AMLC 检测医学论文的特点及期刊的应用对策

刘清海 王晓鹰 孙慧兰 张恩健 徐 杰
(中山大学学报编辑部/编辑学与出版研究中心,510080,广州)

摘 要 学术不端文献检测系统(AMLC)的成功开发和免费推广,为学术期刊抵制学术期刊中的学术不端行为提供了锐利的武器;但在多种学术不端行为中,AMLC 仅能检测一稿多投(或多发)、重复或抄袭、不正当署名等。本文检测了 178 篇来稿和 2 期共 55 篇已发表论文,发现约有 40% 的医学论文存在或多或少的文字重复现象,其中重复文字比 >30% 者达 16.7%。分析发现,医学论文"方法"和"讨论"部分重复文字较多、AMLC 对数字和符号不敏感等,也发现了一些较易重复的论文或项目。提出了 AMLC 的一些改良建议和期刊编辑部在使用过程中的一些对策。

关键词 学术不端行为;学术不端文献检测系统(AMLC);医学论文

Characteristics of AMLC in checking medical articles and its application in periodical editorial offices//LIU Qinghai, WANG Xiaoying, SUN Huilan, ZHANG Enjian, XU Jie

Abstract Academic Misconduct Literature Check (AMLC) has provided editors a sharp weapon against misconduct articles since its development and extension. However AMLC can only check a few academic misconducts, such as multiple contribution of same manuscript, pleonasm or plagiarizing, unwarrantable signature, and so on. We checked 178 medical manuscripts and 55 articles published in 2 issues, and found that almost 40% of them repeated other articles´ sentences or passages to some extent, among which those repeating ratio with more than 30% reached 16.7%. Analysis found that methods and discussion sections of articles got higher repeating ratio, and AMLC was non-sensitive to numbers and signs. Kinds of articles and items prone to repetition are also described. Improving measures for AMLC are suggested, and applications of AMLC in journal editor offices are also noticed

Key words academic misconduct; AMLC; medical article

Author's address Editorial Department of Journal of SUN Yat-sen University, 510080, Guangzhou, China

学术不端行为由来已久,国内外皆有之;但国内这些年来,这种行为有愈演愈烈之势,甚至有学术不端行为者反告揭露其行为者竟而胜诉的事件[1]。这种趋势促使学术界、学术管理界和高等教育界皆呼吁——要坚决反对和打击学术不端行为。2008 年 10 月,武汉第 7 届全国综合类人文社会科学期刊高层论坛一致通过了由 50 家期刊共同签署的《关于坚决抵制学术不端行为的联合声明》[2]。2009 年 3 月,教育部发出《关于严肃处理高等学校学术不端行为的通知》,并声明年底前教育部会对该通知的执行情况进行专项检查[3]。2008 年底,中国学术期刊(光盘版)电子杂志社与同方知网(北京)技术有限公司合作成立的 CNKI 科研诚信管理系统研究中心成功开发出旨在检测学术文献当中不端行为的 AMLC 系统,包括科技期刊学术不端文献检测系统(AMLC)、社科期刊学术不端文献检测系统(SMLC)和学位论文学术不端行为检测系统(TMLC),这无疑是学术期刊界反对学术不端行为的锐利武器。

我们自 2009 年初开始使用该系统,对来稿和已发表文献进行了检测,发现该系统对医学论文学术不端行为的检测有一定帮助,本文对医学期刊编辑部使用该系统提出了一些建议,也对该系统提出了一些改进建议。

1 AMLC 对学术不端医学文献的划分

学术不端行为的种类很多,概念的界定也不尽一致,其共同特点是涉及学术领域的、违反基本学术规范、违反学术道德的行为[4]。教育部《关于严肃处理高等学校学术不端行为的通知》中列举了 6 种外加其他学术不端行为[3],而在综合类人文社科期刊高层论坛签署的《关于坚决抵制学术不端行为的联合声明》则列举了学术期刊中的 5 种学术不端行为[2]。马勇进等[4]认为学术成果低水平重复也是一种学术不端行为。潘淑君等[5]更进一步认为学术期刊的编委头衔及增刊和加大页码也属于学术不端行为。可见,学术不端行为在学术期刊中的表现多样,不局限于期刊论文本身。

AMLC 在"用户说明"[6]中明确提出,用户稿件的检测主要检测是否存在抄袭、一稿多投和已发表文献的不正当署名等学术不端行为,并依严重程度给出一个权值进行度量。可见,AMLC 难以检测和防止其他类型的学术不端行为。该系统按照重复文字的相似比例,以 10%、30%、50% 为界划分为轻度句子抄袭(<10% 且各连续重合文字均 <200)、句子抄袭(≥10% 且各连续重合文字均 <200)、轻度段落抄袭(≥10% ~ <30% 且存在连续重合文字≥200)、段落抄袭(≥30% ~ <50% 且存在连续重合文字≥200)和整体抄袭(≥50%,重复文字大于总字符数的 1/2),并有相应的绿、黄、橙、红等颜色醒目提示。该系统按是否有部分

作者相同判断是否自我抄袭,如有且文字重合度≥75%则认为是一稿多投,若<75%且≥50%则认为是自抄,若<50%且≥30%则认为是轻度自抄。系统比对参考文献部分,从而确定被检测文献是否引用了被重复的文献。系统还对被重复文献是来自于某 1 篇或多篇而将重复的性质划分为单源和多源。

2　AMLC 检测医学论文的总体情况

我们检测了 2009 年前几个月来稿的 178 篇文献,发现 73 篇文献存在一定的问题,同时检测了已发表文献 2 期共 55 篇论文,发现有 21 篇存在一定的问题。可见来稿中约有 40%的文献存在或多或少的文字重复现象,有个别文献的重复文字比甚至达到 81%,与单篇文献的重复文字比达 77%,被判断为"整体抄袭(多源)",检测文献与重复文献间没有引证关系。也有不少文献重复文字比<10%:在我们检测出的 73 篇可能有问题的来稿中,重复文字比>10%的文献占 48 篇,另 25 篇为重复文字比≤10%。在所有来稿和已发表的 2 期的 233 篇文献中,总重复文字比≥30%的文献有 39 篇,占 16.7%,其中重复文字比≥50%的文献 18 篇,占 7.7%。然而,该系统并不能对所有文献都给出明确的诊断。在我们的检测结果中,只有后 4 种诊断才有显示,即句子抄袭、轻度段落抄袭、段落抄袭和整体抄袭,分别为 13(5.6%)、5(2.1%)、8(3.4%)和 3(1.2%)篇,另有 11 篇(4.7%)轻度自抄和 2 篇(0.9%)自抄。可见,有重复文字的医学论文中,句子抄袭现象和轻度自抄现象比较多,整体抄袭则少见,而段落抄袭反而多于轻度段落抄袭的结果可能与抽样有关。

3　AMLC 检测医学论文的一些特点

为了解 AMLC 检测医学论文的特点,以便医学期刊编辑部有针对性地使用,我们对每一篇总重复文字比>20%且与单篇文献重复文字比>10%的文献进行了全文比对,发现 AMLC 检测医学论文有以下特点。

1)总体上,医学论文的重复文字在"方法"和"讨论"部分较多,而在"摘要""引言""结果"部分较少,说明多数有重复文字的论文是参照了其他论文方法的写法,而有自己的结果,但讨论的大部分内容却没有新意。这与医学科研的现状是吻合的——方法没有创新,仅用他人方法重复自己的实验;因此结果的数据略有不同,但讨论和结论却是类似的。

2)AMLC 不仅比对公开发行的科技期刊,还比对学位论文、内部刊物、重要会议论文、报纸、年鉴、工具书等,因为 AMLC 以 CNKI 收录的 9 000 余种国内刊物和其他资料为基础进行文字重复的检测。

3)AMLC 对数字和符号不敏感,因此,同一句话即使其中的数据或符号不同,也会被认为文字重复。如一个课题的系列研究中需要研究多个因子的状况时,则容易被认为文字重复,甚至被认为句子抄袭或段落抄袭。

4)AMLC 对重复段落的划分与自然段没有关系,有可能发生不同自然段甚至不同层次被划分为同一重复文字段的现象。如医学论文的第 1 层次通常为"材料与方法",临床研究的论文在该层的上面通常以"现报告如下"结尾,而该层下则常以"研究病例选自××院自××年—××年的患某病的住院病例"。由于 AMLC 对数字不敏感,因此,这个地方的文字较易被认为文字重复,且在全文比对的重复文字标记时被认为是同一个重复段落。

5)AMLC 的检测结果有几种不一致的情况。

①方法与讨论部分的重复段落通常较长,而在引言与结果部分则重复段落较短;因此,出现方法部分与讨论部分标记的重复文字段落数还可能不如结果部分的标记段落数多。

②与上一条相关,总体文字复制比可能与标记的重复段落数不成正比。

③中英文摘要的重复文字可以不一致,有些论文中文摘要与某论文重复,而英文摘要却可能没有重复。

④单篇文献比对时发现,重复句子或段落的顺序可与被重复的论文出现的顺序不一致。

⑤仅与单篇文献重复时,总体的文字复制比与列出的该单篇文献的重复文字比常常不同,多数为稍大于单篇文献的重复文字比,可能是因为与未列出的文献仍有小部分重复。

6)AMLC 不仅仅比对题名、摘要和正文,还比对作者单位及参考文献。我们在检测的过程中发现作者单位相同也可被认为是文字重复。引用相同参考文献也可被认为是文字重复。

7)几种较易出现重复文字的论文。

①不同作者研究同一批资料的不同方面时,较易发生抄袭资料描述,如同一批病例,医生撰写了有关资料的临床效果的论文,则护士可能抄袭有关资料描述。

②同一课题组的系列研究时,后续论文可能抄袭已发论文的引言、方法和讨论等内容。曾有一篇论文重复文字比为 34%,经仔细比对,发现大部分都是在方法部分重复了课题已发表的论文。

③研究性论文在讨论部分容易抄袭综述性文章。

④博士毕业后发表的论文容易抄袭博士生答辩的毕业论文。

⑤论文在内部刊物上发表后,相同主题正式发表

时,会有较多重复内部刊物论文的现象。

⑥由于 AMLC 对数字和符号不敏感,因此,会把类似的图题表题和注释认为是文字重复——这也是结果部分较多见的重复文字标记。

⑦统计方法部分,多有重复统计软件和一般统计方法的描述。

⑧较长的并列名称较易被发现重复,如我们曾检测出眼科学"全周RNFL厚度、上方 RNFL 厚度和下方 RNFL 厚度……"在同一文章的讨论中被标记重复 10 次之多,实际不应算是重复。

8)医学论文的遣词造句比较简单和通用,因此,在各个部分都容易出现文字重复的现象;但大多数被认为有较大重复文字比例的文章都是"多源"重复,即与多篇文献有类似的文句,而不是抄袭某一篇的句段。我们曾检测到一篇文献总体复制比大于50%,但经仔细比对,发现与单篇文献的重复文字比最大也仅为17%,其他多数为10%以下。这篇论文未被判断为"整体抄袭",可能是由于没有某段连续的重复文字>200字。

9)大多数论文不引用被抄袭的论文,仅少数抄袭较轻者有引用。我们检测的73篇有抄袭嫌疑的文献中,仅有2篇引用了被重复文字的论文,一篇被判为"句子抄袭",另一篇为作者自己的论文,因重复文字较少,系统尚未判断为学术不端行为类型。

10)短文章容易导致重复文字比例偏大,如仅输入某发表文献的摘要时,常被判断为"一稿多投(主观)"。

4 关于 AMLC 检测医学论文的一些建议

马勇进[4]认为,学术不端行为在学术期刊中大量存在,主要有社会与个人的原因,缺乏有效惩戒措施,以及学术期刊自身的问题(如把关不严、自律缺失),因此,抵制学术不端行为是学术期刊的神圣职责。曹亚军等[7]也认为,出版与编辑环节存在的漏洞及某些新闻媒体的推波助澜为科研不端行为摇旗呐喊,助长了学术造假之风。在 AMLC 开发以前,要鉴别来稿是否重复或抄袭他人文章,多采用文献检索的方法,但是,检索出来的相关文章较多,编辑难以逐句逐段地对照每一篇文章。AMLC 的开发成功,对学术期刊编辑鉴别这些学术不端文献提供了强有力的武器。我们建议,所有具备条件的编辑部都应在收到稿件之时就利用 AMLC 进行学术不端文献的检测;但是,在使用过程中,我们发现 AMLC 自身还难以完全替代编辑的鉴别,并且其自身也还存在一定的问题,某些问题不仅是对医学论文,而且对所有学术期刊的论文都可能存在。

鉴于上述的 AMLC 检测医学论文的一些特点,我们对 AMLC 系统和医学期刊编辑部使用该系统时提出如下建议。

1)对于 AMLC 系统,我们建议:

①由于医学研究性论文(综述与讲座除外)都是按照"引言""材料与方法""结果""讨论"的章节层次撰写的,因此,AMLC 系统应提供入口使得医学论文可以分层次进行重复文字比例的计算,关键的"结果"部分是编辑必须特别注意的,而"方法"与"讨论"部分的少许重复并不要紧。

②标记重复文字段落时同时考虑论文自然段的划分,不要将不同自然段甚至不同层次的内容混合在一起统计和标记,层次标题不计入重复文字比例。

③AMLC 系统不要将作者单位和参考文献列入重复文字的检测范围,因为不同论文是可以引用相同文献的。

④系统增加对图表与数字和符号的识别功能。

⑤在给出总结性数据时,同时给出与某单篇文献重复文字比的最大比例,以方便期刊编辑考察可疑论文与某单篇文献的最大重复比例。

2)对于医学期刊编辑部,我们建议:

①对所有来稿,在登记前就应进行学术不端文献的检测,而将检测出的学术不端文献即行退稿,以避免人力物力浪费。

②在每一期发稿之前再行一遍学术不端文献的检查;因为从来稿到发稿,国内大部分期刊都要经过约半年的时间,检测过的文献有可能半年后发现有重复或抄袭行为。这一次检测可采用将所有发稿论文压缩后打包而进行批处理的方法检测。

③编辑在检测文献时,不要单纯按照 AMLC 给出的总体复制比对来稿进行学术不端行为的判断,而应深入分析其与某几篇单篇文献的重复文字情况,如仅是方法上的重复,可以建议作者大力精简方法部分,仅引用文献即可。

④编辑部应制订统一标准,以便各编辑按照相同的标准处理来稿。

⑤使用了远程采编系统的编辑部可与采编系统设计部门协商,看是否可以将来稿在 AMLC 的比对设计成自动化程序。

⑥编辑在处理重复文字的稿件时,应慎重对待,如退稿,则退稿理由不可过于武断地认为作者是抄袭;因为医学论文的遣词造句确实比较通用,容易发生重复现象。

⑦学术不端行为的检测不可替代文献查新,经 AMLC 检测出来没有多少重复文字比例,并不说明论文具有新颖性或创造性。