面向语义出版的学术文本词汇语义功能自动识别*

程齐凯[1,2]，李信[1,2]
（1. 武汉大学信息管理学院，武汉 430072；2. 武汉大学信息检索与知识挖掘研究所，武汉 430072）

摘要：为提高学术文献语义出版水平，既需要在写作和出版模式方面进行研究，也需要探索学术文本语义理解技术，以实现对学术文献，特别是存量学术文献的语义化处理。本文在学术文本词汇功能分析框架基础上，提出一种基于条件随机场的学术文献问题和方法识别模型，该模型使用词法特征、句法特征、组块特征等27个特征。实验表明，该方法具有优于当前最佳的识别效果。

关键词：词汇功能；语义出版；序列标注；学术文本
中图分类号：G23
DOI：10.3772/j.issn.1673-2286.2017.08.004

1 引言

科研大数据时代的来临，使科研工作者处于科研信息过剩的状态，以往单纯依靠人工搜索、阅读和分析学术文本来获取有价值的科研信息方式已经越来越不现实。为辅助解决这一问题，语义出版作为一种新型的出版方式和信息呈现技术，正发挥越来越重要的作用[1-2]。

语义出版以语义化表示技术呈现文献的内容、逻辑、结构，并将文本内容与现实世界的对象进行关联[3]。学术文献语义出版是语义出版技术在学术文献上的应用，以实现学术文献呈现方式的语义化和文献内容的机器可理解性。为推进语义出版的研究和实现水平，人们既需要对写作出版模式进行探索，也需要从技术研究视角出发，有针对性地研发面向文献内容理解的语义分析技术。

本文提出一种基于条件随机场和多语义特征的学术文本词汇功能识别方法。词汇功能在不同的文本领域有不同的理解，本文的词汇功能指词汇概念所映射的现实对象在科研活动中体现的作用。程齐凯已对学术文本词汇功能的定义和显现机理进行分析，构建了一个包含领域无关词汇功能和领域相关词汇功能的学术文本词汇功能框架，并通过标注得到学术文本词汇功能标注集[4]。

出于实用性考虑，本文采用一个简单实用的词汇功能分类方案，将词汇功能简单界定为研究问题和研究方法两类，前者指论文或者论文片段所期望解决的问题、构建的应用，后者是为解决问题而提出的方法。识别学术文献中词汇的语义功能，有助于机器更好地理解文献与抽取知识信息，辅助实现学术文献的语义化。

2 相关研究

有关学术文本词汇功能识别的研究还较少，关于识别方法研究的文献不到10篇。Knodo等最早对该问题进行探索，提出一种面向标题的学术文献研究问题、研究方法、研究领域识别方法，并且在日文和英文数据集上分别评测提出方法的效果，取得0.780和0.816的平均准确率[5]。由于标题构造具有一定的规律，面向标题的抽取方法常能取得较好的准确度表现，但在覆盖率

* 本研究得到中国博士后科学基金项目（编号：2016M602371）和国家自然科学基金青年项目“基于深度语义挖掘的引文推荐多样化研究”（编号：71704137）资助。

和召回率上有所不足，且难以处理构造不规律的标题。在后续研究中，Nanba等进一步对识别的对象范围进行了扩展，试图从摘要中识别文献的研究问题、研究方法等[6]，其将识别问题转换为分类问题加以解决，验证实验F1为0.24（准确率与召回率的加权平均值）；Gupta等借助模板和重抽样方法解决识别问题，通过不断扩展候选词和候选模板，得到用于识别词汇功能的句法模板，在ACL数据集上的实验结果表明，该方法在主要问题、技术、领域三个功能类别上取得的F1值分别为0.553、0.367和0.373[7]；Tsai等将词汇功能划分为技术、应用两类，采用重抽样策略和多特征结合的方法[8]，在Gupta数据集上，Tsai方法在技术和应用两个类上F1值分别为0.485和0.456；Tateisi等将学术文本词汇功能区分为方法、任务和其他三类，利用马尔科夫逻辑网络方法进行词汇功能识别，也取得一定效果[9]。

目前学术文本词汇功能自动识别的研究还处于初步探索阶段，已有方法的实际效果难以保证，识别方法的性能和效果都有不足，难以付诸实际的语义分析应用。为此，本文提出一种基于序列标注和多特征融合的词汇功能识别方法，试图从学术文本中有效地识别研究问题和研究方法。

3 研究方法

本文使用的学术文本词汇功能框架详见程齐凯前期的研究成果[4]。该框架将学术文本词汇功能区分为领域无关词汇功能和领域相关词汇功能。本文提出的方法主要关注领域无关词汇功能的两个重要类别，即研究问题和研究方法，采用条件随机场模型识别学术文本中体现的研究问题和研究方法。

3.1 标注问题表示

序列标注的第一个问题是采用何种标记。本文采用三种标记将词汇功能识别问题转化为标注问题，对于每个词汇，标注模型需要为词汇标上“M”“T”“O”标签中的一种。其中，“M”为“mehod”的缩写，用于标记方法词；“T”为topic的缩写，用于标记问题词；“O”用于标记其他词汇。

例如，对于文本“We propose a SVM based method for text categorization”，模型期望得到标签序列“OO OMMMOTT”。

得到标记序列后，通过反查词汇序列，即可发现该文本中“问题”概念为“text categorization”，“方法”概念为“SVMbased method”。

3.2 标注模型

基于三标记方法，机器学习模型需要针对输入文本生成期望的三标记标注序列。为完成这项工作，本文使用条件随机场（Conditional Random Fields，CRFs）模型。CRFs由Lafferty等提出，是一个应用广泛的序列标注模型[10]。本文将使用线性链条件随机场，条件随机场基本公式如下：

$$LL(D) = \sum_j log(P(s_j|o_j)) - \sum_k \frac{\lambda_k^2}{2\sigma^2}$$

模型训练的目标是基于训练数据学习得到模型各参数的最优取值。完成模型训练后，CRFs方法会对输入特征序列进行计算，以获得标注结果。

3.3 特征构造

在序列标注任务中，特征的构造会直接影响模型的标注效果。本文共构造27个特征，分为6个类别，即词/词组合、词性（POSTAG）、动词信息、组块特征、句法特征以及动词角色特征。

3.3.1 词汇特征

对于给定句子S，其对应词汇序列T=[LB，t_0，t_1…t_n，RE]，LB和RE为占位符，分别表示句子的开始和结尾。令t_i为目标词汇，0≤i≤n。针对目标词汇t_i，构造特征：

（1）t_i（当前词）；

（2）t_{i-1}（当前词的前一个词）；

（3）t_{i-2}（当前词的前第二个词，i=0时为空）；

（4）t_{i+1}（当前词的后一个词）；

（5）t_{i+2}（当前词的后第二个词，i=n时为空）；

（6）t_{i-2}_t_{i-1}（当前词前面两个词以“_”为拼接符的拼接结果）；

（7）t_{i+1}_t_{i+2}（当前词后面两个词以“_”为拼接符的拼接结果）。

此外，为标记当前词的大小写形态，构造两个布尔型的特征以标记当前词是否为全大写形式或者全小写

形式。

例1：*We propose a SVM based method for text categorization*.

在此例中，若当前词为“SVM”，则可以生成特征“SVM”“a”“propose”“based”“method”“propose_a”“based_method”。

“SVM”为全大写形式，可以得到特征“UPPER:TRUE”和“LOWWER:FALSE”。

如果标注粒度为词组，除上述特征外，还将构造3个新特征，分别为词组的第一个词、词组的最后一个词以及词组的长度。

以例1给出的句子为例，当标注粒度为词组且当前词组为“SVMbasedmethod”时，可以得到“FIRSTWORD:SVM”“LASTWORD:method”以及“LENGTH:3”3个特征。

3.3.2 POSTAG特征

对于给定句子S和对应词汇序列T，通过词性标注得到词性序列P=[LB，p_0，p_1…p_n，RE]，设需要为下标为i的词汇构造特征，则构造特征p_i、p_{i-1}、p_{i-2}、p_{i+1}、p_{i+2}。

以例1给出的文本为例，为“SVM”构造POSTAG特征，得到“NN”“DT”“VBP”“VBN”“NN”5个特征。

同词特征一样，如果标注粒度为词组，则构建3个新特征，分别是词组首词词性、词组最后一个词的词性、词组内各词词性的拼接形式。在例1所示文本中，就“SVMbased method”可得到特征“fPostag:NN”“lPostag:NN”和“iPostags:NN-VBN-NN”。

3.3.3 动词相关特征

给定文本和目标词汇，可构造3个动词相关特征，分别是目标词汇左边最近的第一个动词、目标词汇右边最近的第一个动词、距目标词汇最近的动词。

当目标词汇左边或者右边没有动词，则将对应的上述特征标记为“<NONE>”。以例1中的“SVM”为例，可构造“lVerb:propose”“rVerb:based”和“nVerb:based”。

3.3.4 组块分析特征

组块分析特征用于记录词汇所在组块的属性，常见组块类型有“NP”“VP”“PP”“PRT”等。目标词汇w所在组块的类型将构成其组块分析特征。

本文使用句法解析方法间接获得词的组块信息。对于给定句子S，其对应的句法树记为Pt，S中的词构成Pt的叶子节点，对每个叶子节点w，Pt中距w节点最近的组块标记将被用作w的组块标记。

图1给出了一个例子。其中，组块节点用方框标出。文本的组块识别结果见表1。

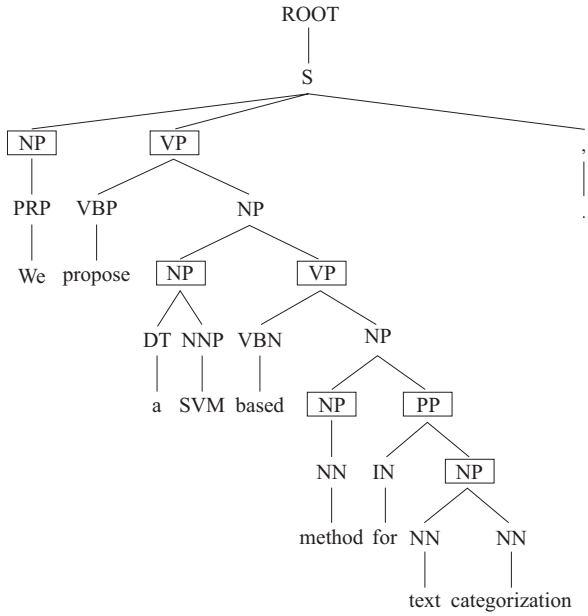


图 1 句子句法树解析及组块标注示意图

表 1 组块分析结果示例

序 号	组块标记	词 汇
1	NP	We
2	VP	propose
3	NP	a
4	NP	SVM
5	VP	based
6	NP	method
7	PP	for
8	NP	text
9	NP	categorization

3.3.5 句法特征

（1）Head词汇。用于记录词汇或词组的关键成分。如果标注对象为单一词汇，则Head词汇是其自身。

若对象是词组，则进行以下操作。

①构建有向网络，将词组中的单词加入网络，然后依据单词间的依存关系为节点构建边；②遍历节点，如果节点出度为0且入度大于0，则从网络中剔除该节点；③重复步骤②，直到网络中剩下的节点都是孤立节点；④如果词汇网络中仅余1个节点，返回该节点，否则返回“<MULI_HEAD>”。

如图2所示，为找出“DLmodel for named entity recognition”的Head词汇，首先利用词汇以及词汇间依存关系构造有向网络；其次，遍历删除网络中出度为0且入度大于0的节点，并重复此操作多次；最后，经过多轮遍历操作，网络中仅有1个节点，返回该节点对应的词汇“model”。

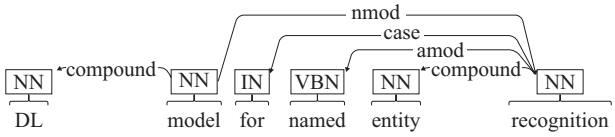


图 2 Head词汇识别结果示意图

（2）词汇到Root的依存路径。本文使用的依存路径是从目标词到句子Root词汇的最短路径[4]。对于词组，本文使用词汇组合的Header词到Root词的路径作为特征。

给定句子S，路径提取结果表示为（w1，p1:r:p2，w2）+，*+表示重复。w表示一个词，p是其词性标记，r表示w1到w2的句法依存关系。如果标注粒度为词组，且词组中存在多个Header词汇，则返回“<NO_PATH>”。

（3）Verb only ROOT依存路径。Verb only路径同（2）描述的结构一致，但在（2）生成的路径中去除所有非动词w文本。

（4）词汇直接关联的依存关系特征。给定目标词汇w，特征构造方法：①如果w与词汇t间存在dobj依存关系，w是支配词，则返回“dobj:t”，否则返回“dobj-r:t”；②如果w与词汇t间存在obj依存关系，参照①构造特征，但将特征中“dobj”替换为“obj”；③如果w不存在直接关联的obj或dobj依存关系，则对每一个与w存在依存关系的词汇t，构造从w到t的路径p。顺序拼接所有生成的p路径表示，返回拼接结果。

依存路径特征是一类非常重要的特征，但由于词汇（特别是动词）的多样性，上述方法构建的依存路径存在特征稀疏的问题。为此，本文使用动词的角色聚类类别替换路径中原始的动词词汇，以得到对提升模型效果更有帮助的特征。

3.4 基于Word2vec的动词角色聚类

句式如“this paper use <OBJ>”“<OBJ> is utilized”在词汇功能识别工作中特征明显。一旦找到这样的句式，学习模型可以马上确定<OBJ>是被使用的对象，更倾向于是方法而不是问题。这意味着，找到词汇通过动宾关系关联的动词将有利于提升模型标注效果。然而，同一个动作在文本中可能表现为多个词汇，直接使用原始的动词作为特征，会带来特征稀疏的问题。为此，需要对动词进行聚类。

本文动词聚类的目标是找出表示相同或者类似动作的动词词汇。这些词汇在词典中可能有不同的含义，但在学术文本的上下文中却扮演同样的角色。

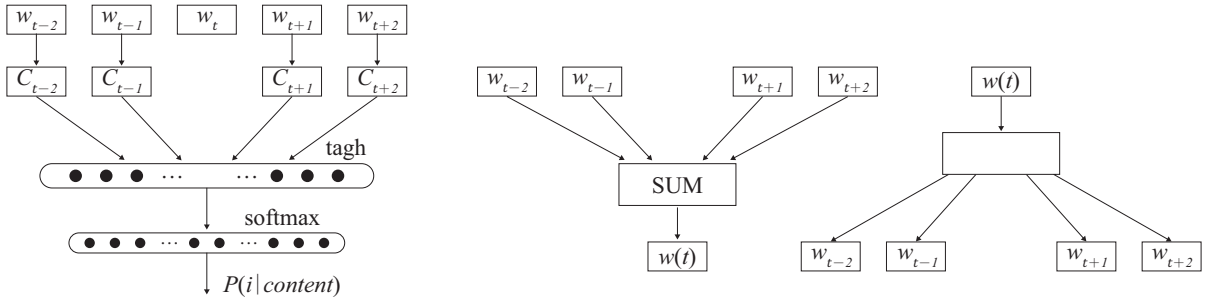
如表2所示，“present”和“propose”的原始含义不一样，但在特定的上下文里却扮演同样的语义角色，即“提出”一种基于SVM的方法；“use”和“employ”意义也有所不同，但在列出的文本中都表示“应用”。

表 2 动词角色示例

角 色	表 示
present	*This paperpresentsa* SVM based model
	This paperproposesa SVM based model
use	*This paperuses* SVM for ……
	weemploy SVM
	we utilized SVM

为识别动词的角色，一种可行的方法是编制词典。但是，人工完成这一个工作并不现实。首先，人工编制成本过高；其次，编制的词典很难具有领域通用性。因此，有必要探索自动化的动词角色聚类方法。基于深度学习的研究成果[11-13]，本文提出一种利用Word2vec聚类相同角色词汇的方法。

Word2vec是Mikolov等提出的一种利用深度学习思想学习词嵌入表示的工具[13]。Word2vec词嵌入模型以向量表示词汇，词汇间的语义相似性可通过向量距离加以衡量。Word2vec在模型实现上有两种主流结构，分别是CBOW模型和Skip-gram模型，如图3所示。本文使用Word2vec的CBOW模型训练词嵌入模型，词嵌入表示的向量维度设为100。



(1)NNLM　　(2)CBOW　　(3)Skip-gram

图 3 神经网络语言模型的三种模式

Word2vec模型衡量的是词在上下文的可替代性，这与LDA等主题模型有所区别[14-15]。如针对学术文本，Word2vec模型倾向于给“propose”和“present”一个较高的相似性得分，因为这两个词通常具有类似的上下文结构。

本文使用ACM计算机科学论文的摘要数据作为语料，对摘要进行句子切分，并使用StanfordPOS Tagger对句子文本进行词性标注，得到训练语料集[16-17]。本文使用的最终训练语料大小为258MB。表3给出一条语料数据的样例。基于语料，本文使用Word2vec工具进行词嵌入学习，在学习结果中过滤掉训练结果中的非动词词汇，得到动词词汇的词嵌入表示。

经过图4聚类算法，得到词汇类别649类，部分词汇及对应类别如图5所示。

上述聚类方案仅对高频词汇进行处理，存在大量未被聚类的动词。对这些词汇，本文使用Word2vec工具提供的聚类功能进行聚类，聚类数量设定为1 000。如果动词词汇不在高频词列表（频率≥2 000），则返回词汇在Word2vec原生聚类结果中的类别标签。

表 3 训练语料样例

We	present	a	SVM	based	method
We_PRP	present_VBP	a_DT	SVM_NN	based_VBN	method_NN

Algorithm 1 构建概念关系图的递归算法buildChunkGraphNode(G,word,node)

Input：并查集UT，阈值th，阈值maxCount，阈值maxSize，词汇initword，word关联词汇列表list（初始为空）
Onput：
1：将initword加入list
2：while list长度小于maxSize do
3：　for all词汇word in list do
4：　令array为同word余弦相似度大小排列的前maxsize个动词列表
5：　for i=0；i<array.length;i++ do
6：　令词汇temp为array的第i个元素
7：　if temp 同word的相似度小于th且list的长度大于1
8：　break；
9：　end if
10：　if如果temp未被加入UT then
11：　向UT中加入词对word和temp
12：　end if
13：　end for
14：　end for
15：　threshold+=0.1；
16：end while

图 4 词汇角色聚类算法

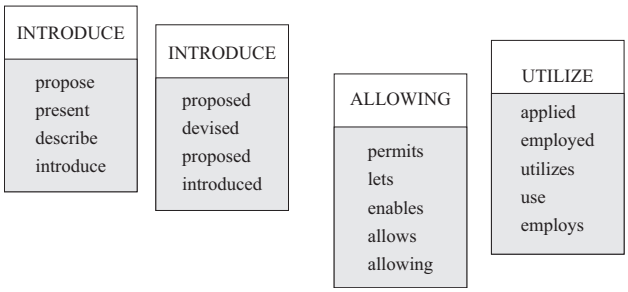


图 5 动词聚类效果示意图

4 对比方法

4.1 Gupta识别方法

Gupta[5]提出一种基于重抽样的标注方法。该方法试图抽取满足一定标准的句法模板，并将模板匹配到的词汇标注为“问题”或“方法”。如模板“employ（dobj xxx）”可识别出“we employ SVM for text categorization”中的方法词汇“SVM”。

Gupta等提出的方法从种子模板开始，利用设定的种子模板标注匹配到的文本，然后从匹配到的文本中学习新的模板[7]。重复“利用模板匹配文本-从文本中学习模板”这一步骤，直到完成标注工作。Gupta方法的关键是候选模板的重要性评分方法，评分超过一定阈值的模板将被用于后续的文本标注。该评分方法可简单表述为对类别（问题或者方法）C下的候选模板q，记q正确识别到词汇集合为p，q的得分 $socre\ (q) = \sum_{p} \in P \frac{1}{Z_p} count\ (p \in C)$，$Z_p$是p集合包含词汇在语料中出现的总频次[7]。

4.2 Tsai识别方法

Tsai方法同样基于重抽样策略，与Gupta方法在设计上非常类似[7]，差异在于Tsai方法从设定种子词开始[8]。Tsai方法利用种子词从语料中学习满足一定要求的匹配特征，然后用学习得到的匹配特征扩充种子词，重复这一过程，以不断扩充种子词和匹配特征集合，直到完成标注工作。

5 实验与讨论

5.1 实验数据

本文通过数据实验验证提出方法的效果。实验数据集为自建数据集和Gupta等提出的数据集[7]。

自建数据集的数据来源为ACM数据库收录的200篇计算机科学会议论文，对论文的标题和摘要数据进行人工标注，标注粒度为组块。自建数据集共包括1 002个句子，其中，标记为“问题”的名词组块为604个，标记为“方法”的名词组块为1 059个。Gupta数据集原始数据来源于ACL数据库，Gupta等标注了474篇文档的摘要和标题，标注粒度为单词[7]。Gupta构造的数据集有三个类别标记，本文仅使用其中的“问题”和“方法”两个类别。Gupta数据集中句子数为2 647个，其中，标记为“问题”的名词组块为3 851个，标记为“方法”的名词组块为4 042个。

5.2 实验设定

数据预处理包括使用OpenNLP进行句子切分，使用正向最大匹配算法及术语词典（包含131 917个术语）对文本进行术语识别，使用Stanford POS Tagger对文本进行词性标注，使用StanfordParser对句子进行句法分析。

在不同的数据集上，模型进行自动标注的粒度也不一样。自建数据集采用组块作为标注单元，而对Gupta数据集采用词汇粒度进行标注。因为自建数据集的人工标注粒度是组块，而Gupta数据集在单词粒度上进行人工标注。

本文使用CRF^{++}工具训练CRF模型[18]，规范化处理选用L2模式，cutf-off参数设为1，评测使用五折交叉检验方法，每次选取在训练集上最好的hyper-parameter参数用于测试集的效果测试，最后报告的结果为每一轮测试结果的平均值。评测指标为准确率、召回率和F1值。

5.3 实验结果及讨论

5.3.1 自建数据集中的实验效果

本文提出的方法在自建数据集上的评测结果见表4。从评测结果看，方法类词汇的识别效果在三个指标上都要优于问题类词汇的识别效果。准确率指标上，两个类别词汇识别效果类似，但在召回率上，问题类词汇的识别效果相对较低。

本文未试图在自建数据集上将提出的方法同参照

表 4 本文提出方法在自建数据集上的效果

类 别	召回率/%	准确率/%	F1
方法	0.584	0.667	0.622
问题	0.397	0.582	0.454

方法进行对比。Gupta和Tsai的方法都基于重抽样策略提出，在这种策略下，种子的选择以及参数的设定都将直接影响最终实验效果。本文自建数据集同Gupta和Tsai使用的评测数据属于不同的研究领域，因此，Gupta和Tsai给出的算法、种子以及参数设定不能直接应用。从这点看，在自建数据集上将本文的方法与Gupta和Tsai进行比较是没有意义的。

5.3.2 Gupta数据集中的实验效果

在Gupta数据集上进行评测的结果如表5所示。

表 5 在Gupta数据集上3种方法的评测效果

	方 法			问 题		
	准确率/%	召回率/%	F1	准确率/%	召回率/%	F1
Gupta	0.305	0.467	0.369	0.276	0.575	0.373
Tsaiet	0.482	0.488	0.485	0.440	0.473	0.456
本文方法	0.542	0.448	0.489	0.517	0.477	0.495

从评测结果看，本文提出的方法在各类别的F1指标和准确率指标的结果都最优。在召回率指标上，Tsai的方法效果最好，而Gupta在问题识别上取得最高的召回率。从整体看，本文提出的方法要优于Gupta和Tsai的方法。

本文提出的方法在方法类和问题类两个类别上取得了0.489和0.495的F1值，从绝对值来看，这一表现并不好。但需要说明的是，在这一数据集上，人工标注的一致性也仅为0.723。

6 结语

学术文本语义出版的发展对文本语义理解技术提出更高的要求。为辅助实现学术文本语义出版，本文提出一种基于序列标注思想的学术文本问题与方法功能的识别方法。实验表明，该方法具有优于当前最佳的识别效果。此外，本文还提出一种基于Word2vec的动词词汇角色聚类方法，能够将原本词义不同但在一定上下文环境下表现出同样功能的词汇聚类到一起。同时，通过对实验结果分析发现，学者在表述问题和方法时，对于词汇的组织和运用存在一定的共性。

本文仅着眼于问题和方法词汇的识别，而没有讨论词汇到底承担何种功能，是核心问题还是一般问题。因此，接下来的研究应进一步探索如何实现更加细分的问题和方法功能词汇的自动识别。另外，本文仅从识别技术的角度进行研究，如何更进一步将识别技术以及识别结果应用于语义出版，也需要后续更进一步的探索。

参考文献

[1] 王晓光,陈孝禹.语义出版的概念与形式[J].出版发行研究,2011(11):54-58.

[2] 王晓光,陈孝禹.语义出版:数字时代科学交流系统新模型[J].出版科学,2012(4):81-86.

[3] 苏静,曾建勋.国内外语义出版理论研究述评[J].中国科技期刊研究,2016,28(1):33-38.

[4] 程齐凯.学术文献词汇功能识别[D].武汉:武汉大学,2015.

[5] KONDO T,NANBA H,TAKEZAWA T,et al.Technical trend analysis by analyzing research papers' titles[M]//Human Language Technology.Challenges for Computer Science and Linguistics.[S.l.]:Springer Berlin Heidelberg,2009:512-521.

[6] NANBA H,KONDO T,TAKEZAWA T.Automatic creation of a technical trend map from research papers and patents[C]//International Workshop on Patent Information Retrieval.ACM,2010:11-16.

[7] GUPTA S,MANNING C.Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers[C]//Proceedings of ijcnlp.Thailand:The Association for computer Linguistics,2011:1-9.

[8] TSAI C T,KUNDU G,ROTH D.Concept-based analysis of scientific literature[C]//ACM International Conference on Information & Knowledge Management.[S.l.]:[s.n.],2013:1733-1738.

[9] TATEISI Y,SHIDAHARA Y,MIYAO Y,et al.Relation Annotation for Understanding Research Papers[C]//Linguistic Annotation Workshop and Interoperability with Discourse,2013:140-148.

[10] LAFFERTY D,MCCALLUM A,PEREIRA N.Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Eighteenth International Conference on Machine Learning.Morgan Kaufmann Publishers Inc,2001:282-289.

[11] HINTONG E,OSINDERO S,TEH Y.A fast learning algorithm for deep belief nets[J].Neural Computation,2006,18(7):1527-1554.

[12] BENGIO Y.Learning deep architectures for AI[J].Foundations and Trends Extregistered in Machine Learning,2009,2(1):1-127.

[13] MIKOLOV T,CHEN K,CORRADO G,et al.Efficient estimation of word representations in vector space[C].In Proceedings of Workshop at ICLR,2013:11-12.

[14] BLEID M,NG A Y,JORDAN M I.Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003(3):993-1022.

[15] BLEI D M.Probabilistic topic models[J].Communications of the ACM,2012,55(4):77-84.

[16] TOUTANOVA K,MANNING C D.Enriching the knowledge sources used in a maximum entropy part-of-speech tagger[C]//Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora:Held in Conjunction with theMeeting of the Association for Computational Linguistics.Association for Computational Linguistics,2000,25(6):63-70.

[17] TOUTANOVA K,KLEIN D,MANNING C D,et al.Feature-rich part-of-speech tagging with a cyclic dependency network[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics,2003:173-180.

[18] Taku-ku.CRF++:Yet another CRF toolkit[CP/DK].http://crfpp.sourceforge.net,2005.

作者简介

程齐凯，男，1989年生，博士，讲师，研究方向：自然语言处理、文本挖掘、信息检索，E-mail：chengqikai0806@gmail.com。
李信，男，1991年生，博士研究生，研究方向：大数据分析、语义计量、医学知识发现，E-mail：lucian@whu.edu.cn。

Automatic Recognition of Term Function in Academic Text for Semantic Publishing

CHENG QiKai[1,2], LI Xin[1,2]
(1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China)

Abstract: To enhance the development of semantic publishing of academic text, it is necessary to do more research on writing/publishing model and academic text understanding. Text understanding is a key technology for the semantic processing of academic text, especially stock academic text. This paper proposes a method for term function identification of academic text based on CRF model and term function analysis framework. Twenty-seven features (such as morphology features, syntax features, and chunk-based features) are employed in the sequence-labeling model. Experimental results show that the method obtains better results than the state of the art.

Keywords: Term Function; Semantic Publishing; Sequence Labeling; Academic Text

（收稿日期：2017-08-09）

书讯

《汉语主题词表》（工程技术卷）

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有的作用，曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要，2009年由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，完成《汉语主题词表（工程技术卷）》的重新编制工作。

全书共收录优选词19.6万条，非优选词16.4万条，等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用，《汉语主题词表》（工程技术卷）按专业分13个分册出版，同时建立《汉语主题词表》服务系统，提供在线概念检索和辅助标引服务，通过可视化技术展示各类概念关系，是图书馆、档案馆、出版社、期刊杂志社、文献信息中心等专业工作者及科研、教育及工程技术领域人员必备的参考书。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，全书2 300余万字，总定价3 880元，可分册购买。