面向领域科技文献的句子级创新点抽取研究*

张　帆[1,2]　乐小虬[1]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【**目的**】抽取领域科技文献中句子级创新点。【**方法**】面向文献中的句子, 以领域词表和本体中的关系为基础构建识别规则, 采用基于主题词重叠度的冗余度计算方法过滤创新点候选集。【**结果**】选取肿瘤领域的数据集进行实验, 抽取结果的准确率为 89.42%, 召回率为 60.14%。【**局限**】规则有待进一步完善, 提高召回率。【**结论**】利用领域词表和本体中的关系能有效地抽取科技文献中的句子级创新点。

关键词: 科技文献　语言学特征　结构式摘要　创新点抽取　冗余度计算
分类号: TP393

1　引　言

创新点是科学论文的灵魂[1], 可以直接体现一篇论文的学术价值和发表价值[2]。从知识论(Theory of Knowledge)角度进行界定, 论文创新点的表现形式为论文中作者使用的“知识主张”(Knowledge Claim)[3-4]。科技文献旨在为同一问题的其他研究者提供新知识[5-7], 因此作者写作时会采用特定描述方式声明其首创性以及创新性。论文作者使用的“新知识主张语句”为读者提供所做研究的新知识, 可以揭示论文的创新点。

抽取领域文档集中的论文创新点, 能有效揭示领域的研究进展, 可以帮助研究者快速发现和评估领域内各研究方向出现的新技术、新方法。本文以肿瘤领域的期刊文献、主题词表和知识本体为基础, 旨在探索利用领域词表和本体抽取科技文献中的句子级创新点的有效方法, 为研究领域创新点深层次挖掘和利用提供途径。

2　创新点抽取方法研究概况

识别科技文献中的具体创新点隶属于句子级知识抽取的研究范畴。近年来, 关于知识抽取的研究引起了国内外研究者的广泛关注, 研究对象从实体或实体之间的关系等短文本[8-9]扩展到可以揭示更完整语义的句子级。目前, 抽取创新点采用的主要方法有基于语言学特征的方法、基于本体或词表的方法以及基于句子分类的方法三种:

(1) 基于语言学特征的方法。其核心思想为通过分析创新点的语言特征, 选择句子的特征项进行抽取或制定相应规则抽取。文献[10-11]在构建知识元时根据语言特点定位描述创新点知识元的句子, 进而获取句子中的特征词(实词)作为抽取特征项; 文献[12-15]采用基于规则的方法抽取科技文献中的创新研究内容、性能参数、概念及其学术定义等重要信息。该方法的不足为: 完善规则的制定需要语言学专家的参与; 特征的选取和规则的制定无法覆盖抽取目标的所有语言学现象。

(2) 基于本体或词表抽取的方法。通过词表或本体中实体之间的关联可以发现潜在的新知识(如已知关系 A 影响 B, B 引发 C, 则可以推理出潜在的新关系

收稿日期: 2014-05-14
收修改稿日期: 2014-06-23
*本文系国家科技支撑计划子课题“基于文献知识网络的领域学术关系研究与示范”(项目编号: 2011BAH10B06-04)的研究成果之一。

A 影响 C)[16-18]。该方法对实体间关系的标注主要依据实体在句子中的共现关系，新知识通过文献中现存实体及其关联推理得出，其可靠性需进一步证实。Chowdhury 等[17]从用户的检索式中提取假说，并利用本体标注检索式和句子得到实体及其之间的关联，将假说中的实体和关联与现存的实体和关联进行匹配，通过统计学的方法对该假说进行验证；Cohen 等[18]利用 SemRep 标注文本中的实体和关系，借助基于谓语的语义索引(PSI)快速推断药物与疾病的可能关系。该方法领域针对性强，但偏向词表或本体中存在的概念，对本体中没有的新概念、新术语的揭示能力较弱。

(3) 基于句子分类的方法。其核心思想是将创新点抽取问题转化为分类问题：依据 Schema 中体现创新点的类别(如作者新贡献)选择句子的分类特征，利用标注集训练分类器，最后利用分类器识别句子所属类别。可以选取词频、句长、动词特征、元话语特征、线索词等[19-20]作为分类特征。其中，文献[19, 21-22]将作者所做的新贡献(OWN)作为基础标注 Schema 之一进行标注，获得关于作者新贡献的相关句子；Huang 等[20]依据“PICO”的 Schema 抽取医学文献中关于病人(问题)、干预、对比和结果的句子，在此基础上，Demner-Fushman 等[23]主要抽取结果相关的句子，并进一步细化 Schema。该方法劣势在于有监督的方法(决策树等)需要人工参与完成对训练集的标注，分类器的分类效果受标注结果的影响显著。

综合以上研究，本文借助领域词表对文本进行标注，运用基于规则的方法对描述创新点的句子进行准确定位，并考虑结构式摘要等其他特征对规则进行优化，提高召回率。此外，采用基于领域主题词重叠度的计算方法判断所抽取创新点的有效性。

3 论文创新点抽取方法

科技文献中的“知识主张”可以体现论文的创新点，针对科技文献的语言特征和体裁特征，利用基于规则的抽取方法可以准确识别论文中的“知识主张”并且避免人工标注训练集的工作。此外，在以领域文档集为抽取对象时，对抽取结果进行过滤，去除冗余的创新点，进一步保证抽取结果的准确性。

本文提出的创新点抽取方法的处理流程主要包括文档集标注、候选创新点抽取以及冗余创新点过滤三个部分，如图 1 所示。首先，选择论文摘要作为抽取对象，分句后利用领域词表和引导词对文档集进行标注，得到标注语句集；其次，在基于领域词表标注的基础上，依据创新点句子的语言学特征制定抽取规则，并从不同角度对规则进行优化和扩展，将符合规则的语句抽取出来得到候选创新点句子集；最后，利用新颖探测中重叠度的计算方法结合领域词表去除创新点句子集中的冗余，得到最终创新点句子集。

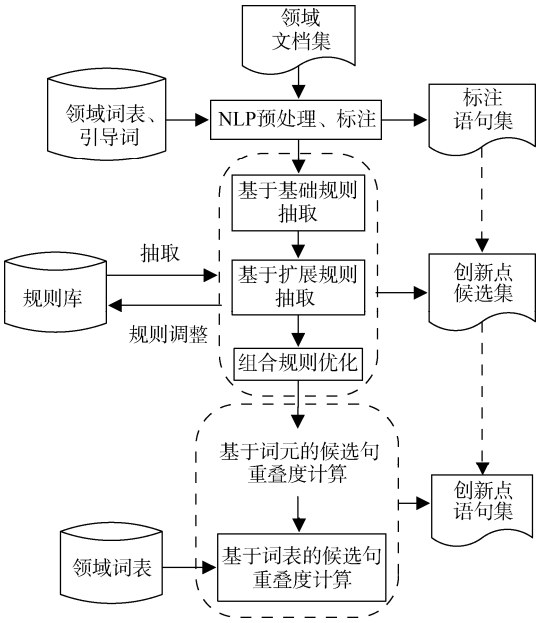

图 1 创新点抽取处理流程

3.1 文档集语义标注

论文创新点的特征主要分为分布特征和语言学特征两类。文献[2,11]对论文创新点分布特征进行总结，认为论文摘要、引言、结论等部分可以集中体现创新点。创新点语言特征主要体现在特征词(引导词)和常用表达方式两方面[3,6,7,24]。本文选取传统摘要以及结构式摘要作为研究对象，并针对结构式摘要的特征，制定相应抽取策略改善抽取效果。

预处理过程主要包括分句、词元化、去除停用词以及语义标注。本文结合引导词和领域词表对句子集进行标注，借助领域词表中同义词、近义词等可以最大限度揭示句子研究主题。参考 Trine [6]和 Parkinson[7]的研究成果，选取创新点的语言学特征引导词类型为以下 7 类，如表 1 所示。

表 1　创新点语言学标注引导词

类型	标注符号	词例
指代作者	RF	I, We, Our
指代文章	TP	this paper, this article, this report, our study 等
标志性动词	VB	find, reveal, illustrate, suggest, argue, indicate 等
标志性名词	NN	insight, finding, analysis, investigation 等
标志性形容词	AD	new, novel, unused 等
表明目的	OB	aim, objective, purpose, goal 等
主题词、关键词	TW	cytotoxic chemotherapy, antitumor 等

最终得到标注完成的例句如下:

NVP-ADW742 monotherapy or its #combination#TW with #cytotoxic chemotherapy#TW had significant #antitumor#TW #activity #TW in an #orthotopic#TW #xenograft#TW MM #model#NN, providing #in vivo#TW proof of principle for #therapeutic#TW #use# VB of selective IGF-1R inhibitors in cancer.

3.2　候选创新点识别

本文采用基于规则的方法抽取科技文献创新点候选集, 抽取方法主要包括三个部分: 规则的构建、基于规则的抽取以及规则的组合与优化。在构建规则时, 利用创新点句子的语言特征以及领域主题词表初步制定规则; 在此基础上, 通过补充引导词和针对特定抽取对象制定补充规则两种方法, 对基础规则进行补充和调整; 最后, 将基础规则和不同补充规则结合使用, 并通过实验验证抽取效果, 选取最佳规则组合。

抽取时将标注后语句中的标注符号分离出来, 形成以空格相隔的标注符号序列(如例句中标注序列为: TW TW TW TW TW TW NN TW TW VB), 规则构建的重点是将主题词表引入规则构建的过程中, 领域主题词可以揭示该领域的研究重点, 使抽取结果与领域研究主题密切相关, 提高抽取结果的准确度; 此外, 构建规则时综合考虑不同类别线索词出现的频率、位置以及不同线索词之间的组合关系, 并对某些规则设定优先匹配; 基于领域词表的抽取规则并未过多涉及复杂的句法关系, 因此可以避免一些规则的交叉和矛盾, 简化规则制定过程; 最后, 采用正则表达式书写如下:

(TW){1,}((VB)|(NN))(TW){0,1}(([A-Z]{2}){0,})(TW){0,}((RF)|(TP)|(AD)|(OB))(TW){0,}

采用基本规则对整个文档集中的标注语句进行匹配, 初步得到创新点候选集。仅使用基础规则会造成有价值创新点的遗漏, 笔者从多个角度对规则进行补充和调整。其主要思路包括以下两点: 通过分析文献, 扩充引导词, 达到对规则进行优化的效果; 从不同角度对原始句子集进行过滤, 将过滤后的句子作为抽取对象, 制定新的抽取规则进行抽取。利用过滤句子的思路对规则进行优化主要包括以下三个方面:

(1) 利用结构式摘要的体裁特征

结构式摘要的提出和使用源于医学领域期刊[25], 与传统摘要相比, 结构式摘要通常将论文的主要部分(背景、目的、方法、结果、结论等)精要、清晰地罗列出来, 为读者提供关于论文内容具体、直观的概述。通过结构式摘要可以准确定位论文中描述方法、材料和实验部分, 而这些部分通常可以体现作者研究中所采用的创新方法和实验步骤。进行规则扩展时, 以结构式摘要中方法、材料和实验(Method、Material、Experimental Design)部分作为抽取对象, 抽取基础规则未发现的创新点句子, 补充规则如下:

((TW)(([A-Z]{2}){0,})((VB)|(NN)))

(2) 利用新词特征

句子中出现新词的个数(New Word Count)可以作为衡量该句子新颖性的指标之一, 该指标在 TREC 2002 新颖探测任务中具有良好的鲁棒性[26], 尽管新词不能完全代表创新性, 但新词很可能揭示论文作者所采用的新材料、提出的新概念、新指标以及做出的新发现等。处理方法为: 对句子进行词性标注, 提取每个句子中的名词; 以年为间隔, 将当前句子出现年份之前所有句子作为历史集, 探测当前句子中出现的新词; 将所有包含新词的句子作为抽取对象, 采用补充规则进行抽取。

(3) 利用标题–主题词特征

论文标题是对整篇论文研究工作的高度概括, 论文标题中包含的词或术语可以揭示研究主题, 而论文中的创新点通常与论文主题相关。因此, 笔者利用领域词表以及关键词词表对论文标题及句子进行处理, 旨在获取与主题相关的语句作为抽取对象。

3.3　冗余创新点过滤

判断文本流中信息的新颖性时, 可以从信息的冗余度入手, 如果文本的冗余度越小则新颖度越高。常用的句子冗余度计算方法包括: 基于几何距离计算

(余弦距离、Manhattan 距离等)[26-27]，词差集计算(Set Difference)[26,28]以及词重叠度计算(Overlap)[29]等。为保证候选创新点的有效性(新颖性)，本文通过计算创新点之间的冗余度对候选创新点进行过滤，该计算方法将句子视为词的集合，两个句子共同包含的词数越多，则两个句子之间信息冗余的可能性越高。仅利用分词词元计算冗余度没有考虑到不同词对句子的贡献不同，在此基础上引入基于主题词表的冗余度计算，避免错误过滤表述相似但包含不同重要信息的句子。以Zhang等[29]提出的词重叠度计算方法为基础，分别计算句子间的词元重叠度和主题词重叠度，公式如下:

$$\mathrm{Overlap}B_{A_i} = \max\{\frac{A_i \cap B}{B}\} \quad (1)$$

其中，A_i为历史集中句子的词元/主题词向量，B 为当前句子的词元/主题词向量。

首先以分词词元为匹配单位，采用公式(1)的计算方法，将当前句与历史句(出现时间先于当前句的句子)共同包含词元的个数与当前句包含词元个数的比值作为重叠度，选取重叠度的最大值作为当前句的冗余度，将冗余度超过阈值的当前句视为候选冗余句子，阈值可通过实验得到或依据经验设定。

其次，计算得到候选冗余句子对的主题词冗余度，利用领域词表标注句子中的主题词，采取公式(1)计算二者之间的主题词冗余度，将主题词重叠度超过阈值的创新点句子加入冗余句子集，最后将冗余句子统一从候选集中删除，得到最终的创新点句子集。

4 实验结果与分析

4.1 实验数据来源与预处理

实验数据来源于 Web of Knowledge 中 20 种肿瘤领域期刊的文章摘要，将实验数据存入 MySQL 数据库中，使用 LingPipe 自然语言处理软件包[30]进行分句处理，采用 Stanford 自然语言处理工具[31]进行分词和词性标注。编写 Java 程序对文档集进行标注，标注过程中引入关键词词表以及领域术语词表 NCIt(National Cancer Institute thesaurus)[32]。NCIt 由美国国立癌症研究所(NCI)编制，可以提供肿瘤领域相关主题词的上位词、下位词以及同义词。

选取 NCIt 中“Childhood Neoplasm”及其所有下位词作为检索词，对数据集进行检索，选取前 100 篇论文摘要进行手工标注，得到创新点句子 281 个。采用信息检索中的准确率与召回率对抽取结果进行测评，利用剩余创新点句子优化、调整规则以及作为冗余创新点过滤时的历史句子集。

4.2 实验结果与分析

论文创新点的抽取实验主要分为两个部分: 基于规则的方法识别摘要中的创新点句子，以及基于重叠度计算的方法去除创新点候选集中的冗余。

(1) 候选创新点识别

该部分实验重点在于通过对比不同规则优化组合策略的抽取结果，选择最佳组合抽取策略。首先对单独使用基础规则以及三种补充规则策略的实验效果进行测评，并通过对抽取结果的抽样分析对 3.1 节中引导词进行补充，进而完善标注集的标注效果，得到基本规则和各种优化策略的抽取结果如表 2 所示:

表 2 规则单独使用时实验结果

规则策略	候选集			最终结果集		
	准确率	召回率	F 值	准确率	召回率	F 值
基本规则	80.00%	46.98%	59.19%	90.54%	47.69%	62.47%
基于结构式摘要补充规则	77.78%	14.95%	25.07%	87.50%	14.95%	25.53%
基于新词补充规则	51.79%	10.32%	17.21%	55.56%	10.68%	17.91%
基于标题–主题词补充规则	53.60%	47.33%	50.28%	56.54%	47.69%	51.74%

表 2 结果表明，补充规则中基于结构式摘要优化的规则准确率较高，与基础规则并无较大差异，而召回率很低，其原因为具有结构式摘要的论文数量仅占全部论文数量的 30%，因此会造成召回率的损失；而基于新词和基于标题–主题词补充规则的准确率在50%左右，抽取结果中非创新点句子含量过高，抽取效果不理想，其原因可能为: 医学论文标题通常与主题句相似，揭示论文的主题或立意[33]，但不等同于揭示论文创新点的特征词。

针对单独使用一种规则策略的不足，本文通过实验测评不同类型规则组合后的抽取效果，如表 3 所示。表 3 结果显示，对比单独使用基本规则，将基本规则与补充规则组合使用时召回率都有不同幅度的提升，表明文中三种优化策略对原始抽取结果具有一定改善效果。其中，基本规则和结构式摘要补充规则的组合使用获得的准确率和召回率都较高，抽取效果最好，

说明借助结构式摘要的特征抽取创新点具有可行性; 加入标题–主题词补充规则的组合策略获取的召回率较高, 这是由于摘要语言精简, 并且一般都围绕论文主题展开, 所以在摘要句子中主题词覆盖范围广, 因此采用相同的补充规则抽取会造成准确率的下降; 基于新词补充规则与基本规则的组合策略对召回率的改善结果不够理想, 同时降低抽取结果的准确率, 应考虑和其他策略综合使用。考虑到新词优化与标题–主题词优化方法分开使用的综合效果相近(F 值接近), 但在准确率和召回率各有侧重, 因此考虑将二者的结果取交集, 再与其他策略取并集使用, 即将同时满足新词规则和标题–主题词规则的抽取结果, 加入满足基本规则和结构式摘要规则的结果之中(见表 3 最后一行)。

表 3　组合规则策略抽取结果

组合规则策略	候选集			最终结果集		
	准确率	召回率	F 值	准确率	召回率	F 值
基本规则+结构式摘要优化	80.29%	59.43%	68.30%	89.42%	60.14%	71.91%
基本规则+新词优化	71.78%	51.60%	60.04%	79.46%	52.31%	63.09%
基本规则+标题–主题词优化	56.69%	63.34%	59.83%	61.02%	64.06%	62.50%
基本规则+结构式摘要优化+(新词优化∩标题–主题词优化)	74.89%	61.56%	67.57%	82.55%	62.28%	70.99%

(2) 冗余创新点过滤

采用3.3节所述的冗余度计算方法去除重复的创新点。以整个文档集为实验对象, 选取 5 年为区分当前文本与历史文本的时间间隔, 设定第一次冗余度过滤的阈值为 0.6, 对超过阈值的创新点句子进行二次过滤, 设定阈值为 0.6。通过对比表 2 与表 3 中候选集与最终结果集对应指标可以发现: 经过冗余去除后的准确率、召回率以及 F 值都有所提升。通过对比当前句子与历史句子的词重叠度以及主题词重叠度, 可以有效地去除候选集中冗余的创新点, 进一步优化抽取结果。

5　结　语

本文利用领域词表和本体, 通过分析科技文献中创新点的体裁特征和语言特征, 提出句子级创新点识别规则及其组合策略, 通过实验选取最优组合规则, 并利用主题词重叠度的句子冗余度计算方法过滤冗余创新点。该方法在肿瘤领域抽取结果的准确率为 89.42%, 召回率为 60.14%。

本文的不足之处在于抽取结果存在描述不够细致、具体, 无法准确识别创新点类型等问题, 考虑将抽取对象扩展到引言、讨论、结论等部分, 同时还需对领域词表或本体中主题之间的各种关系做进一步分析和利用。

在后续研究中, 将以现有研究结果为基础, 探索创新点的评估问题, 如判别创新点是否为原始创新、集成创新、理论创新或应用创新等, 使其应用更具实效性。

(致谢: 本文在修改过程中参考了审稿人提出的宝贵意见, 这些建议对本文的研究产生很多启发, 在此表示衷心感谢!)

参考文献:

[1] 温有奎, 徐国华, 赖伯年, 等. 知识元挖掘[M]. 西安: 西安电子科技大学出版社, 2005.(Wen Youkui, Xu Guohua, Lai Bonian, et al. Knowledge Element Mining [M]. Xi'an: Xi'an Electronic Science & Technology University Press, 2005.)

[2] 虞沪生, 张瑞清, 阎为民. 科技论文创新性的审读[J]. 编辑学报, 2006, 18(5): 333-334. (Yu Husheng, Zhang Ruiqing, Yan Weimin. Evaluation of Innovative Attribute of Scientific Papers [J]. Acta Editologica, 2006, 18(5): 333-334.)

[3] Dahl T. The Linguistic Representation of Rhetorical Function: A Study of How Economists Present Their Knowledge Claims [J]. Written Communication, 2009, 26(4): 370-391.

[4] 林浩欣, 阮明淑. 知识管理系统导入的知识主张研究—以软体公司知识管理顾问师为例[J]. 图书馆学与资讯科学, 2012, 38(1): 65-83. (Lin Hauhsin, Yuan Mingshu. A Study of Knowledge Claim in Implementing Knowledge Management System —An Example of Software Company's KM Consultants [J]. Journal of Library and Information Science, 2012, 38(1): 65-83.)

[5] Berkenkotter C, Huckin T N. Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power [M]. Lawrence Erlbaum Associates Inc, 1995.

[6] Trine D. Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics [J]. Journal of Pragmatics, 2008, 40(7): 1184-1201.

[7] Parkinson J. The Discussion Section as Argument: The Language Used to Prove Knowledge Claims [J]. English for Specific Purposes, 2011, 30(3): 164-175.

[8] Liu X, Guo C, Zhang L. Scholar Metadata and Knowledge Generation with Human and Artificial Intelligence [J]. Journal of the American Society for Information Science and Technology, 2014, 65(6): 1187-1201.

[9] Gonzalez E, Turmo J. Unsupervised Relation Extraction by Massive Clustering [C]. In: Proceedings of the 9th IEEE International Conference on Data Mining, Miami, FL, US.IEEE, 2009: 782-787.

[10] 温有奎, 温浩, 徐端颐, 等. 基于创新点的知识元挖掘[J]. 情报学报, 2005, 24(6): 663-668. (Wen Youkui, Wen Hao, Xu Duanyi, et al. Knowledge Element Mining in Knowledge Management [J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 663-668.)

[11] 杨硕, 崔蒙, 赵英凯, 等. 基于知识元的中医药信息知识标引[J]. 中国中医药信息杂志, 2011, 18(8): 24-25. (Yang Shuo, Cui Meng, Zhao Yingkai, et al. Knowledge Index about TCM Information Based on Knowledge Element [J]. Chinese Journal of Information on Traditional Chinese Medinine, 2011, 18(8): 24-25.)

[12] 冷伏海, 白如江, 祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作, 2013, 57(11): 112-119. (Leng Fuhai, Bai Rujiang, Zhu Qingsong. A Hybrid Semantic Information Extraction Method for Scientific Research Papers [J]. Library and Information Service, 2013, 57(11): 112-119.)

[13] Klavans J L, Muresan S. DEFINDER: Rule-based Methods for the Extraction of Medical Terminology and Their Associated Definitions from On-line Text[C]. In: Proceedings of the AMIA Symposium on American Medical Informatics Association, 2000:1049.

[14] 刘一宁, 郑彦宁, 化柏林. 学术定义抽取系统实现及实验分析[J]. 情报理论与实践, 2012, 34(12): 15-19. (Liu Yining, Zheng Yanning, Hua Bolin. Analysis and Realization of the Academic Definition Extraction System and Experiment [J]. Information Studies: Theory & Application, 2012, 34(12): 15-19.)

[15] Liu B, Chin C W, Ng H T. Mining Topic-Specific Concepts and Definitions on the Web [C]. In: Proceedings of the 12th International Conference on World Wide Web. ACM, 2003: 251-260.

[16] Swanson D R. Medical Literature as a Potential Source of New Knowledge [J]. Bulletin of the Medical Library Association, 1990, 78(1): 29-37.

[17] Chowdhury M N, Paul S, Sultana K Z. Statistical Analysis Based Hypothesis Testing Method in Biological Knowledge Discovery [J]. International Journal on Conputational Sciences & Applications, 2013, 3(6): 21-29.

[18] Cohen T, Widdows D, Schvaneveldt R W, et al. Discovering Discovery Patterns with Predication-based Semantic Indexing [J]. Journal of Biomedical Informatics, 2012, 45(6): 1049-1065.

[19] Teufel S, Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [J]. Computational Linguistics, 2002, 28(4): 409-445.

[20] Huang K C, Liu C C H, Yang S S, et al. Classification of PICO Elements by Text Features Systematically Extracted from PubMed Abstracts[C]. In: Proceedings of the 2011 IEEE International Conference on Granular Computing, Kaohsiung, Taiwan, China. IEEE, 2011: 279-283.

[21] Teufel S, Moens M. Discourse-level Argumentation in Scientific Articles: Human and Automatic Annotation [C]. In: Proceedings of the ACL Towards Standards and Tools for Discourse Tagging Workshop. 1999.

[22] Teufel S, Moens M. What's Yours and What's Mine: Determining Intellectual Attribution in Scientific Text [C]. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000: 9-17.

[23] Demner-Fushman D, Few B, Hauser S E, et al. Automatically Identifying Health Outcome Information in MEDLINE Records [J]. Journal of the American Medical Informatics Association, 2006, 13(1): 52-60.

[24] 温有奎, 温浩. 关键词与创新点词句群分布分析[J]. 情报学报, 2007, 26(1): 50-55. (Wen Youkui, Wen Hao. Sentence Group Distribution of Keywords and Innovation Idea Words [J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(1): 50-55.)

[25] Lock S. Structured Abstracts [J]. British Medical Journal, 1988, 297(6642): 156.

[26] Allan J, Wade C, Bolivar A. Retrieval and Novelty Detection at the Sentence Level [C]. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2003: 314-321.

[27] Kwee A T, Tsai F S, Tang W. Sentence-level Novelty Detection in English and Malay [C]. In: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowlege Discovery and Data Mining. Berlin Heidelberg: Springer, 2009: 40-51.

[28] Zhang Y, Callan J, Minka T. Novelty and Redundancy Detection in Adaptive Filtering [C]. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2002: 81-88.

[29] Zhang M, Song R, Lin C, et al. Expansion-based Technologies in Finding Relevant and New Information: THU TREC 2002 Novelty Track Experiments [C]. In: Proceedings of the 11th Text Retrieval Conference. 2002: 586-590.

[30] LingPipe 4.1.0 [CP/OL]. [2008-10-01]. http://alias-i.com/lingpipe/.

[31] The Stanford Natural Language Processing Group [EB/OL]. [2013-09-24]. http://nlp.stanford.edu.

[32] National Cancer Institute Thesaurus[EB/OL]. [2014-04-28]. http://ncit.nci.nih.gov/.

[33] 韩英，梁建莉. 英语医学论文标题的类型与翻译[J]. 新疆医科大学学报，2002, 25(1): 115-117. (Han Ying, Liang Jianli. Type and Translation of English Medical Paper Headline [J]. Journal of XinJiang Medical Universtity, 2002, 25(1): 115-117.)

作者贡献声明：

张帆：设计并实施技术方案、技术路线，数据采集、数据清洗，实验的分析和验证，论文的起草、撰写以及最终版本的修订；

乐小虬：提出论文研究方向和主要研究思路，指导研究方案及技术路线的设计，文章部分修改。

(通讯作者：张帆 E-mail: zhangf@mail.las.ac.cn)

Research on Innovation Points Extraction from Scientific Research Paper Based on Field Thesaurus

Zhang Fan[1,2]　Le Xiaoqiu[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This article aims to extract innovation points of sentence-level from scientific research paper of specific domain. [**Methods**] The field thesaurus and Ontology are used in constructing rules to extract innovation points from sentences in research papers, and a redundancy computing method based on keyword-overlap computing is used to filter redundant innovation points. [**Results**] The experiment is undertaken on data set of Neoplasm and the result shows that the accuracy rate is 89.42% and the recall rate is 60.14%. [**Limitations**] The rules need to be further improved, and the recall rate needs to be improved. [**Conclusions**] Using field thesaurus and the relationships in Ontology is effective in extracting innovation points from scientific research paper.

Keywords: Scientific research paper　Linguistic feature　Structured abstract　Innovation point extraction　Overlap computing