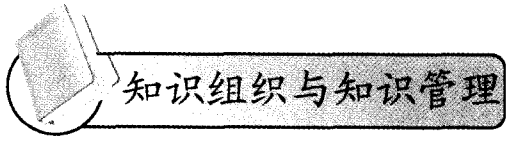
国家哲学社会科学学术期刊数据库
National Social Sciences Database



CSSCI 语料中短语结构标注与自动识别*

谢 靖[1] 苏新宁[2] 沈 思[2]
[1](南京中医药大学经贸管理学院 南京 210046)
[2](南京大学信息管理学院 南京 210093)

【摘要】将短语结构标注引入 CSSCI 期刊论文题录信息分析,在关键词、术语构成上从语法角度深度探讨各组成词汇之间的语法关系,力图通过语法功能分析揭示其所蕴含的语义知识。在进行一定规模语料标注基础上,通过短语词汇、词性统计及短语语法功能分析获取学术文献中短语结构构成特征,并将这部分特征与清华树库语料短语特征混合,提高短语结构在科技文献中的识别率。

【关键词】短语结构标记 CSSCI 语料 混合特征 自动识别
【分类号】TP391

Chinese Phrase Tagging and Automated Annotation Based on CSSCI Corpus

Xie Jing[1] Su Xinning[2] Shen Si[2]
[1](School of Economics and Management, Nanjing University of Chinese Medicine, Nanjing 210046, China)
[2](School of Information Management, Nanjing University, Nanjing 210093, China)

【Abstract】The paper introduces a new syntax method as the solution of term phrase identification on CSSCI corpus, and obtains the inter - relationship among terms in academic literature from the linguistic aspect based on phrase components, such as words, part - of - speech, grammar functions, etc. These linguistic features are mixed with phrase features which are extracted from Tsinghua Treebank so as to leverage the accuracy of phrase auto - identification in academic corpus.

【Keywords】Phrase annotation CSSCI corpus Multi - feature Auto - identification

1 引 言

目前的科技文献检索在内容检索方面主要是通过关键词检索完成,其中涉及到词面相似度计算、关键词概念层次计算等方面。在专家的帮助下,可以通过主题法、分类法、概念描述体系、主题分类一体化等传统检索语言方式实现对检索的扩充与优化,如可以通过对检索词的分析,寻找其中的上下位类关系,对关系最为密切的词汇概念进行二次检索。通过关键词的共现也可以寻找这些词汇之间的潜在关系,并对共现较多的词汇进行联合查询。但无论主题法、分类法、概念描述,还是词汇共现都有一定的缺陷,即对于主题法、分类法、概念描述,需要领域专家和语言学家的介入,需要消耗大量的人工与精力,同时主题法、分类法、概念描述体系具有一定的滞后性,不能反映出当前学科关键词及知识的创新,在使用时用户也需要对分类法进行学习后才能正确标注与查询。

收稿日期:2012 - 11 - 14
收修改稿日期:2012 - 12 - 14
*本文系国家自然科学基金面上项目"面向知识服务的知识组织模式与应用研究"(项目编号:71273126)、高技术研究发展计划(863 计划)项目"以科技文献服务为主的搜索引擎研制"(项目编号:2011AA01A206)和江苏省教育厅高校哲学社会科学研究基金项目"基于本体的高校突发事件网络舆情监控预警模式研究"(项目编号:2010SJB870003)的研究成果之一。

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

在计算语言学取得重大进展的前提下,可以通过将语言学知识引入信息检索领域,借助短语知识反映关键词、术语内部成分间的语法关系,也可以通过对标题、摘要甚至正文中小句的短语分析获取关键词、术语之间的语法联系,这种联系可以作为语义信息的组成部分。

本文在清华树库标注体系基础上,将其应用于CSSCI语料短语标注,选取"知识组织"、"知识服务"为检索词,将相关文献中的标题、关键词进行短语识别与标注。在此基础上,考察在学术语料中汉语短语组成及其分布情况,这主要从组成词汇、词性序列及短语语法功能等方面进行考察。这些特征在随后被加入以清华树库短语知识为基础的训练语料,提高短语标注在领域科技文献中的识别率。在人工标注过程中,标题、关键词短语中定中结构出现频次最高,本实验即选取定中结构为研究对象,并以最短定中结构(无嵌套定中结构)设定相关语言学特征。实验结果精确率达86.79%,召回率82.51%,F值为84.60%,表明这种混合标注策略具有一定实用价值。

2 国内外研究现状

在计算语言学方面,Chomsky[1]在1957年提出了转换生成语法,将语言学由经验主义引入理性主义,通过转化生成理论可以对句子进行机器规则识别。Abney[2]在1991年提出句子可以被划分为更小的组成即Chunk,这种基于组块的句法与基于转换规则的句法有很大不同,通过Chunk可以将句子划分为若干更小的结构,对这些小的结构可以进行进一步的观测。基于组块的语料库使得计算语言学得到飞速发展,通过统计的方法可以完成对语言规则的抽取,同时也避免了基于规则研究的局限性。美国宾夕法尼亚大学在1989年启动了"The Penn Treebank Project",并于20世纪90年代推出英文U-Penn树库,其语料来源为华尔街语料(WSJ)、布朗语料(BROWN)以及两个口语语料SWBD和ATIS[3]。周强[4]通过多层标注设计并标注了清华大学中文树库(Tsinghua Chinese Treebank, TCT),该树库是国内比较成熟的中文短语树库,也是国内第一个基于平衡语料的短语树库。

树库的出现为计算语言学提供了学习素材和实验平台,尤其是其中短语的语法功能可以作为研究词汇间语义关系的基础。通过对中文短语树库的统计与观测,可以对其中短语结构进行自动识别,通过对短语的语言特征统计,基于统计的模型可以对短语进行序列标注,进而实现机器学习,最后在语料库上对标注和训练情况进行对比分析。陈静等[5]在大规模语料的基础上,通过对兼语结构语言学特征统计,使用条件随机场进行短语自动识别。朱丹浩等[6]通过对清华树库中介宾结构内外部语言特征的统计,使用条件随机场模型对介宾结构进行自动识别。

近年来,在术语相关的研究中,开始引进自然语言的计算机处理方法和技术,出现了"计算术语学"(Computational Terminology)学科。冯志伟[7]在1988年就注意到术语的自动处理问题,他在德国夫琅禾费研究院(Fraunhofer Institute)使用计算机对汉语的词组型术语进行了自动结构分析,他也是较早将计算术语学引入我国的学者,介绍了国外学者对术语的发现、术语的充实、术语的受控标引、术语的自由标引等问题的研究[8]。在词汇及词组的概念上,冯志伟[9,10]对中文单词型术语及短语型术语均进行了结构分析,并从语言学的经济原则角度探讨了单词型术语和短语型术语在术语库中的分布,提出了术语形成的经济律(FEL公式),从数学公式上完成了对术语系统的经济指数、单词的术语构成频率F和术语的平均长度L的公式表达[11]。这些研究将语言学知识引入术语短语研究,通过这种研究可以在语义上获取更多关于术语概念、成分等信息。

3 CSSCI标注语料

在清华树库标注体系基础上,通过对CSSCI关键词、标题的简单分词及词性标注,借助清华树库短语知识进行辅助标注。标注语料为以"知识组织"、"知识服务"为关键词对所有标题进行检索,将相关文献关键词及标题通过ICTCLAS进行初步分词及词性标注,共涉及文献369篇。在分词及词性标注基础上,通过清华树库相关词汇及短语前后词汇、词性知识进行人工辅助标注,具体标题标注样例如下:

[np-DZ [np-DZ [sp-FW [np-DZ 新/a 环境/n] 下/f] 的/u [np-DZ 图书馆/n [np-DZ 知识/n 服务/vN]]] 探讨/vN]

对应文献关键词标注为:

[np-DZ 网络/n 环境/n];[np-DZ 知识/n 服务/vN];[np-DZ 信息/n 资源/n];

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

表1 CSSCI语料辅助标注结果
—以“知识组织”、“知识服务”为例

标题标注	短语标注	修正
[np-DZ [np-DZ [vp-PO 面向/v [np-DZ 技术/n 创新/vN]] 的/u [np-DZ 知识/n 服务/vN]] 研究/vN]	[np-DZ 技术/n 创新/vN];[np-DZ 知识/n 服务/vN];企业/n;	
np-DZ [np-DZ [np-DZ 中小型/b [np-DZ [np-DZ 科研/n 系统/n] 图书馆/n]] [np-DZ [np-DZ 知识/n 服务/vN] 模式/n]] 初探/n]	[np-DZ 科研/n 系统/n];图书馆/n;[np-DZ 知识/n 服务/vN];[np-DZ [np-DZ 知识/n 服务/vN] 模式/n];[np-DZ 应用/vN 策略/n];	
[np-DZ [np-DZ [pp-JB 基于/p 博客/n] 的/u [np-DZ 图书馆/n [np-DZ [np-DZ 知识/n 服务/vN] 模式/n]]] 研究/vN]	博客/n;图书馆/n;[np-DZ 知识/n 服务/vN];	
[np-XX [np-DZ 知识/n 服务/vN] [dlc-BC -/w -/w [np-DZ [np-DZ [tp-DZ 21/m 世纪/n] 图书馆/n] 的/u [np-DZ 发展/vN 方向/n]]]]	[np-DZ 知识/n 经济/n];[np-DZ 知识/n 服务/vN];[np-DZ 信息/n 服务/vN];	
[np-LH [np-DZ 网格/n 技术/n] 与/c [np-DZ [np-DZ 数字/n 图书馆/n] [np-DZ 知识/n 服务/vN]]]	[np-DZ 数字/n 图书馆/n];网格/n;[np-DZ 知识/n 服务/vN];	
[vp-PO 构建/v [np-DZ [pp-JB 基于/p [np-DZ [vp-AD 学科/n 化/k] [np-DZ 知识/n 服务/vN]]] 的/u [np-DZ 信息/n 门户/n]]]	[np-DZ 知识/n 服务/vN];[np-DZ 学科/n 门户/n];[np-DZ 信息/n 门户/n];	[学科/n 化/v]->[学科/n 化/k]
[np-DZ [pp-JB 基于/p [np-DZ 网格/n 技术/n]] 的/u [np-DZ [np-DZ 数字/n 图书馆/n] [np-DZ 知识/n 服务/vN]]]	[np-DZ 数字/n 图书馆/n];网格/n;[np-DZ 知识/n 服务/vN];	
[np-DZ [np-DZ [sp-FW [np-DZ 新/a 环境/n] 下/f] 的/u [np-DZ 图书馆/n [np-DZ 知识/n 服务/vN]]] 探讨/vN]	[np-DZ 网络/n 环境/n];[np-DZ 知识/n 服务/vN];[np-DZ 信息/n 资源/n];	
[np-DZ [np-DZ 我国/n [np-DZ [np-DZ 知识/n 服务/vN] 研究/vN]] [np-DZ 论文/n [np-DZ 定量/b 分析/vN]]]	[np-DZ 知识/n 服务/vN];[np-LH 统计/vN 分析/vN];[np-DZ 计量/vN 分析/vN];	[定量分析/l]->[定量/b 分析/vN]
[np-DZ [np-DZ [sp-FW [np-DZ [np-DZ 知识/n 服务/vN] 体系/n] 中/f] [np-DZ [np-DZ 自主/d 协同/vN] 机制/n]] 的/u 实现/vN]	[np-DZ 自主/d 协同/vN];[np-DZ 知识/n 服务/vN];[np-DZ 智能/n 代理/vN];	[自主/v]->[自主/d];[协同/v]->[协同/vN]

需要说明的是,在辅助标注的同时,本文对其中分词及词性标注错误进行了修正,主要包括几个方面:

(1)分词错误,如“[定量分析/l]->[定量/b 分析/vN]”,这类错误较少,可在今后标注中通过导入领域词典减少切分错误;

(2)后缀词汇切分及标注错误,如“[学科/n 化/v]->[学科/n 化/k]”,这类问题在分词上存在争议,有学者认为应当将其视为一个词汇,但在本研究中,仍将这类词汇以后缀形式进行分词及词性标注,以便对此种构成的短语术语概念进行语言学分析;

(3)词性标记错误,例如“[协同/v]->[协同/vN]”,在关键词及标题中,对于研究对象涉及动词,科技语料与通用语料有很大不同,动词名词化并与研究对象构成术语现象较多,大部分词性标记错误均与此相关。

表1为部分标题及对应关键词短语标注结果,同时也附上相关错误校正记录。在标注基础上,通过对标注语料解析获取相关标题、关键词中的短语知识,这些知识将在今后研究中用于统计及机器训练。

4 CSSCI标注语料分析

4.1 关键词短语分析

在这部分标注语料中,可以通过关键词短语标注获取与“知识组织”、“知识服务”相关的术语词汇、短语,并分析其语言学构成,进而进行语义分析。表2为相关关键词单词型、词组型分布情况,表3为关键词短语层次分布情况。

表2 标注CSSCI语料关键词类型及所占比例

类型	单词型	所占比例	词组型	所占比例
未去重	209	15.64%	1 127	84.36%
去重后	85	15.18%	475	84.82%

表3 标注CSSCI语料关键词短语层次分布情况

层次	频次	样例
1	320	[np-DZ 知识/n 服务/vN]、[np-DZ 知识/n 管理/vN]
2	120	[np-DZ [np-DZ 知识/n 服务/vN] 模式/n]
3	34	[np-DZ 读者/n [np-DZ [vp-AD 个性/n 化/k] 服务/vN]]
4	1	[np-DZ [np-DZ [np-DZ [np-DZ 知识/n 组织/vN] 工具/n] 类型/n] 划分/vN]

从表2可知,在“知识服务”、“知识组织”相关文献中,短语型关键词占绝大多数,所占比例为84.36%,经过去重后,所占比例为84.82%。在科技文献中,短语型术语占绝大多数,因而可以通过短语标注研究其构成。从表3可知,在词组型关键词中,一层短语占大多数,二层短语次之,三层短语及以上较少。这个数据表明,在短语型关键词中,学者较多使用一层短语,从短语结构及词性序列中,可以看出这与科学研究本身有密切联系。

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

表 4 标注 CSSCI 语料关键词短语结构分布情况

序号	短语结构	标记	频次	样例
1	定中结构	DZ	449	[np - DZ 知识/n 服务/vN]、[np - DZ 知识/n 组织/vN]
2	附加结构	AD	11	[np - AD 主题/n 法/n]、[vp - AD 个性/n 化/k]
3	状中结构	ZZ	5	[vp - ZZ 互/d 操作/v]
4	联合结构	LH	4	[np - LH 参考/vN 咨询/vN]
5	主谓结构	ZW	3	[dj - ZW 服务/vN 流畅/a]
6	述宾结构	PO	2	[vp - PO 开发/v 链接/vN]
7	标号结构	BH	1	[np - BH《/w 中分表/j 》/w]

表 5 标注 CSSCI 语料关键词短语内部词性序列前 10 位

序号	词性序列	频次	样例
1	n + n	118	[np - DZ 高校/n 图书馆/n]
2	n + vN	87	[np - DZ 知识/n 组织/vN]
3	vN + n	48	[np - DZ 发展/vN 趋势/n]
4	n + vN + n	32	[np - DZ [np - DZ 知识/n 组织/vN] 系统/n]
5	n + n + n	25	[np - DZ [np - LH 图书/n 情报/n] 机构/n]
6	vN + vN	21	[np - LH 参考/vN 咨询/vN]
7	n + n + vN	17	[np - DZ 图书馆/n [np - DZ 知识/n 服务/vN]]
8	a + n	8	[np - DZ 客观/a 知识/n]
9	n + k	8	[np - AD 图书馆/n 法/k]
10	b + n	5	[np - DZ 内在/b 结构/n]

从表 4 和表 5 可知，在短语型关键词中，定中结构占绝大多数，其词性序列主要有“n + n”、“n + vN”、“vN + n”、“a + n”等。在结构上，定中结构更能反映科学研究对象以及相关概念词汇间的组合关系。例如，词性序列为“n + n”的定中结构中，两个名词词汇往往是研究对象，如“[np - DZ 高校/n 图书馆/n]”；词性序列为“n + vN”的定中结构中，名词为研究对象，通过与名词化的动词结合，形成特定概念术语，如“[np - DZ 知识/n 组织/vN]”，这类术语中，动名词与名词间往往可以通过转化形成述宾结构，即“[vp - PO 组织/v 知识/n]”。但是，并非所有“n + vN”定中结构都可以转化为述宾结构，例如“[np - DZ 知识/n 服务/vN]”，这里就名词与涉及动词的支配关系有关，这里只能理解为“提供知识的服务”或是“基于知识的服务”。

在以“知识组织”、“知识服务”为检索词的文献关键词中，“知识”是研究对象，通过与动词“组织”、“服务”的名词化相结合，形成独立的概念，在此基础上又可以继续与“工具”、“系统”等词汇继续结合形成新的概念。在进一步结合的过程中，由于关注点不同，学者对于其中词汇的倾向也不同，在短语词汇结合上会产生差异。其中由研究对象词汇与其他词汇结合为新的概念，并继续构建相关概念理论、方法、工具的过程如图 1 所示，示例如图 2 所示。

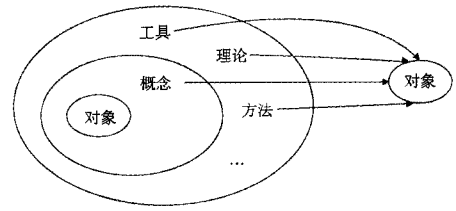


图 1 标注 CSSCI 语料关键词短语构造过程

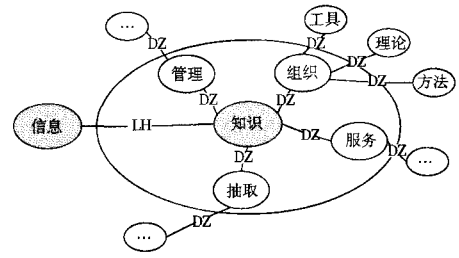


图 2 标注 CSSCI 语料关键词短语构造示例

4.2 标题短语分析

在研究标题中这些词汇无关的概念之间的短语关系之前，首先对其进行基本情况统计，主要包括短语结构统计、短语功能统计、层次分布情况、词性序列统计等。表 6 为标题中短语层次统计，表 7 为主要短语结构统计。

表 6 标注 CSSCI 语料标题中短语层次分布情况

层次	频次	样例
1	1 017	[np - DZ 知识/n 服务/vN]、[np - DZ 知识/n 管理/vN]
2	457	[np - DZ 图书馆/n [np - DZ 知识/n 服务/vN]]
3	311	[np - DZ 图书馆/n [np - DZ [np - DZ 知识/n 服务/vN] 模式/n]]
4	212	[np - DZ [pp - JB 在/p [np - DZ [np - DZ 知识/n 服务/vN] 方面/n]] 的/u 应用/vN]
5	183	[np - DZ [sp - FW [np - DZ [np - DZ 高校/n 图书馆/n] 服务/vN] 中/f] 的/u [np - DZ 人文/n 关怀/vN]]
6	138	[np - DZ [np - DZ [pp - JB 基于/p 博客/n] 的/u [np - DZ 图书馆/n [np - DZ [np - DZ 知识/n 服务/vN] 模式/n]]] 研究/vN]
7	116	[fj - BL [dj - ZW 信息/n [vp - ZZ [pp - JB 以/p 人/n] [vp - PO 为/vC 本/n]]] [dj - ZW 知识/n [vp - PO 服务/v 大众/n]]]
8	93	[np - DZ [np - LH 知识库/n 和/c [np - DZ [np - DZ 知识/n 发现/vN] 技术/n]] [np - DZ [pp - JB 在/p [np - DZ [np - DZ 知识/n 服务/vN] 方面/n]] 的/u 应用/vN]]
9	58	[np - DZ [np - DZ [sp - FW [np - DZ [np - DZ 知识/n 经济/n] 环境/n] 下/f] [np - DZ 图书馆/n 的/u [np - LH [np - DZ 知识/n 管理/vN] 与/c [np - DZ 知识/n 服务/vN]]]] 研究/vN]
10	30	[np - DZ [sp - FW [np - DZ [np - DZ 语义/n 网格/n] 环境/n] 下/f] [np - DZ [np - DZ 数字/n 图书馆/n] [np - DZ [np - DZ 知识/n 组织/vN] 的/u [np - DZ [np - DZ 语义/n 互联/vN] 策略/n]]]]

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

表7 标注 CSSCI 语料关键词短语结构分布情况

序号	短语结构	标记	频次	样例
1	定中结构	DZ	1 938	[np - DZ 科研/n 系统/n]、[np - DZ 知识/n 组织/vN]
2	联合结构	LH	132	[np - LH [np - DZ 网格/n 技术/n] 与/c [np - DZ 数字/n 图书馆/n] [np - DZ 知识/n 服务/vN]]
3	介宾结构	JB	130	[pp - JB 基于/p [np - DZ 网格/n 技术/n]]
4	述宾结构	PO	122	[vp - PO 面向/v [np - DZ 技术/n 创新/vN]]
5	状中结构	ZZ	61	[vp - ZZ 更/d [vp - PO 需/v [np - DZ 团队/n 协作/vN]]]
6	方位结构	FW	60	[sp - FW [np - DZ [np - DZ 知识/n 服务/vN] 体系/n] 中/f]
7	缺省结构	XX	52	[np - XX [np - DZ 知识/n 服务/vN] [dlc - BC -/w -/w [np - DZ [np - DZ [tp - DZ 21/m 世纪/n] 图书馆/n] 的/u [np - DZ 发展/vN 方向/n]]]]
8	补充结构	BC	48	[dlc - BC -/w -/w [np - DZ [np - DZ [tp - DZ 21/m 世纪/n] 图书馆/n] 的/u [np - DZ 发展/vN 方向/n]]]
9	附加结构	AD	46	[np - AD [np - DZ 知识/n 服务/vN] 型/k]
10	主谓结构	ZW	24	[dj - ZW 知识/n [vp - PO 服务/v 大众/n]]

从表6可以看出，在以“知识组织”、“知识服务”为检索词的文献标题中，随着短语层深的增加，其分布数呈不均匀下降趋势。结合表7以及关键词短语标注情况，在关键词共现短语间若不存在共现词汇时，通过定中、介宾、方位、补充等短语结构，可以将这些词组型术语短语进行语法连接，不同的短语结构具有不同的语义特征。例如，定中结构可以表现研究对象概念的扩展，如“知识”通过定中结构扩展为“知识管理”、“知识抽取”、“知识组织”、“知识服务”，进而扩展到“知识组织工具”、“知识组织方法”等。相关涉及概念可以通过介宾与定中结合，完成两个及以上术语概念的语法联系，如“[np - DZ [pp - JB 基于/p [np - DZ 网格/n 技术/n]] 的/u [np - DZ [np - DZ 数字/n 图书馆/n] [np - DZ 知识/n 服务/vN]]]”，“网格技术”通过介宾短语，与“数字图书馆”、“知识服务”这两个概念词组联系起来。介宾结构在标题级短语标注中，通常可以表现概念间的方法、工具、理论等语义，方位结构则可以表现概念间的条件及上下位概念关系，补充结构通常为论文标题副标题，即对论文主标题的进一步语义解释等。通过结构组成分析，可以获取相关词汇、概念短语之间的语法联系，进一步分析其语义联系。

需要说明的是，对标题中的短语标注时以相应关键词短语为参考标准，不同学者对相同文本的理解存在差异，其关注点也会随之变化，对于相同序列的词汇短语标注也应随其认知而改变。在图2基础上，图3对内容相关的术语短语间通过更高层次短语知识进行关联的展示。通过高层次的定中、介宾、方位、补充等，可以实现多个术语概念短语的共现分析，这种分析可以获取概念短语间的语法关系，从而获取其修饰关系。如在定中结构中，充当定语部分的概念短语修饰充当中心语部分的概念短语，介宾结构中作为宾语部分的概念短语通过更高层次的定中结构，充当高层中心语短语的修饰部分，语义上可以进一步分析为工具、理论、方法等。

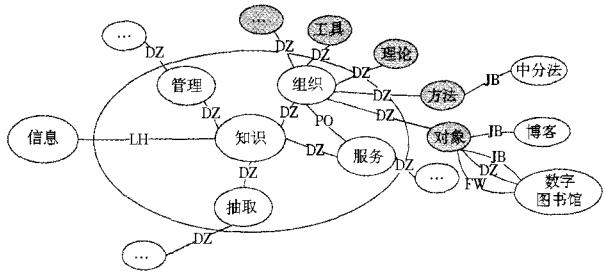


图3 标注 CSSCI 语料标题中术语短语间语法关联

5 基于 CSSCI 短语自动识别

在以上标注语料统计中，通过“知识组织”、“知识服务”对标题进行检索的结果中，关键词短语大部分为一层短语，且定中结构为多数。本实验以一层定中结构为识别对象，以清华树库中无嵌套定中结构为训练样本，同时将关键词短语中定中结构短语作为辅助，进行混合样本训练，用于识别标题中最短定中结构。由于人工标注工作量较大，而只是用通用语料特征进行训练会缺失专业领域知识。通过少量关键词短语标注，辅以通用语料短语内外部特征，进行 CSSCI 标题中最短定中结构识别。选择 CRFs 为训练模型[12]，训练模板设置如表8所示，识别情况如表9所示，结果统计如表10所示。

从表10可知，在标注语料标题中，共有最短定中短语848个，识别结果为892个。其中，前后边界一致、前界一致后界错误的短语均为736个，其精确率为86.79%，召回率为82.51%，F值为84.60%。通过对最小定中结构的识别，可以获取标题中与关键词短语

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

表 8 条件随机场特征选择

序号	特征	说明
1	Current Word - 3	当前观察词前三位词汇
2	Current Word - 2	当前观察词前二位词汇
3	Current Word - 1	当前观察词前一位词汇
4	Current Word	当前观察词
5	Current Word + 1	当前观察词后一位词汇
6	Current Word + 2	当前观察词后二位词汇
7	Current Word + 3	当前观察词后三位词汇
8	Current Word - 1, Current Word	当前观察词、当前观察词前一位词汇组合
9	Current Word, Current Word + 1	当前观察词、当前观察词后一位词汇组合
10	Current Word - 1, Current Word + 1	当前观察词前一位词汇、当前观察词后一位词汇组合
11	Current POS - 2	当前观察词前二位词性
12	Current POS - 1	当前观察词前一位词性
13	Current POS	当前观察词性
14	Current POS + 1	当前观察词后一位词性
15	Current POS + 2	当前观察词后二位词性
16	Current POS - 1, Current POS	当前观察词前一位词性组合、当前观察词词性
17	Current POS, Current POS + 1	当前观察词性、当前观察词后一位词性组合
18	Current Word, Current POS	当前观察词、当前观察词性组合

表 9 标注语料标题中最短定中结构识别情况

序号	词汇	词性标记	训练标记	结果标记
1	面向	v	S	S
2	技术	n	B	B
3	创新	vN	E	E
4	的	u	S	S
5	知识	n	B	B
6	服务	vN	E	E
7	研究	vN	S	S
8	中小型	b	S	S
9	科研	n	B	B
10	系统	n	E	E
11	图书馆	n	S	S
12	知识	n	B	B
13	服务	vN	E	E
14	模式	n	S	S
15	初探	n	S	S
16	基于	p	S	S
17	博客	n	S	S
18	的	u	S	S
19	图书馆	n	S	B
20	知识	n	B	E
21	服务	vN	E	S
22	模式	n	S	B
23	研究	vN	S	E

表 10 标注语料标题中最短定中结构自动识别结果统计

	精确率	召回率	F 值
前后边界识别一致	86.79%	82.51%	84.60%
前界一致	86.79%	82.51%	84.60%

最为相关的一层定中结构，其中绝大多数被作者标记为关键词。这些短语可以作为作者及用户关注焦点，通过更高层次的短语知识进行组合，从而通过语言网络进行语义扩充。

6 结 语

本文提出面向 CSSCI 的短语自动识别方法，并通过 CSSCI 语料标注及机器训练验证了这些思想在应用中的现实依据。在研究中，通过对语言学知识的引入，将 CSSCI 中关键词、术语短语之间建立语法功能关系，同时又通过对 CSSCI 数据的计量验证了短语语法功能等语言学思想，最后通过条件随机场对 CSSCI 中的最短定中结构完成基于混合训练语料的自动识别，取得很好的效果，其精确率为 86.79%，召回率为 82.51%，F 值为 84.60%。该数据表明，可以通过少量人工标注领域文本，辅以通用语料库知识进行领域语料中相应术语短语结构的识别，从而通过语法研究其构成词汇隐藏的语义知识。

参考文献：

[1] Chomsky N. Syntactic Structures[M]. Berlin: Mouton de Gruyter, 1957.

[2] Abney S P. Parsing by Chunks[A]. // Berwick R C, Abney S P, Tenny C L. Principle - Based Parsing[M]. Springer, 1991.

[3] The Penn Treebank Project[EB/OL]. [2012 - 09 - 12]. http://www.cis.upenn.edu/~treebank/.

[4] 周强. 汉语句法树库标注体系[J]. *中文信息学报*, 2004, 18(4):1 - 8. (Zhou Qiang. Annotation Scheme for Chinese Treebank[J]. *Journal of Chinese Information Processing*, 2004, 18(4):1 - 8.)

[5] 陈静，王东波，谢靖，等. 基于条件随机场的兼语结构自动识别[J]. *情报科学*, 2012, 30(3):439 - 443. (Chen Jing, Wang Dongbo, Xie Jing, et al. Automatic Identification of Concurrent Structure Based on Conditional Random Field[J]. *Information Science*, 2012, 30(3):439 - 443.)

[6] 朱丹浩，王东波，谢靖. 基于条件随机场的介宾结构自动识别[J]. *现代图书情报技术*, 2010(7 - 8):79 - 83. (Zhu Danhao, Wang Dongbo, Xie Jing. Automatic Identification of Prepositional Phrase Based on Conditional Random Field[J]. *New Technology of Library and Information Service*, 2010(7 - 8):79 - 83.)

[7] Feng Z W. Analysis of Chinese Terms in Data Processing[R]. Report in Fraunhofer Institute, 1988.

国家哲学社会科学学术期刊数据库
National Social Sciences Database

国家哲学社会科学学术期刊数据库
National Social Sciences Database

[8] 冯志伟. 一个新兴的术语学科——计算术语学[J]. 术语标准化与信息技术, 2008(4):4-9. (Feng Zhiwei. A New Scientific Domain in Terminology——Computational Terminology[J]. *Terminology Standardization & Information Technology*, 2008(4):4-9.)

[9] 冯志伟. 汉语单词型术语的结构[J]. 科技术语研究, 2004, 6(1):15-20. (Feng Zhiwei. Structure of Word Terms in Chinese Language[J]. *Chinese Science and Technology Terms Journal*, 2004, 6(1):15-20.)

[10] 冯志伟. 汉语词组型术语的结构[J]. 科技术语研究, 2004, 6(2):35-37. (Feng Zhiwei. Structure of Chinese Phrase Term[J]. *Chinese Science and Technology Terms Journal*, 2004, 6(2):35-37.)

[11] 冯志伟. 术语形成的经济律——FEL 公式[J]. 中国科技术语, 2010, 12(2):9-15. (Feng Zhiwei. Economic Law of Term Formation— FEL Formula[J]. *Chinese Science and Technology Terms Journal*, 2010, 12(2):9-15.)

[12] CRF++: Yet Another CRF Toolkit[EB/OL]. [2012-09-11]. http://crfpp.sourceforge.net/.

(作者 E-mail:bmy_xj@163.com)

Data Harmony 3.8 被选为 KMWorld 2012 年度引领潮流产品之一

Data Harmony 软件组件是数字化内容组织领域带头人 Access Innovations 的产品。最近,该软件组件被选为 KMWorld 2012 年度引领潮流产品。KMWorld 致力于知识管理、内容管理和文档管理领域新闻、趋势和案例研究。KMWorld 列举了 Data Harmony 软件组件的一些新功能,包括使用 SharePoint 2010 整合 Thesaurus Master® (叙词表和元数据管理工具)的能力,改进的多语言支持,一个动态的以交互的树形式展示叙词表的视图模式,以及一个更容易使用的管理模块。

Data Harmony 软件组件提供了一套知识管理解决方案,基于系统的分类法和叙词表对信息资源进行组织。该软件工具提供了一个强大的、有效的进行分类、标引、及数据过滤的系统。

Data Harmony 的应用程序提供卓越的内容管理功能,能够有效、灵活、可扩展地进行内容管理,使得用户能够迅速并准确地找到所需要的信息。

Access Innovations 公司总裁 Marjorie M. K. Hlava 指出,"经过约 30 年的发展,进入数字时代,我们仍然处于知识管理的最前沿,这是一项伟大的成就。我们很高兴能够为客户提供 Data Harmony 3.8 这个完整的工具包。"

KMWorld 的引领潮流产品评比活动始于 2003 年。今年,KMworld 的评审小组评估了 700 多个相关的产品。该评审小组成员由编辑人员、分析师、系统集成商、供应商、业务程序经理和用户组成。

"每一个参加评比的公司的产品都是整个大市场中的一员,这个市场对我们的读者来说是非常重要的。"KMWorld 杂志首席编辑 Hugh Mckellar 说。

Data Harmony 3.8 通过云提供服务,是一个托管的 SaaS 服务。欲了解更多有关 Data Harmony 的信息,请访问:http://www.dataharmony.com。

(编译自:http://www.accessinn.com/library/news/12-09-04-data-harmony-3.8-selected-as-a-kmworld-2012-trend-setting-product-of-the-year.html)

(本刊讯)

国家哲学社会科学学术期刊数据库
National Social Sciences Database