利用小样本量机器学习实现学术文摘结构的自动识别*

白光祖[1,3] 何远标[2,3] 马建霞[1] 刘建华[2,3] 邹益民[4]
[1](中国科学院兰州文献情报中心 兰州 730000)
[2](中国科学院文献情报中心 北京 100190)
[3](中国科学院大学 北京 100049)
[4](浙江师范大学经济与管理学院 金华 321004)

摘要: 【**目的**】通过在小样本量下基于机器学习算法实现文摘语句的自动分类, 以此实现学术文摘结构的自动识别。【**方法**】设计多种学术文摘的文本表示特征, 利用自然语言处理技术实现特征的自动提取, 以此指导朴素贝叶斯、支持向量机模型进行训练, 并利用训练模型自动识别文摘结构。【**结果**】实验证明该方法较之于同类方法能够在较少训练语料下实现较好的识别准确率。【**局限**】由于文摘中"方法"类别语句缺乏固定的类别特征词与核心动词, 导致算法对该类别语句识别准确率较低。【**结论**】所提方法是一种小样本量情况下行之有效的学术文摘结构自动识别方法。

关键词: 学术文摘 结构识别 机器学习
分类号: G356.7

1 引 言

结构化学术文摘可通过研究目标(Objective)、研究方法(Method)、实验结果(Result)、研究结论(Conclusion)等结构化内容来简明扼要地归纳概括正文中相应部分内容信息[1], 其对于自动摘要、信息检索、信息抽取、自动问答系统等研究领域具有重要意义, 目前已被广泛应用到学术期刊出版领域[2]。但众多学术期刊数据库仍存有大量论文并未采用结构化文摘行文, 如若采用人工回溯改写将耗费大量人力物力。因此越来越多的研究尝试利用机器学习的方法来实现学术文摘结构的自动识别, 藉此来弥补这一缺憾。

2 相关研究

2.1 文摘结构表示模型

主流的文摘结构表示模型主要有基于分段名称(Section Names)的 S1 模型[3]、基于论证分区(Argumentative Zoning)的 S2 模型[4,5]以及基于核心科学概念(Core Scientific Concepts)的 S3 模型[6], 各个模型所包含的内容表示元素如表 1 所示。

其中, S1 模型是专为表示文摘结构而设计, S2、S3 模型由文献正文结构表示模型演化而来, 由于 S2、S3 模型表示元素过于庞杂而无法从文摘内容中建立一一对应关系, 因此在本研究中采用 S1 模型来表示文摘结构。

收稿日期: 2013-10-08
收修改稿日期: 2014-04-10
*本文系中国科学院西部之光联合学者项目"基于计算情报方法的甘肃省战略新兴产业技术创新竞争与发展研究"(项目编号:Y200201001)的研究成果之一。

表 1　S1、S2、S3 模型内容表示元素对照

表示元素	简写	元素含义	模型所含元素		
			S1	S2	S3
Hypothesis	HYP	研究假说			√
Motivation	MOT	研究动因			√
Background	BKG	研究背景		√	√
Goal	GOAL	研究目的			√
Objective	OBJ	研究目标	√	√	√
Experiment	EXP	实验细节			√
Model	MOD	理论模型、框架			√
Method	METH	研究方法、手段	√	√	√
Observation	OBS	实验中的数据、现象记录			√
Result	RES	主要研究结果	√	√	√
Conclusion	CON	分析、讨论以及主要结论	√	√	√
Related word	REL	与相关研究工作的对比		√	
Future word	FUT	下一步开展的工作		√	

2.2　文摘结构识别研究

在国外研究中, 用于文摘结构自动识别的机器学习算法有朴素贝叶斯模型(Naive Bayesian Model, NBM)、支持向量机模型(Support Vector Machines, SVM)等, 主要方法及其性能表现(训练数据来源均为 Medline 数据库)如表 2 所示:

表 2　已有的结构识别方法及其性能表现

方法出处	训练文摘量(单位: 篇)	学习模型	文摘结构表示模型	识别效果(F-scores)
Ruch 等[7] (2007)	10 800	NBM	OBJ - METH - RES - CON	85%
McKnight 等[8] (2003)	7 253	SVM	BKG - OBJ - METH - RES - CON	85.7%
Guo 等[9] (2010)	1 000	SVM	OBJ - METH - RES - CON	89%
Guo 等[9] (2010)	1 000	NBM	OBJ - METH - RES - CON	82%
Yamamoto 等[10] (2005)	8 383	SVM	BKG - OBJ - METH - RES - CON	87.98%

其中比较有代表性的是Guo等[9]与Yamamoto等[10]的研究。Guo等[9]基于上述三种文摘表示模型, 将文摘语句用词、二元词串以及句子上下文信息等特征来表示, 分别利用支持向量机、朴素贝叶斯算法实现了文摘语句的自动分类并对比了算法对各文摘表示模型的分类精度, 其中支持向量机对S1模型准确率达到89%, 朴素贝叶斯算法对S1模型准确率达到82%。Yamamoto等[10]也是利用支持向量机算法来表示文摘语句用词、核心动词、动词时态、语句位置信息、语句评分等特征。实验证明, 该方法准确率为87.98%, 特别是对于由背景(Background)与目的(Purpose)组成的引言(Introduction)语句的识别率高达91.3%。但对于文摘结构的自动识别工作来说, 某些领域结构化文摘数量偏小或难以获取, 而上述方法所需训练样本数量较大, 限制了其在这些领域的有效应用。

在国内相关研究中, 比较典型的有霍东云等[11]基于支持向量机算法, 通过医学主题词表——MeSH (Medical Subject Headings)分类体系标记Medline数据库文摘, 构建了较好的文摘文本分类模型。潘华山等[12]基于支持向量机算法, 选用句法和语义层面特征, 实现了对越语新闻文本的有效分类。

本研究在上述研究方法的基础上, 重新设计了文摘语句文本特征提取方法, 并以决策树算法为基准方法对朴素贝叶斯算法、支持向量机算法进行了实验测试与效果对比, 结果证明该方法在较小训练样本数量下对S1模型文摘结构识别效果较好, 因而具有较强的实践应用价值。

3　文摘结构识别方法

3.1　问题的提出与转化

学术文摘是非结构化的自由文本, 结构识别是为文本中每一个语句分配元素标签(如S1模型的OBJ/METH/RES/CON), 并将相同标签的语句组合成为一个语块, 用于表征某个文摘结构内容的过程。由此可将文摘结构自动识别问题转化为基于文摘语句的文本自动分类问题。

3.2　分类算法

本研究分别采用朴素贝叶斯算法[13]与支持向量机算法[13]来实现语句自动分类, 并以决策树算法[13]为基准方法(BaseLine Method)进行对比评测。

决策树算法是一种逼近离散函数值的方法。首先对训练数据进行处理, 利用归纳算法生成可读的规则和决策树, 然后使用决策树对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程。

朴素贝叶斯分类器基于独立假设、应用贝叶斯定

理与最大似然估计来为待分类特征数据集 F 选择具有最大后验概率[13]的类别 C:

$$\arg\max_C P(C|F) = \arg\max_C \frac{P(C)\cdot P(F|C)}{P(F)} = \arg\max_C P(C)\cdot P(F|C) = \arg\max_C P(C)\cdot \prod_{f\in F} P(f|C)$$

支持向量机算法能够在有限样本信息、模型的复杂性及学习能力之间寻求最优, 从而达到最好的推广能力。支持向量机的原理是通过事先选择的非线性映射将输入向量 x 映射到一个高维特征空间中, 在这个空间构造一个最优分类超平面, 将两类样本无错误地分开, 且要使两个类别(标记为$y\in\{-1,1\}$)的分类空隙最大, 如此便可以使在原始空间中非线性可分的问题变为高维空间中线性可分的问题, 从而达到分类的目的[14]。

3.3 特征提取

本研究将以非结构化文本形式存在的文摘转化为由一系列文本特征表示的数据集合, 用以描述和替代自由文本并指导机器学习算法进行模型训练和识别判断。研究选取的文本特征如下所示:

(1) 位　置

位置是指一个语句在文摘中所处的位置信息。通常情况下OBJECTIVE语句会出现在摘要的靠前部分, 而CONCLUSION语句会出现在摘要的靠后位置, 中间部分一般为METHOD语句和RESULT语句, 而METHOD语句通常又在RESULT语句之前。根据此类规律赋予每个语句一个位置属性值, 并将其作为一项分类特征。

(2) 类别词相似度

类别词是指对某一类别(OBJ/METH/RES/CON)表征显示度最高的一个特征词集。本研究采用χ^2统计度量特征项和类别之间的独立性(即相关度), 特征词对某类χ^2统计值越高, 它与该类之间的相关性越大, 表征的类别信息就越多, 提取特定类别中χ^2值最高的前若干个词组成类别词集合。

在此基础上, 将文摘语句进行预处理后形成语句词集合, 与每个类别的类别词集合在向量空间模型中做余弦相似度计算, 并将其得分分别作为一项分类特征。

(3) 核心动词

句子中核心的词性是动词, 动词是支配其他词性成分的中心成分, 而它本身却不受其他任何词性成分的支配, 并且与上下文环境无关, 所有受支配成分都以某种依存关系从属于动词词性成分[15]。本研究将文摘训练语料中的核心动词抽取出来并进行词频统计及排序处理后构成类别动词词表。将语句中的核心动词、词性、时态以及其是否在特定类别的动词词表中出现分别作为独立的特征项。同时, 本研究也将受支配成分与中心成分(即语句的主谓宾成分)抽取出来作为特征项之一。

(4) 上下文信息

文摘语句所处语境及其上下文线索信息对类别判定具有提示意义, 因此本研究将前一句类别信息、前一句核心动词、词性及其时态均作为分类特征项。

处理完的两条文本特征语料实例如表 3 所示。

4 实验及其评估

4.1 语料来源

1987年起, 部分生物学和医学期刊开始采用结构化文摘(Structured Abstracts)[16], 即作者需在其文摘中标示出能够表征文本结构的分段内容, 如Background、Objective、Method、Result、Conclusion等。本研究选取Medline数据库中经过标注的结构化文摘用作模型训练及结果测试语料。

4.2 语料处理

在语句特征提取时需要对语句进行分词、词干还原、词性标注等处理, 该步骤主要基于斯坦福大学的Stanford-CoreNLP[17]工具包来扩展编程实现。文本语句中核心动词及其词性的识别利用词法关系模式匹配工具 Tregex[18]配合 StanfordParser[19]来完成。语句中主谓宾成分的抽取利用依赖关系模式匹配工具 Semgrex[18]来完成。

4.3 模型训练

为与已有研究方法识别效果进行对比, 本研究采用两组语料分别进行训练测试。A组随机选取500篇文摘语料(含5 375个句子)作为训练语料, 500篇文摘语料(含5 289个句子)作为测试语料。B组随机选取1 390篇文摘语料(含15 110个句子)作为训练语料, 512篇文摘语料(含5 860个句子)作为测试语料。借助Waikato大学的开源机器学习软件Weka(Waikato Environment for Knowledge Analysis)[20]来完成模型训练过程, 分别利用决策树算法(J48)、朴素贝叶斯算法、支持向量机算法来对语料进行模型构建和识别测试。

表 3　语料实例

特征名称	含义	语料实例 1	语料实例 2
Location	该语句在文摘中位置	0.5	0.23
Double PreClass	前两个语句的类别	START	OBJECTIVE
PreClass	前一个语句的类别	OBJECTIVE	OBJECTIVE
SimilarityToO	与 OBJECTIVE 类别词相似度	0	0.02
SimilarityToM	与 METHOD 类别词相似度	0.01	0
SimilarityToR	与 RESULT 类别词相似度	0	0
SimilarityToC	与 CONCLUSION 类别词相似度	0	0.01
DoublePreVbWord	前两个语句的核心动词	null	do
PreVbWord	前一个语句的核心动词	do	be
VbWord	该语句的核心动词	receive	be
VbBelongO	核心动词是否属于 OBJECTIVE 类别动词表	no	no
VbBelongM	核心动词是否属于 METHOD 类别动词表	yes	no
VbBelongR	核心动词是否属于 RESULT 类别动词表	yes	no
VbBelongC	核心动词是否属于 CONCLUSION 类别动词表	no	no
PreVbPos	前一语句核心动词的词性	VBZ(现在时)	VBD(过去式)
VbPos	该语句核心动词的词性	VBZ(现在时)	VBZ(现在时)
PreVbVoice	前一语句核心动词的时态	present	past
VbVoice	该语句核心动词的时态	present	present
S	该语句的主语成分	patient	OCA
P	该语句的谓语成分	receive	be
O	该语句的宾语成分	test	group
Class	该语句的类别	METHOD	OBJECTIVE

4.4　评价方法

本研究采用开放测试方法, 以决策树算法为基准方法, 对朴素贝叶斯、支持向量机算法识别结果进行评价, 用正确率(Precision)、召回率(Recall)以及F-测度值(F-measure)指标进行结果衡量[21]。以OBJ类为例:

$$正确率=\frac{被正确识别为OBJ的语句数量}{所有被识别为OBJ的语句数量}$$

$$召回率=\frac{被正确识别为OBJ的语句数量}{所有OBJ类语句数量}$$

$$F\text{-}测度值=\frac{正确率\times召回率\times 2}{正确率+召回率}$$

4.5　实验结果与分析

(1) 算法对比

如表4所示, 从A组实验测试结果可以看出, 支持向量机算法准确率(90.7%)略高于朴素贝叶斯算法(88.9%), 它们均优于基准方法——决策树算法的分类准确率(86.3%)。算法正确率、召回率由高到低排列为支持向量机(90.7%, 90.7%)、朴素贝叶斯算法(89.2%, 88.9%)、决策树算法(87.0%, 86.2%)。

表 4　A 组实验测试结果

分类算法	类别标签 (Class)	正确率 (Precision)	召回率 (Recall)	F-measure
决策树算法 BaseLine	OBJ	0.948	0.793	0.864
	METH	0.77	0.901	0.831
	RES	0.88	0.858	0.869
	CON	0.94	0.869	0.903
均值		0.87	0.862	0.863
朴素贝叶斯算法	OBJ	0.966	0.878	0.92
	METH	0.844	0.903	0.872
	RES	0.872	0.903	0.887
	CON	0.946	0.846	0.893
均值		0.892	0.889	0.889
支持向量机算法	OBJ	0.937	0.928	0.932
	METH	0.868	0.916	0.892
	RES	0.916	0.89	0.903
	CON	0.925	0.904	0.914
均值		0.907	0.907	0.907

从识别效率来看，支持向量机算法建模用时301.22秒，测试数据用时4.62秒。朴素贝叶斯算法建模用时0.1秒，测试数据用时0.65秒。决策树算法建模用时1.2秒，测试数据用时0.27秒。其中，支持向量机算法建模耗时最多，决策树算法次之，朴素贝叶斯算法最少。

表5　B组实验测试结果

分类算法	类别标签(Class)	正确率(Precision)	召回率(Recall)	F-measure
决策树算法BaseLine	OBJ	0.840	0.919	0.878
	METH	0.782	0.817	0.799
	RES	0.907	0.873	0.890
	CON	0.950	0.861	0.903
均值		0.869	0.866	0.867
朴素贝叶斯算法	OBJ	0.930	0.912	0.921
	METH	0.826	0.899	0.861
	RES	0.918	0.891	0.904
	CON	0.926	0.887	0.906
均值		0.899	0.897	0.897
支持向量机算法	OBJ	0.917	0.921	0.919
	METH	0.849	0.87	0.86
	RES	0.918	0.903	0.91
	CON	0.913	0.91	0.911
均值		0.90	0.90	0.90

如表5所示，从B组实验测试结果可以看出，支持向量机算法准确率(90.0%)略高于朴素贝叶斯算法(89.7%)，它们均优于基准方法——决策树算法的分类准确率(86.7%)。算法正确率、召回率由高到低排列为支持向量机(90.0%，90.0%)、朴素贝叶斯算法(89.9%，89.7%)、决策树算法(86.9%，86.6%)。

从识别效率来看，支持向量机算法建模用时1 410秒，测试数据用时4.62秒。朴素贝叶斯算法建模用时0.05秒，测试数据用时0.19秒。决策树算法建模用时2.12秒，测试数据用时0.07秒。其中，支持向量机算法建模耗时最多，决策树算法次之，朴素贝叶斯算法最少。

对比A、B组实验结果，从识别准确率来看，在相同数量的训练、测试语料下，支持向量机算法识别准确率略高于朴素贝叶斯算法；但从识别效率来看，支持向量机算法建模时间开销要远大于朴素贝叶斯算法，因此在实际工作中应根据具体需求考虑选用合适的识别算法。

支持向量机算法与朴素贝叶斯算法对结构元素的识别准确率由高到低排序均为OBJ>CON>RES>METH。其中，对于OBJECTIVE类别语句识别准确率均达到91%以上，主要是由于OBJECTIVE类别特征词与核心动词较多，特征较为显著。而METHOD类别语句含有大量工具、指标、单位、试剂等类型专用名词，缺乏较为固定的类别特征词与核心动词，因此算法较难识别。典型示例如表6所示，该条语句的正确类别应为METHOD，但由于其与METHOD类别特征词集相似度(SimilarityToM)为零且语句核心动词不属于METHOD类别核心动词集(VbBelongM值为no)，模型将该语句类别误判为RESULT。

表6　识别错误的语料示例

Location	Double PreClass	PreClass	SimilarityToO	SimilarityToM	SimilarityToR	SimilarityToC
0.31	METHOD	METHOD	0.0	0.0	0.0	0.0
DoublePreVbWord	PreVbWord	VbWord	VbBelongO	VbBelongM	VbBelongR	VbBelongC
be	be	be	no	no	no	no
PreVbPos	VbPos	PreVbVoice	VbVoice	S	P	O
VBD	VBD	PAST	PAST	reason	be	withdrawal

(2) 效果对比

虽然均采用Medline文摘作为样本数据，但由于测试集合和测试条件的差异，表7中指标数值只能作为方法效果的参考。较之于文献中识别效果最优的方法[7,9]，本研究提出的方法在较少训练样本量下即可实现较高准确率，表明本方法是一种小样本量情况下行之有效的学术文摘结构自动识别方法。

表 7　与同类方法识别准确率对比

算法名称	文献中识别效果最优的方法				本文提出的方法		
	方法出处	训练样本数量(篇)	测试方法	平均准确率	训练样本数量(篇)	测试方法	平均准确率
支持向量机	Guo 等[9] (2010)	1 000	十折验证	89.0%	500	开放测试	90.7%
朴素贝叶斯	Ruch 等[7] (2007)	10 800	开放测试	85.0%	1 390	开放测试	89.7%

5　结　语

本研究基于自然语言处理方法设计抽取了学术文摘文本表示特征，并利用机器学习算法实现了文摘结构的自动识别方法。实验结果表明本方法在较小规模训练样本下取得了较好的识别效果。下一步研究将通过限定方法应用的特定领域来提升训练语料的主题集中度，预计会得到更好的识别效果。

参考文献：

[1] U.S. National Library of Medicine. Structured Abstracts [EB/OL]. [2014-04-01]. http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html.

[2] 张晓林, 彭希珺. 用高水平学术规范保障论文学术质量[J]. 现代图书情报技术, 2014(1): 1-3. (Zhang Xiaolin, Peng Xijun. Secure the Quality of Academic Papers by High-level Academic Norms [J]. New Technology of Library and Information Service, 2014(1): 1-3.)

[3] Hirohata K, Okazaki N, Ananiadou S, et al. Identifying Sections in Scientific Abstracts Using Conditional Random Fields [C]. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08). 2008: 381-388.

[4] Teufel S, Siddharthan A, Batchelor C. Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics [C]. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09). Stroudsburg: Association for Computational Linguistics, 2009: 1493-1502.

[5] Mizuta Y, Korhonen A, Mullen T, et al. Zone Analysis in Biology Articles as a Basis for Information Extraction [J]. International Journal of Medical Informatics, 2006, 75(6): 468-487.

[6] Liakata M, Teufel S, Siddharthan A, et al. Corpora for the Conceptualisation and Zoning of Scientific Papers [C]. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. 2010:2054-2061.

[7] Ruch P, Boyer C, Chichester C, et al. Using Argumentation to Extract Key Sentences from Biomedical Abstracts [J]. International Journal of Medical Informatics, 2007, 76(2-3): 195-200.

[8] McKnight L, Srinivasan P. Categorization of Sentence Types in Medical Abstracts [C]. In: Proceedings of the 17th Annual Symposium of the American Medical Informatics Association. 2003: 440-444.

[9] Guo Y, Korhonen A, Liakata M, et al. Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes [C]. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. 2010: 99-107.

[10] Yamamoto Y, Takagi T. A Sentence Classification System for Multi-document Summarization in the Biomedical Domain [C]. In: Proceedings of the International Workshop on Biomedical Data Engineering (BMDE'05). 2005: 90-95.

[11] 霍东云, 聂峰光, 郭力. 利用 Medline 文摘数据库研究文本分类[J]. 计算机与应用化学, 2007, 24(9): 1281-1284. (Huo Dongyun, Nie Fengguang, Guo Li. Text Categorization Research by Using Medline Database [J]. Computers and Applied Chemistry, 2007, 24(9): 1281-1284.)

[12] 潘华山, 严馨, 余正涛, 等. 基于支持向量机的越语新闻文本分类方法[J]. 山西大学学报：自然科学版, 2013, 36(4): 505-509. (Pan Huashan, Yan Xin, Yu Zhengtao, et al. Vietnamese News Text Classification Method Based on Support Vector Machine [J]. Journal of Shanxi University: Natural Science Edition, 2013, 36(4): 505-509.)

[13] Alpaydin E. 机器学习导论[M]. 范明, 昝红英, 牛常勇译. 北京：机械工业出版社, 2009. (Alpaydin E. Introduction to Machine Learning [M]. Translated by Fan Ming, Zan Hong ying, Niu Changyong. Beijing: China Machine Press, 2009.)

[14] Hsu C W, Lin C J.A Comparison of Methods for Multiclass Support Vector Machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[15] 张艳. 汉语句法分析的理论、方法的研究及其应用[D]. 北京：中国科学院自动化研究所, 2003. (Zhang Yan. Research on Theory and Methods of Chinese Syntactic Parsing and Application [D]. Beijing: Institute of Automation, Chinese Academy of Sciences, 2003.)

[16] Huth E J. Structured Abstracts for Papers Reporting Clinical Trials [J]. Annals of Internal Medicine, 1987, 106(4): 626-627.

[17] The Stanford Natural Language Processing Group. Stanford

CoreNLP [EB/OL]. [2014-04-01]. http://nlp.stanford.edu/software/corenlp.shtml.

[18] The Stanford Natural Language Processing Group. Tregex, Tsurgeon and Semgrex [EB/OL]. [2014-04-01]. http://nlp.stanford.edu/software/tregex.shtml.

[19] The Stanford Natural Language Processing Group. The Stanford Parser: A Statistical Parser [EB/OL]. [2014-04-01]. http://nlp.stanford.edu/software/lex-parser.shtml.

[20] WEKA. Weka3: Data Mining Software in Java [EB/OL]. [2014-04-01]. http://www.cs.waikato.ac.nz/ml/weka/.

[21] Yang Y. An Evaluation of Statistical Approaches to Text Categorization [J]. Information Retrieval, 1999, 1(1-2): 69-90.

作者贡献声明：

白光祖：研究命题的提出、研究过程的实施、主要数据的获取与分析处理、论文的起草与修订；
白光祖，何远标：研究方案的设计；
马建霞，刘建华，邹益民：实施方案的设计。

(通讯作者：白光祖 E-mail: baigz@llas.ac.cn)

Application of Machine Learning with Limited Corpus to Identify Structure of Scientific Abstracts Automatically

Bai Guangzu[1,3] He Yuanbiao[2,3] Ma Jianxia[1] Liu Jianhua[2,3] Zou Yimin[4]
[1](Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000, China)
[2](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)
[4](College of Economics & Management, Zhejiang Normal University, Jinhua 321004, China)

Abstract: [**Objective**] This study aims to identify structural contents of scientific abstract automatically by classifying the academic abstracts sentences based on machine learning with limited samples. [**Methods**] This paper designs a variety of text features to represent scientific abstract sentences, then extracts these features from the academic abstracts based on natural language processing techniques so as to instruct Naive Bayesian Model and Support Vector Machines in training, and ultimately identifies the structure of academic abstracts automatically by using these models. [**Results**] Experiments show that the method can achieve fairly even better recognition accuracy compared with previous methods by using less training corpus. [**Limitations**] Due to the lack of feature words and core verbs in abstract sentences with "METHOD" class label, it resulted in a lower recognition accuracy on these sentences. [**Conclusions**] This method is an effective approach to achieve the automatic recognition of academic abstracts structure by using limited corpus.

Keywords: Science abstract Structure identifying Machine learning