• 医学信息研究 •

精准医学文本语料库构建研究*

周永称 范少萍 晏归来 安新颖

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 利用 BRAT 标注工具及人工标注方法，根据初步制定的文本语料标注规范，按照语料选择、语料预处理、语料标注、标注校对、一致性检验的流程开展语料标注工作，构建包含 2 000 篇文本语料的精准医学文本语料库，为相关实体与关系抽取算法研究以及精准医学知识网络构建提供有效的资源支持。

〔关键词〕 文本语料库；精准医学；语料库构建

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2019.12.010

Study on the Building of Precision Medicine Text Corpus *ZHOU Yongcheng, FAN Shaoping, YAN Guilai, AN Xinying, Institute of Medical Information, CAMS&PUMC, Beijing 100020, China*

〔**Abstract**〕 Using BRAT annotation tool and manual annotation method, according to the preliminary text corpus annotation specifications, the corpus annotation is carried out in accordance with the procedures of corpus selection, corpus preprocessing, corpus annotation, annotation proofreading, and consistency check. The precision medicine text corpus containing 2 000 text corpora has been built, providing effective resources support for the study of related entity and relation extraction algorithm and the building of precision medicine knowledge network.

〔**Keywords**〕 text corpus; precision medicine; corpus building

1 引言

随着生物医学文献数量快速增长，大量有价值的生物医学知识被隐藏在海量文献中，如基因与疾病、基因与药物、疾病与药物、基因与蛋白质之间的作用关系等。生物医学文本信息抽取已成为生物医学信息学领域与计算机领域的研究热点[1]。D. Swanson 基于生物医学文献利用实体间的共现关系发现深海鱼油与雷诺氏病之间存在相关关系，这为雷诺氏病的治疗提供新方法[2]。此后基于生物医学文献陆续发现镁元素与偏头疼之间、血液中生长调节素水平与精氨酸之间均存在先前未知的相关关系[3]。利用监督或半监督学习方法构建学习器的过程中需要标注语料作为机器学习算法的训练数据，以验证算法的性能。标注语料的规模与质量对评测机器学习算法的效果与性能至关重要。国外已有的生物医学文本语料库根据标注语料的实体类型及关系类型的不同可分为多种类型，如标注蛋白质及其

〔修回日期〕 2019－07－26

〔作者简介〕 周永称，硕士研究生，发表论文 2 篇；通讯作者：安新颖，博士，研究员，发表论文 40 余篇。

〔基金项目〕 国家重点研发计划“精准医学文本知识网络构建”子课题“精准医学文本语料库构建”（项目编号：2016YFC0901902－2）；国家自然科学基金项目“面向精准医学的基因－疾病－药物语义关义抽取研究”（项目编号：71704188）；中央级基本科研业务费项目“临床医学高层次人才科研能力评价研究”（项目编号：2018TX63002）。

相互关系的 AIMED[4]、BioInfer[5]、HPRD50[6]、IEPA[7]和 LLL 语料库[8]等；标注基因型及表型的 MKH[9]、Phenominer 语料库[10]等以及标注文本语料中缩写实体的 Ab3P[11]、BIOADI[12]、Medstract 语料库[13]等。这些语料库的规模、标注的实体数、关系数等，见表 1。上述语料库已被用于文本挖掘竞赛中的评测语料库，如 HPRD50 语料库作为 RelEx 系统的测试语料库；LLL 语料库被用于第 4 届逻辑学习语言研讨会生物医学文本关系挖掘竞赛的共享数据集；Medstract 语料库则是国际上生物医学命名实体识别领域应用较为广泛的评测数据集，为相关机器学习算法的开发与评测提供标准与依据。由表 1 可知现有生物医学文本语料库的规模较小，标注的实体、关系类型单一且数量较少，以此为基础训练的机器学习算法对实体识别和关系抽取会局限在特定的实体及关系类型，鲁棒性较差，无法全面挖掘出文献中隐含的知识。但随着基因组学、蛋白质组学、代谢组学等多组学的发展以及精准医学项目对于疾病、基因、分子等多种实体类型之间关系的研究，生物医学文献中涉及的实体、关系类型不再单一。因此构建一个更大规模、覆盖面更广的文本语料库对于机器学习算法的研究大有裨益，以支持精准医学知识网络相关研究。基于此本研究在国家重点研发计划“精准医学”的项目资助下，构建规模更大、实体与关系类型覆盖面更广的精准医学文本语料库，以支撑相关机器学习算法与知识图谱的开发与构建工作，从而推动精准医学研究的进一步发展。

表 1 部分语料库分类、名称、规模、标注内容

语料库分类	语料库名称	语料规模	标注内容
蛋白质相互作用关系	AIMED	1 955(sentence)	1 000(相互作用关系正例)
	BioInfer	1 100(sentence)	2 534(相互作用关系正例)
	IEPA	486(sentence)	335(相互作用关系正例)
	HPRD50	145(sentence)	163(相互作用关系正例)
	LLL	77(sentence)	164(相互作用关系正例)
缩写实体	Ab3P	1 250(PubMed abstract)	1 221(SF - LF)
	BIOADI	1 200(PubMed abstract)	1 668(SF - LF)
	MEDSTRACT	199(PubMed citations)	168(PFS - PLF)
基因型及表型	Phenominer	1 976(sentences)	1 611(基因) 472 (表型)
	MKH	2 755(sentences)	4 233(表型)

2 精准医学文本语料库构建

2.1 概述

通过调研多种语料库，笔者总结出文本语料库的构建流程大致可分为标注语料选择、文本标注和一致性检验 3 个步骤，其中文本标注包括标注工具选择、语料预处理、语料标注和标注校对。本研究按照这 3 个步骤开展精准医学文本语料库构建工作。选取 2 000 篇癌症领域研究文献作为标注语料，利用课题组提供的精准医学本体开展精准医学文本语料库构建研究。

2.2 标注文献筛选方法与流程

课题组精准医学本体 V1.0 版涉及 6 大类实体类型：细胞机制（Cellular Mechanisms）、化合物和药物（Chemicals and Drugs）、疾病（Diseases）、遗传机制（Genetic Mechanisms）、人类表型（Human Phenotypes）、分子机制（Molecular Mechanisms）。为保证文献对实体类型的覆盖面，尽量避免实体数量分布不均衡，本研究制定以下文献筛选流程：第一，利用比较毒理组学数据库（Comparative Toxicogenomics Database，CTD）[14]检索相关文献，根据文献所在期刊 IF 值、发表年份、文献所含实体类型对相关文献进行筛选。CTD 由北卡罗莱纳州立大学的国立环境健康科学研究所资助开发，数据经过人工编审并且每年更新数次，可以确保数据质量与有效性。同时可显示文献所包含的疾病、化合物及基因类型，可在一定程度上确保所检文献的实体覆盖面。第二，将筛选得到的训练集文献作为输入集，

利用文献相似性算法检索相似文献。通过 CTD 检索到相关文献后，经过文献筛选最终得到的训练集文献数量较少，直接将其作为标注文献集无法满足精准医学文本语料库对语料规模的需求，因此需要进行文献相似性检索以扩展标注文献集的内容。文献相似性检索工具有多种，如 Medline Ranker[15]、PubFinder[16]、MScanner[17]等，上述工具可自动从输入的一组摘要中提取主题并在 Medline 数据库中进行相似性检索。但 MScanner 工具在进行摘要的判断性特征抽取时主要利用摘要的注释，如 MeSH 主题词或期刊标识符；PubFinder 工具开发的时间较早，近年来缺少更新；而 Medline Ranker 直接从文本中提取特征词，减少对摘要的依赖性，且工具较新，因此选用 Medline Ranker 工具获取相似文献。第三，将训练集文献及相似文献提交给专家判读，确定最终的标注文献集。训练集文献包括 210 篇经筛选的肝癌相关文献及 158 篇经筛选的肠癌相关文献，相似文献集包括 1 000 篇肝癌研究相似文献及 1 000篇肠癌研究相似文献，共计 2 368 篇。将2 368 篇文献提交给专家判读，专家主要包括中国医学科学院基础医学研究所及吉林大学从事肝癌研究以及医学分子生物学研究的专家共 3 名。3 位专家从研究内容及实体覆盖面两个方面确定纳入文献，经过专家判读后形成规模为 2 000 篇的标注文献集。文献筛选流程，见图 1。

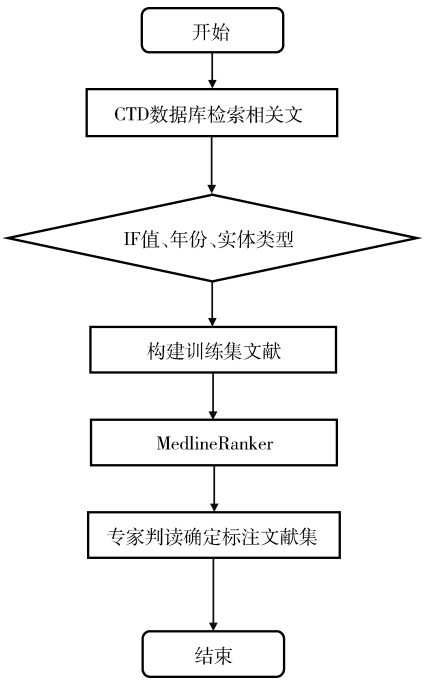

图 1 标注文献筛选流程

2.3 文本标注

2.3.1 标注工具 利用标注工具可提高语料标注效率及准确性，因此选择或者开发语料标注工具是构建大规模语料库的前提。目前用于生物医学文本语料标注工作的标注工具有多种，其中开源的并且已被用于生物医学文本标注的工具有 Argo[18]、CALLISTO[19]、GAET[20]、MyMiner[21]、Semantator[22]、BRAT[23]等，其性能对比，见表 2[24]。

表 2 部分生物医学文本标注工具对比

标注工具	发行时间(年)	开发团队	标注实体	标注关系	在线标注	响应速度
Argo	2012	曼彻斯特大学国家文本挖掘中心	Yes	No	Yes	Fine
CALLISTO	2004	美国 MITRE 公司	Yes	Fine	No	Fine
GAET	2010	英国谢菲尔德大学	Yes	Fine	Yes	Good
MyMiner	2012	澳大利亚莫纳什大学、西班牙国家癌症研究中心、法国地中海大学、印度理工学院	Yes	Poor	Yes	Good
Semantator	2011	美国梅奥诊所、美国利哈伊大学	Yes	Fine	Yes	Good
BRAT	2012	日本东京大学、英国曼彻斯特大学	Yes	Very good	Yes	Very good

可知相比于其他标注工具，BRAT 在对关系的标注及响应速度上有较大优势。此外 BRAT 工具在安装与配置上较其他标注工具更简单。基于此，本文选择的标注工具为 BRAT 文本标注工具。BRAT

基于 Web，其生成的标注结果可以将非结构化的原始文本结构化，实现对文本的结构化标注并供计算机处理。BRAT 既支持用户对文本进行手工标注，也可以利用其配置的工具对文本进行自动标注，或者对其他标注工具的标注结果进行可视化展示。通过对配置文件进行修改可定义标注的实体名称以及实体间的关系类型。该工具已被成功用于多种语料库的构建（包括 Anatomical Entity Mention（AnEM）corpus、CellFinder corpus、Multi – Level Event Extraction corpus）。本文利用该标注工具进行 2 000 篇标注集文献的标注。

2.3.2 标注流程及结果 语料标注流程分为语料预处理和语料标注两部分。语料预处理目的在于将待标注的生物医学文献进行格式转换，将语料格式转换为 BRAT 工具可读取的 .TXT 格式文件并将文献标题及摘要部分进行断句切分。由于本研究的目的在于构建一个可作为金标准的精准医学文本语料库，因此语料标注研究采用人工标注方法。人工标注模式可分为 3 种：传统的领域专家标注、众包标注以及团体标注[25]，根据精准医学文本语料标注工作专业性强且工作量大的特点，分别从中国医学科学院医学信息研究所、中国军事医学科学院、中国医科大学、吉林大学、广东药科大学等院校招募具有医学背景的硕士研究生若干名作为标注人员分批次进行标注培训。在标注过程中，BRAT 工具配置 4 个外部链接：MeSH、NCIt、SNOMEDCT、Wikipedia 以帮助标注人员遇到自由词或模糊概念词时快速有效地确定正确的概念与语义类型。借助外部链接确定实体类型并标注，标注效果，见图 2。由于配置的外部链接中不能完全涵盖精准医学本体涉及的实体类型，标注过程中需要标注人员更多地利用标注规范开展实体及关系标注。语料库标注规范是标注人员遵循的最重要的原则，直接影响标注结果质量。为确保标注结果的科学性、有效性与可用性，本研究在精准医学本体 V1.0 版本以及 2 000 篇文献标注的基础上逐步完善语料库标注规范，形成精准医学语料库标注规范 V1.0 版本并根据形成的标注规范对 2 000 篇文献开展 2 次标注。精准医学语料库标注规范 V1.0 版本主要包括 3 部分：精准医学本体语义类型、关系来源及概念解析、实体标注规范以及关系标注规范。第 1 部分主要根据精准医学本体对概念词的来源进行标准化，以辅助标注人员在遇到模糊自由词时可以快速有效地确定正确的概念及语义类型；第 2 部分与第 3 部分主要针对 6 大类实体和下位类实体以及关系类型给出其定义以及标注示例，辅助标注人员了解概念内容与语义场景。此外，针对 6 大类实体类型中均存在的“全称（缩写）”形式词语的标注，初步确定标注 3 次的原则，即“全程”、“缩写”、“全称（缩写）”均需标注；针对“名称 + 类型”形式词语的标注，经过与后续机器学习算法开发项目组成员的确认，确定“名称 + 类型”形式词语一起标注为一个实体的原则，见图 3。

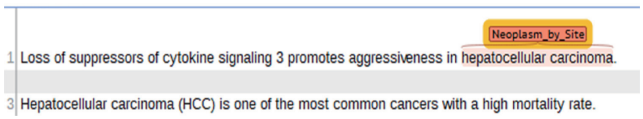

图 2 实体标注效果

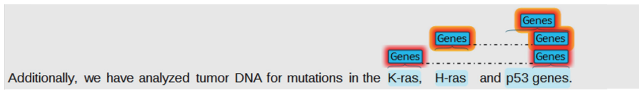

图 3 “名称 + 类型”形式词语标注

2 000 篇标注语料中标注的实体及关系数量，见表 3、表 4。可看出所选的 2 000 篇语料标注的实体及关系总量较大，各个实体类型均有涉及且数量较多，确保语料对于实体的覆盖面。

表 3 2 000 篇语料实体及关系总量统计

实体总量	不重复实体总量	关系总量
92 917	23 725	22 622

表 4 500 篇语料各类型实体数量统计

实体类型	总量	不重复数量
Cellular Mechanisms	11 939	3 006
Molecular Mechanisms	5 203	2 484
Genetic Mechanisms	8 170	3 453
Chemicals and Drugs	45 970	11 300
Diseases	14 505	2 335
Human Phenotype	7 024	2 495

2.4 一致性检验

用来对语料的标注质量进行把关。在语料标注过程中，为确保语料标注的正确性以及方便后续一致性检验工作，每篇文献均由两名标注者“背靠背”标注完成。常见的一致性检验方法有 Kappa 检验[26]、F - measure[27]等。Kappa 检验用 Kappa 统计量来反应一致性程度的高低，计算方法见式（1）、式（2）、式（3）：

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

$$Po = \frac{a + d}{n} \quad (2)$$

$$Pe = \frac{(a + b)(a + c) + (c + d)(b + d)}{n^2} \quad (3)$$

Kappa 检验经常用于临床中对两种诊断方法诊断某疾病的一致性评判，n 表示接受两种方法诊断的患者数。如用 A、B 方法针对 C 疾病检测，其结果，见表 5。

表 5 A、B 方法针对 C 疾病的检测结果

方法		A 方法		
		阳性	阴性	汇总
B 方法	阳性	a	b	$a+b$
	阴性	c	d	$c+d$
	汇总	$a+c$	$b+d$	$n\ (a+b+c+d)$

用 F - measure 方法进行一致性检验的前提是存在金标准，因为 F - measure 是关于精确率和召回率的加权调和平均值，在没有金标准存在的语料标注工作中各标注者标注工作的精确率和召回率无法计算，因此 F - measure 方法也不适用于本研究的一致性检验环节。基于此，用 Jaccard Score 算法[28]进行语料标注的一致性检验，其计算方法，见式（4）：

$$S_{Ai,Aj} = \frac{Ai \cap Aj}{Ai \cup Aj} \quad (4)$$

其中 A_i 表示在语料实体类型的标注过程中被标注者 i 标注为实体类型 A 的实体集；Aj 表示在语料实体类型的标注过程中被标注者 j 标注为实体类型 A 的实体集；$A_i \cap A_j$ 表示对于实体类型 A，两个标注者标注完全一致的实体集；$A_i \cup A_j$ 表示两个标注者标注为实体类型 A 的所有实体。为确保标注结果的可靠性及标注质量，在一致性检验的计算过程中使用严格一致性检验，即在确认两个标注者对于某个实体的标注是否完全一致时必须综合考量标注的实体类型、出现位置、覆盖范围，当两个标注者对该实体的标注在这 3 方面完全一致时才视为标注一致。由于每篇文献均由两名标注者“背靠背”标注完成，在标注者严格遵守标注规范的前提下，一致性检验结果可以反映标注规范的有效性及准确性，一致性较低的文献进行再次加工以提高标注质量。在已标注语料文献集中随机抽取 300 篇完成“背对背”标注的文献，利用 Jaccard Score 算法计算一致性，各类实体及关系的平均一致性，见表 6。可知部分实体的一致性较高，如“Diseases”类型实体，部分实体的一致性程度略有不足，如“Molecular Mechanisms”及“Human Phenotype”类型实体。针对一致性较低类型剖析其原因：从本体角度来说，这两类实体的下位类实体类型较多，部分下位类与其他大类下位类内容存在交叉，且下位类实体层级结构较深，这导致不同标注人员在标注时容易将同一概念词标注为不同类型或不同层级的实体；从文献角度来说，同一篇文献无法涵盖本体涉及的全部实体类型，对不同类型实体的覆盖各有侧重，相较于其他实体类型而言，文献对于“Molecular Mechanisms”类型及“Human Phenotype”类型实体的覆盖不足，这导致标注为这两个类型实体的数量较少。根据 Jaccard Score 算法计算一致性时，当标注不一致的实体数量相同时文献中实体数量越少的实体类型其标注一致性计算结果越低。

表 6 300 篇癌症文献实体及关系的平均一致性

类型	平均一致性（%）
总体	69. 86
Cellular Mechanisms	75. 38
Molecular Mechanisms	59. 31
Genetic Mechanisms	67. 49
Chemicals and Drugs	71. 42
Diseases	81. 74
Human Phenotype	60. 03
关系	60. 56

3 结语

国内对精准医学文本语料库构建研究较少，本研究在参考与调研多个国外经典标注语料库的基础上，经过语料选择、语料预处理、语料标注、一致性检验等过程完成 2 000 篇癌症相关文献的语料标注，构建一定规模的精准医学文本语料库。该语料库有以下特点：一是语料规模更大。目前已经完成 2 000 篇语料的标注工作，标注的语料句子约 20 000 个。二是实体、关系覆盖面更广。在进行语料选择时充分考虑文献对实体的覆盖面，标注结果中，包含约 23 725 个不重复实体与 22 622 种关系，更利于潜在的有价值生物医学知识的挖掘。三是标注结果具有较好的一致性。利用“背靠背”标注方式进行语料的人工标注并检验一致性，对一致性低的语料进行再加工，保证语料的标注质量。在今后的标注工作中将从以下 3 个方面进一步完善：第一，开发新的文献相似性算法，完善标注文献选择工作，使得实体与关系覆盖面更加均衡有效。标注文献的选择是精准医学文本语料库构建工作的第 1 步，也是语料库构建的基础。标注语料质量直接影响最终的标注效果。因此开发新的文献相似性算法以完善标注文献的选择工作对精准医学文本语料库的构建大有裨益。第二，调整文献筛选策略。随着项目组精准医学本体版本的变更与完善，本体涵盖的实体及关系范围越来越广、越来越细，且精准医学研究较为注重生物通路及遗传突变，因此在后续语料库标注过程中需要根据精准医学研究重点及本体内容及时调整文献筛选策略。第三，进一步扩大精准医学文本语料库覆盖疾病类型。精准医学的本质是利用生物信息、现代遗传技术以及基因组等组学技术精确定位疾病的致病位点以及治疗靶点，以实现疾病的个性化治疗。其短期目标致力于为癌症疾病提供更好的治疗方法，长期目标则致力于为多种疾病的治疗提供服务。因此精准医学文本语料库覆盖的类型也需相应增加。目前标注疾病类型仅限于癌症，未来将增加呼吸系统疾病、心血管疾病等类型，充分满足精准医学疾病谱研究需要。

参考文献

1 顾静航. 面向生物医学领域的实体关系抽取研究［D］. 苏州：苏州大学，2017.

2 Swanson D R. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge［J］. Perspectives in biology and Medicine, 1986, 30（1）: 7 – 18.

3 崔雷. 生物医学文本挖掘：步骤与工具［J］. 中华医学图书情报杂志，2017，26（3）：1 – 5.

4 Bunescu R, Ge R F, Kate R J, et al. Comparative Experiments on Learning Information Extractors for Proteins and Their Interactions［J］. Artificial Intelligence in Medicine, 2005, 33（2）: 139 – 155.

5 Pyysalo S, Ginter F, Heimonen J, et al. BioInfer: a corpus for information extraction in the biomedical domain［J］. BMC Bioinformatics, 2007, 8（1）: 50.

6 Fundel K, Kueffner R, Zimmer R. RelEx – relation extraction using dependency parse trees［J］. Bioinformatics, 2007, 23（3）: 365 – 371.

7 Ding J, Berleant D, Nettleton D, et al. Mining MEDLINE: abstracts, sentences, or phrases?［EB/OL］.［2019 – 12 – 20］. https://www.ncbi.nlm.nih.gov/pubmed/?term=11928487.

8 Pyysalo S, Airola A, Heimonen J, et al. Comparative Analysis of Five Protein – protein Interaction Corpora［J］. BMC Bioinformatics, 2008, 9（3）: 1 – 11.

9 Khordad M, Mercer R E, Rogan P K, et al. A Machine Learning Approach for Phenotype Name Recognition.［EB/OL］.［2019 – 12 – 20］. https://cn.bing.com/academic/profile?id=79957832e95697f9b7a743ff4ca263ad&encoded=0&v=paper_preview&mkt=zh – cn#.

10 Collier N, Groza T, Smedley D, et al. PhenoMiner: from text to a database of phenotypes associated with OMIM diseases.［EB/OL］.［2019 – 12 – 20］. https://www.ncbi.nlm.nih.gov/pubmed/?term=26507285.

11 Sohn S, Comeau D C, Kim W, et al. Abbreviation Defini-

tion Identification based on Automatic Precision Estimates [J]. BMC Bioinformatics, 2008, 9 (1): 402.

12 Kuo C, Ling M H T, Lin K, et al. BIOADI: a machine learning approach to identifying abbreviations and Definitions in biological Literature [J]. BMC Bioinformatics, 2009, 10 (15): 1 - 10.

13 Pustejovsky J, Castano J, Cochran B, et al. Automatic Extraction of Acronym - meaning Pairs from MEDLINE Databases [J]. Studies in Health Technology and Informatics, 2001, 84 (Pt1): 371 - 375.

14 Mattingly C J, Rosenstein M C, Colby G T, et al. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies [J]. Journal of Experimental Zoology Part A Ecological Genetics & Physiology, 2003, 111 (6): 793 - 795.

15 Fontaine J, Barbosa - Silva A, Scgaefer M, et al. MedlineRanker: flexible ranking of biomedical literature [J]. Nucleic Acids Research, 2009 (37): W141 - W146.

16 Goetz T, Von Der Lieth C W. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts [J]. Nucleic Acids Research, 2005 (332): W774 - W778.

17 Poulter G L, Rubin D L, Altman R B, et al. MScanner: a classifier for retrieving medline citations [J]. BMC Bioinformatics, 2008, 9 (1): 108.

18 Rak R, Rowley A, Black W J, et al. Argo: an integrative, interactive, text mining - based workbench supporting curation [EB/OL]. [2019 - 12 - 20]. https://www.ncbi.nlm.nih.gov/pubmed/? term = 22434844.

19 Rigouste L, Szpakowicz S, Japkowicz N, et al. An Automatic Evaluation Framework for Improving a Configurable Text Summarizer [EB/OL]. [2019 - 12 - 20]. https://cn.bing.com/academic/profile? id = 8ae14e824448e1977e934940ef349352&encoded = 0&v = paper_ preview&mkt = zh - cn#.

20 Cunningham H, Maynard D, Bontcheva K, et al. GATE: an Architecture for Development of Robust HLT applications [EB/OL]. [2019 - 12 - 20]. http://apps.webofknowledge.com/full_ record.do? product = WOS&search_mode = GeneralSearch&qid = 1&SID = 5DhuHpaQa8MNzHtWYaz&page = 1&doc = 1.

21 Salgado D, Krallingger M, Depaule M, et al. MyMiner: a web application for computer - assisted biocuration and text annotation [J]. Bilinformatics, 2012, 28 (17): 2285 - 2287.

22 Tao C, Song D, Sharma D, et al. Semantator: semantic annotator for converting biomedical text to linked data [J]. Journal of Biomedical Informatics, 2013, 46 (5): 882 - 893.

23 Stenetorp P, Pyysalo S, Topic G, et al. Brat: a web - based tool for NLP - assisted text annotation [EB/OL]. [2019 - 12 - 20]. http://cn.bing.com/academic/profile? id = 7a3599adf5dfb6d9c1cdbc0f69034166&encoded = 0&v = paper_ preview&mkt = zh - cn.

24 Neves M, Leser U. A Survey on Annotation Tools for the Biomedical Literature [J]. Briefings in Bioinformatics, 2014, 15 (2): 327 - 340.

25 杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建 [J]. 软件学报, 2016, 27 (11): 2725 - 2746.

26 《药学与临床研究》编辑部. 如何正确运用组内相关系数进行一致性检验——药物研究中的统计学 (一) [J]. 药学与临床研究, 2018, 26 (1): 7 - 8.

27 Cheng F, Zhou Y, Gao J, et al. Efficient Optimization of - Measure with Cost - Sensitive SVM [EB/OL]. [2019 - 12 - 20]. http://apps.webofknowledge.com/full_record.do? product = UA&search_mode = GeneralSearch&qid = 5&SID = 5DhuHpaQa8MNzHtWYaz&page = 1&doc = 1.

28 Hennig C. Cluster - wise Assessment of Cluster Stability [J]. Computational Statistocs & Data Analysis, 2007, 52 (1): 258 - 271.

《医学信息学杂志》开通微信公众号

《医学信息学杂志》微信公众号现已开通，作者可通过该平台查阅稿件状态；读者可阅览当期最新内容、过刊等；同时提供国内外最新医学信息研究动态、发展前沿等，搭建编者、作者、读者之间沟通、交流的平台。可在微信添加中找到公众号，输入“医学信息学杂志”进行确认，也可扫描右侧二维码添加，敬请关注！

《医学信息学杂志》编辑部