个与所研究领域关联度较高的小型领域基础知识词典，可在

后续研究中根据需要进行格式输出与生成。

4 领域文本的自动分词与新词发现方法研究

4.1 常见分词方法及应用问题分析

直接采集获取到的纯文本语料并不能直接作为深度学习模型的输入，还必须经过一系列分词处理过程，如分词，词性标注，命名实体识别、新词发现等，分词的准确与否会直接影响下一步词嵌入模型的训练生成精度高低。分词是指基于一定的方法规则将原本连续的字序列再次拆分与组合形成新的词序列的过程。虽然英文文本也存在着一元词组、二元词组、三元词组等 N-Gram 词组切分问题，但由于英文单词之间已有空格作为天然分隔符，在常规的研究应用中，英文文本即使不分词也能够直接作为机器学习的输入语料使用，并取得不错的效果。而中文文本的词与词之间并没有显式的分隔符，且词组之间的划分边界模糊，还会因为专业领域差异和上下文语境不同而存在同一词组的不同划分方法，因而中文分词一直是研究的重点和难点。

目前常见的中文分词方法有基于词典（字符串）匹配的方法、基于语言学规则的方法、基于统计学习（传统机器学习）/深度学习的方法等[16]。基于词典匹配的方法又常被称为机械分词法，是指按照一定的特征扫描策略或标志切分策略，将语料切分成一定长度的字符串并与一个理论上可以无限大的机器词典中的词汇进行匹配识别的过程，常用的扫描匹配策略有正向最大匹配法（左→右）、逆向最大匹配法（右→左）、双向最大匹配法（先左→右，再右→左）、最少切分法等。该方法分词的精度低且需要大规模的机器词典作基础，通常只是作为一种辅助手段，实际应用时需要和其他方法结合起来使用以提高分词的精度。基于语言学规则的方法是指利用机器程序模拟人对语言的理解判断过程，期望通过对语法、句法和语义进行规则分析和消歧而实现分词的方法。该方法需要大量语言学知识作基础，且对单一语言的依赖性强，还不具有良好的通用性，因而目前多处于试验阶段并没有流行开来。基于统计学习/深度学习的方法是指将词频统计、共现概率计算等方法和机器学习的方法结合起来，训练机器学习模型以实现分词的过程。常用的方法有基于隐马尔可夫模型 HMM 的方法、基于条件随机场模型 CRF 的方法、基于神经网络语言模型 NNLM 的方法[17]、基于深度学习模型 BLSTM-CRF 的方法[18]等。目前基于统计学习/深度学习的方法是最常用的方法，

常和基于词典匹配的方法一起构成了主流方法以综合处理分词中可能遇到的各种问题。

4.2 开源分词工具测试与性能评估

目前国内外也已经有一些成型的开源中文分词组件工具，如 Stanford CoreNLP、NLPIR、THULAC、LTP、HanLP、ANSJ、Jieba 等，基本都支持中文分词、词性标注、命名实体识别、关键词/摘要提取、依存句法分析、文本聚类/分类等功能，支持自定义词典分词，有的还提供 Http(s) API，支持直接在线调用相关分词服务或自行训练相关分词模型。本文对上述分词工具的主要支持语言及分词方法/模型进行了调研梳理，如表 1 所示：

表 1　几种主流开源分词工具

工具名称	研发机构	支持语言	主要分词方法/模型
CoreNLP	美国斯坦福大学	Java Python	CRF 模型：构建特征模板，求最大联合概率分布
NLPIR	北京理工大学大数据搜索与挖掘实验室	Java C++ C#	HMM-Bigram：二元切分词图
THULAC	清华大学自然语言处理与社会人文计算实验室	Java Python C++	结构化感知机模型 SP：最大熵准则构建评分函数，双数组 Trie 树存储特征，Viterbi 算法求解。
LTP	哈尔滨工业大学社会计算与信息检索研究中心	Java Python C++ C#	结构化感知机模型 SP：最大熵准则构建评分函数，HashMap 存储特征，Viterbi 算法求解。 提供 Https API 在线训练分词模型
HanLP	上海林原信息科技有限公司	Java Python	HMM-Bigram：最短路径分词、N-最短路径分词 由字构词：结构化感知机 SP 分词、CRF 分词 极速词典分词：Aho Corasick 自动机，双数组 Trie 树多模式匹配
ANSJ	个人（孙健，ansjsun）	Java	HMM-Bigram：双数组 Trie 树 DAT 存储特征，邻接表实现分词 DAG（有向无环图）。 提供 Http API 在线分词服务
Jieba	个人（fxsjy）	Python Java	HMM-Unigram：双 dict 分别存储 Trie 树和词及词频，Dict 实现分词 DAG，动态规划 DP 求解。

同时，为了选出性能最佳的分词方法与工具，本文还对上述分词工具依次进行安装、测试和性能评估比较，主要采用 64 位 Windows10 操作系统、Intel(R) Core i7-7700 CPU @ 3.60 GHz 处理器，64 GB 内存作为安装测试环境，第二届国际中文分词比赛（SIGHAN Bakeoff 2005）提供的经典的开源的标准测评语料库 icwb2-data[19]为测试数据集，自动评分脚本 score 为测试标准和方法。具体测试和评分过程如下：

（1）首先对各分词工具进行下载、安装及所依赖的数据词典、分词模型、其他组件库、系统环境的准备与配置，主要使用当前最新开源的 Java 版本。

（2）编写 Java 程序，对各工具的核心分词方法模型进行分词训练测试。有的分词工具只有一种分词模式，有的还支持多种分词模式，如标准分词、索引分词（全模式）、词典分词（自带核心词典，用户自定义词典）、HMM 分词（最短路径分词，N-最短路径分词）、CRF 分词、SP 分词等，主要测试的是其支持自然语言处理、命名实体识别、未登录词（新词）识别等的 HMM、CRF、SP 分词方法。如 NLPIR 测试的是 ParagraphProcess 方法，THULAC 是分词模型 Model_1，LTP 是分词模型 CWS Model，HanLP 是 NLPTokenizer 方法，ANSJ 是 NlpAnalysis 方法，Jieba 是开启了 HMM 模型的精确模式等。

（3）分别使用 icwb2-data 中微软研究院和北京大学提供的简体中文 UTF-8 编码测试集 msr_test.utf8 和 pku_test.uf8 作为输入语料各进行了两轮测试。研究发现，可能受计算机自身的性能影响，同一轮测试中，即便使用相同语料、相同工具、相同方法，多次执行结果的时间还是有一定差异的。为了缩小这种差异，在同一轮中，本文对相同语料、相同工具、相同方法至少重复 3 次测试，将时间耗费最少的批次时间作为测试的最终分词速率，分词结果作为最终训练集。

（4）调用 icwb2-data 的评分脚本 score，将各工具得到的最终训练集，分别与 icwb2-data 中微软研究院提供的简体中文 UTF-8 编码黄金标准分词词典 msr_training_words.utf8 与黄金标准测试集 msr_test_gold.utf8，相应的北京大学的黄金标准 pku_training_words.utf8 和 pku_test_gold.utf8 一起作为参数输入，执行命令，获得最终评分结果。主要命令格式为：

./score 黄金标准分词词典 黄金标准测试集 工具训练集 > 评分结果.txt

（5）评分结果展示与对比分析。其中 R 表示召回率（Recall），P 表示准确率（Precision），F 表示 F-Measure，是 P 和 R 的加权调和平均，常用的是 1 阶加

权：$F_1 = \frac{2*PR}{P+R}$。F1 综合考虑了 P 和 R 的结果，因而 F1 值越高一般表明方法模型性能越好。T 表示分词速率（Time），单位已转换为秒（s）。OOVR 表示未登录词召回率（Out Of Vocabulary Recall），IVR 表示登录词召回率（In Vocabulary Recall）。OOVR 值越高，表示对新词的识别发现能力越强。在微软研究院 MSR 数据集（548 KB，184358 字符）上的测试结果，如表 2 所示。在北京大学 PKU 数据集（498 KB，172733 字符）上的测试结果，如表 3 所示。

表 2　几种分词工具在 MSR 数据集上的评分结果

Tools	R	P	F1	T	OOVR	IVR
CoreNLP 3.9.1	0.859	0.822	0.840	76.692	0.452	0.870
NLPIR 2016	0.914	0.868	0.890	2.568	0.407	0.927
Thulac4j 3.1.2	0.895	0.862	0.878	0.953	0.457	0.907
Ltp4j 3.4.0	0.899	0.868	0.883	722.19	0.466	0.911
HanLP 1.7.3	0.827	0.865	0.846	4.236	0.615	0.833
Ansj 5.1.6	0.819	0.871	0.844	5.77	0.582	0.825
Jieba 0.39	0.812	0.817	0.815	0.869	0.449	0.822

表 3　几种分词工具在 PKU 数据集上的评分结果

Tools	R	P	F1	T	OOVR	IVR
CoreNLP 3.9.1	0.894	0.901	0.897	75.203	0.778	0.901
NLPIR 2016	0.944	0.939	0.942	1.575	0.702	0.959
Thulac4j 3.1.2	0.938	0.951	0.944	0.914	0.775	0.948
Ltp4j 3.4.0	0.946	0.960	0.953	351.28	0.831	0.953
HanLP 1.7.3	0.811	0.892	0.849	4.027	0.623	0.822
Ansj 5.1.6	0.785	0.879	0.829	4.526	0.602	0.796
Jieba 0.39	0.787	0.853	0.818	0.864	0.583	0.799

从测试结果可以看出，在召回率、准确率、F 度量方面，NLPIR、LTP、THULAC、CoreNLP、HanLP 都比较高，其中最高的是 NLPIR，其次是 LTP 和 THULAC，再次是 CoreNLP 和 HanLP。在对登录词和未登录词的召回方面，最高的是 LTP，其次是 THULAC 和 CoreNLP，再次是 NLPIR 和 HanLP。各工具两次实验的评分排名有一定波动，是由于 MSR 和 PKU 提供的黄金标准分词词典本身也有一定差异，MSR 中大多是由命名实体构成的长单词，粒度比较粗，PKU 恰好相反，粒度比较细，包括甚至将人的姓和名都拆分开的短单词。因而，通过进

一步实验对比发现，NLPIR、LTP、CoreNLP、THULAC 的分词粒度比较细，HanLP 的分词粒度比较大，因而 HanLP 在对命名实体识别方面具有较大优势。如“张三正在学习自然语言处理技术”一句，NLPIR、LTP、CoreNLP、THULAC 会将专有名词“自然语言处理/nz”拆分为“自然/n”、“语言/n”、“处理/v”三个词，而 HanLP 则成功识别。总体来看，HMM 相关的分词模型本质上是基于 DAG 图的词生成模型，速度与精度平衡，整体性能更佳，但对于未登录词的识别能力较差，比较适合于一般快速分词和关键词提取任务。SP 或 CRF 相关的分词模型是基于特征的判别模型，通常需要由大型语料库训练出的模型作为支撑，更侧重于精度，对命名实体和未登录词的识别发现能力更好，但计算开销大，分词速度较慢，更适合于 NLP 预处理任务。

4.3 多策略混合的自动分词与新词发现方法

从上文研究实验已发现，不同的分词方法模型各有优劣，并没有各方面完胜的工具存在。因此，本文从分词的精度，对命名实体、新词的识别能力，核心分词算法模型的开源和可扩展程度综合考量，最终选择以 HanLP 工具为基础，尝试将一些分词方法模型的优点结合起来，提出了基于多策略混合的自动分词方法。

本方法主要是基于 HanLP 中提供的由字构词的结构化感知机(Structured Perceptron, SP)分词模型及标注框架[20]进行扩展实现。HanLP 将传统的以线性二分类标签形式为基础的感知机模型拓展到 BMES 多分类标签形式（其中：B 表示 Begin，词首；M 表示 Middle，词中；E 表示 End，词尾；S，表示 Single，单字词），基于层叠 HMM-Bigram 算法计算多标签之间的转移概率作为转移特征，并结合多种预定义状态特征进行字符特征提取，最后使用改进的 Viterbi 搜索算法进行分类求解以实现分词。目前，HanLP 提供的 SP 基础词法分析模型是训练自将近一亿字的大型综合性中文语料库，是当前已知范围内最大的中文分词模型之一，已具有较高的精度。同时，为了使基础模型能够具有自我革新能力以快速适应新的生产领域和新的术语变化，避免因语料狭小或陈旧而导致的识别遗漏或错误问题，HanLP 还提供了带有 Train 接口和 Learn 接口的 SP 序列化标注框架，支持用户按要求格式自定义领域语料，自行独立训练新分词模型（Train 接口）或在原有模型的基础上实现在线学习新知识和增量训练模型（Learn 接口），并支持模型的序列化压缩与持久化保存。此外，HanLP 还支持用户自定义词典分词，

用户只需通过简单的追加配置即可实现词典的全局加载，并支持基于双数组 Trie 树构建 Aho Corasick 自动机以实现与词典的多模式极速匹配。

因而，首先，本文将应用 HanLP 的 SP 基础分词模型对章节 2 获取到的领域纯文本语料进行命名实体识别和新词发现，将得到的词与章节 3 构建的领域基础知识词典进行匹配比对，将匹配不到的词作为新词补充进词典，实现词典的扩充。随后，将词典按 HanLP 要求的格式输出和追加配置到 HanLP 文件路径中，作为全局自定义词典备用。其次，使用 SP 基础分词模型的词性标注功能，将词典中的词再一次按 SP 框架要求格式标注并生成 txt 文档：①词与词性之间须用英文斜杠“/”分隔，所有词及标点符号都必须标注词性；②词与词之间须用空格分隔；③支持用英文方框“[ ]”将多个单词预处理合并为复合词，如[深度/n 学习/vn]/nz。接着将 txt 文档作为输入，调用 SP 序列化标注框架的在线学习 Learn 接口，实现 SP 基础分词模型的增量训练，以获得模型的泛化能力。最后，以泛化后的新 SP 模型为基础，加载自定义词典，除了章节 3 自定义的领域基础知识词典外，还可以加入机构名词典、人名词典、地名词典、现代汉语词典等，结合调用多模式极速匹配词典分词方法和 SP 分词方法实现对章节 2 的领域纯文本语料的自动分词。同时，在分词过程中，还可通过词性识别与句法分析，并结合同义词词典、停用词词典、特殊字符词典等对常见同义词、停用词、标点符号及特殊符号等进行替换或滤除。如果是英文语料，还要考虑对单词中字母大小写、一些单复数形式进行词干提取和词型还原的统一规范化处理等。

5 实验与结论

基于项目工作需要，本文以海洋科学领域为例开展了实验研究。在多源异构领域语料的自动获取环节，同时采集了以领域科技资讯（新闻、动态、政策）为主的半结构化网络开放领域语料和以期刊论文为主的半结构化科学文献领域语料。其中，领域科技资讯语料的获取主要通过 Incites 分析海洋科学领域全球相关机构排行，从中人工遴选了 17 个中国的重点海洋研究机构（研究所、大学），将它们机构网站的科技进展、科研动态、科技要闻等栏目作为采集源，由于后面发现大多机构网站更新并不是特别频繁，因此又将中国海洋在线网站[18]（中国海洋报主办）的要闻、科技发展、高新技术栏目，国家海洋局&教育部共建的中国海洋发展研究中心网站的海洋要闻栏目等一起加入了采集源，设计 Web Spider

主要对其 2018-2019 年发布的信息进行了自动监测采集和内容抽取。领域科学文献语料的获取主要通过单位订购的 Web of Science 数据库，选择中国科学引文数据库（CSCD），以“SU=OCEANOGRAPHY”作为检索式进行专业检索，编写 Web Services API 调用程序和 XML 解析程序对期刊论文相关的元数据进行自动采集和解析抽取。最终成功获取到海洋科学领域有效网络开放领域语料 1953 篇，中文学术文献语料 23403 篇，共 25356 条记录，初始存储在 MySQL 关系数据库表中，后期可根据需要按格式输出整合在 TXT 文档中，作为学习语料集。在领域基础知识词典的自动构建环节，将 23403 篇中文学术文献的中英文作者关键词提取出来，按章节 3 所提出的方法生成领域基础知识词典 Sea_Dic，共得到有效词组 35197 个，见表 4。在领域文本的自动分词与新词发现环节，基于 Sea_Dic 和 HanLP 提供的 SP 基础分词模型和标注框架，增量训练出经过领域泛化的 SP 词法分析模型，并对学习语料集进行预处理，以一篇资讯中的内容处理示例见图 5，其中灰色的表示基于领域基础词典所识别出的可能的新合成词，其他表示常规的基础分词。

表 4　海洋科学领域基础知识词典示例

中文词	英文词	词频	同义词	上位词
南海	Nanhai Sea; South China Sea	689	南中国海	中国海;北太平洋
数值模拟	Numerical simulation	627	数字模拟;数值拟合;…	数学模拟
沉积物	Sediment	595	淀积物;沉淀物;…	--
长江口	Yangtze estuarine	297	--	--
东海	Donghai Sea; East China Sea	275	--	中国海;北太平洋
海洋生物化学	Marine Biochemistry	253	海洋藻类化学	生物化学;海洋化学
海水淡化	Desalination	251	盐水转化	海水处理
渤海	Bohai Sea	246	--	中国海;北太平洋
天然气水合物	Natural gas hydrate	242	天然气水化物; 可燃冰;…	气体水合物
潮流	Tidal current	219	潮汐水流;波流;…	水流

受 海域 泥质粉砂水合物 地层 传热传质 效率 极低的 制约 ， 目前 水合物 开采 面临 降压 困难 、产能 较低 等 瓶颈 ，且 长期 开采过程 中 必然 面临 地层物质亏空 的 难题 。 降压法 无法 解决 水合物 长期 开采条件 下的 地层亏空 问题 ， 常规 防砂作业 面临着 因为 地层亏空 造成的 防砂 失效的 挑战 。 一次性 裸眼砾石充填 防砂完井作业 虽然 能 在短期内 起到 良好的 作用 ， 但 由于 没有 后续 物源 补给 ， 造成 防砂有效期 短 ， 不足以 满足 海洋天然气水合物 长期 开采的 需求 。长期稳定 的 水合物 生产 迫切需要 对 地层亏空量 进行 及时的 填充 或 置换 。

图 5　海洋领域文本预处理示例

实验表明，本文提出的方法能够有效获取到领域学习语料，并实现分词等预处理任务，过程具有可操作性和可移植性。但也有需要进一步优化的地方，如对于领域基础词典的构建来说，可以现有方法为基础继续进行语义扩充，除了同义词、上位词等，考虑将下位词、相关词等其他语义关系词也逐渐扩充进来，并和已有的中图分类法等分类体系映射起来，形成更为丰富的小型领域基础词典。此外，增量训练领域分词模型的过程比较复杂，未来可考虑结合其它分词方法或工具对分词模型及精度进行优化提升，形成更加强大的更加通用的领域预处理模型。

注释

[1]唐静. 叙词表转换为 Ontology 的研究[J]. 情报理论与实践, 2004, 27(6): 642-645.

[2]何燕, 穗志方, 李素建, 等. 基于专业术语词典的自动领域本体构造[J]. 情报学报, 2007, (1): 65-70.

[3]蔡盈芳, 黄磊. 航空领域本体构建研究[J]. 情报学报, 2010, (2): 223-231.

[4]张翔, 苏晓龙, 吴文辉. 半结构化数据领域本体构建算法及实现[J]. 计算机与信息技术, 2011, (Z1): 37-40+44.

[5]焦晓龙. 基于 Web 数据表抽取的领域本体构建方法研究[D]. 沈阳：东北大学，2012.

[6]宋丹辉. 基于 RDFS 的用户本体的构建与优化研究[J]. 图书馆学研究, 2012,(13): 2-8.

[7]郭瑞. 基于纯文本的领域本体构建与实现[D]. 石家庄：河北科技大学, 2016.

[8]任飞亮, 沈继坤, 孙宾宾, 等. 从文本中构建领域本体技术综述[J]. 计算机学报, 2017, (40): 1-27.

[9]Wisniewski M. Metamodel of Ontology Learning from Text[C].Emergent Web Intelligence: Advanced Semantic Technologies. Advanced Information and Knowledge Processing. Springer, London, 2010: 245-276.

[10]Baroni M, Lenci A. How We BLESSed Distributional Semantic Evaluation[C]. Proceedingsof the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics,

2011.
[11]Lison P, Tiedemann J, Kouylekov M. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora[C]. Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, 2018.
[12]Sketch Engine | Language Corpus Management and Query System[DB/OL]. [2019-06-25]. https://www.sketchengine.eu/.
[13]王思丽, 刘巍, 祝忠明, 等. 基于 CSpace 的科技信息可配置化自动监测功能设计与实现[J]. 数据分析与知识发现, 2017, 1(10): 85-93.
[14]Wu WT, Li HS, Wang HX, et al. Probase: a Probabilistic Taxonomy for Text Understanding[C]. SIGMOD '12 Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, Arizona, USA, 2012: 481–492.
[15]汉语主题词表服务系统[DB/OL]. [2019-04-15]. https://ct.istic.ac.cn/site/organize/index.
[16]中文分词原理及工具[EB/OL]. [2019-04-10]. https://www.cnblogs.com/palace/p/9629614.html.
[17]程志远. 基于神经网络的中文分词研究[D]. 郑州：郑州大学, 2019.
[18]王玮. 基于 Bi-LSTM-6Tags 的智能中文分词方法[J]. 计算机应用, 2018, S2: 107-110.
[19]Second International Chinese Word Segmentation Bakeoff[EB/OL]. [2019-04-10]. http://sighan.cs.uchicago.edu/bakeoff2005/.
[20]HanLP: Han Language Processing[EB/OL]. [2019-06-25]. https://github.com/hankcs/HanLP/blob/master/README.md.

作者简介：

王思丽(ORCID: 0000-0002-2126-3462)，女，1985，中国科学院西北生态环境资源研究院文献情报中心/中国科学院大学，馆员，博士研究生，研究方向：知识发现与知识组织。
祝忠明(ORCID: 0000-0002-2365-3050)，男，1968，中国科学院西北生态环境资源研究院文献情报中心，研究馆员，博士生导师，研究方向：知识发现与知识组织、知识管理系统建设。
刘巍(ORCID: 0000-0001-6387-1709)，男，1980，中国科学院西北生态环境资源研究院文献情报中心，副研究馆员，硕士生导师，研究方向：知识计算与知识挖掘。
杨恒，男，1992，中国科学院西北生态环境资源研究院文献情报中心，助理馆员，硕士，研究方向：分布式大数据系统建设。

定稿日期：2019-07-25