药物疾病语义关系语料库构建方法研究*

孙月萍 侯 丽 李 姣

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 以 BioCreative V CDR 语料库为例，选定现有的领域标注体系作为参照，通过数据质量控制工作流，采用半自动化标注方法实现药物疾病语义关系语料库的标准化构建，总结该语料库的质量及其应用。

〔关键词〕 生物医学；语义关系语料库；化学物质导致疾病关系

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2018.06.011

Study on the Building Method of Drug Disease Semantic Relationship Corpus *SUN Yue－ping, HOU Li, LI Jiao, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Taking the BioCreative V CDR corpus as an example, the paper takes the existing domain annotation system as the reference; through the data quality control workflow, the standardized building of drug disease semantic relationship corpus is realized by adopting the semi－automatic annotation method, and summarizes its quality and application.

〔**Keywords**〕 Biomedical; Semantic relation corpus; Chemical－Induced Disease (CID) relation

1 引言

在生物医学领域，包含实体以及实体语义关系的语料对生物医学文本挖掘研究以及文本挖掘工具评价具有极其重要的意义[1]。目前公开评测提供的标注语料常为百篇或千篇级文本，数据规模偏小[2]。在诸多实体语义关系中，化学物质与疾病之间的关系（Chemical Disease Relation，CDR）尤其是药物副作用关系是重点研究目标之一[3]。药物副作用识别问题在特定研究领域中可以转化为化学物质导致疾病（Chemical－Induced Disease，CID）关系抽取问题。如果某化学物质和某疾病之间存在 CID 关系，表示化学物质的使用导致该疾病，且化学物质并非用于该疾病的治疗。

据调研可经受计算方法检验的用于药物副作用检测的公共资源有限[4]。为研究生物科学文献中的语义关系，Rosario 等[5]提供从 Medline 2001 标题和摘要中抽取的句子，定义包括药物副作用在内的 8 类关系，但该数据集以句子为单元，而现实中的语义关系发现可能涉及到多个句子构成的语义单元。毒性与基因比较数据库（Comparative Toxicogenomics Database，CTD）从科学文献中手工标注化学物质与疾病之间的生物标记/机制关系以及治疗关系[3]。

〔修回日期〕 2018－04－04

〔作者简介〕 孙月萍，助理研究员；通讯作者：李姣，副研究员。

〔基金项目〕 国家重点研发计划“精准医学本体和语义网络构建”（项目编号：2016YFC0901901）；国家自然科学基金委“基于深层网络的药物基因组信息整合与药效预测研究”（项目编号：81601573）；国家社科基金项目“面向知识服务的公众健康知识组织体系构建研究”（项目编号：14BTQ032）；医学融合出版知识技术重点实验室－医学知识服务的关键技术研究项目。

然而该语料库并没有同时提供疾病实体和药物实体标注信息，不能直接用于现有的有监督学习方法。为此 BioCreative V 专为 CID 关系在 CTD 标注的基础上标注 1 500 篇 PubMed 文献，提供疾病实体、化学物质实体以及 CID 关系标注信息，为业界科研人员通过机器学习算法提高关系抽取性能服务。为方便研究，本文对化学物质和药物不作明确区分，文中提及的化学物质，可作为药物理解。作为该评测任务的合作方，本研究团队与美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）团队合作，共同完成了化学物质疾病语义关系评测任务（BioCreative V CDR 任务，简称 CDR 任务）语料库的构建工作，以下内容就该语料库的标注规范、方法和应用加以总结。

2 药物疾病语义关系语料库标注规范

2.1 任务

研究科学文献中的药物副作用发现，为进一步提升基于有监督机器学习算法的 CDR 抽取性能，就 BioCreative V CDR 评测任务提供语料支持，包括疾病、化学物质、化学物质导致疾病关系标注信息，根据评测任务时间节点分别提供训练集、发展集和测试集。

2.2 方式

候选语料标注方式包括领域专家、众包和团体标注。其中团体标注方法在不依赖于专家的情况下能构建高质量的语料[6]。参考 2009 年 I2B2 药品信息抽取评测的语料构建工作[7]，考虑到标注任务需要在固定时间段内构建一定规模的高质量语料，且团队具有组织经验，故选择团体标注方式。

2.3 团队

根据标注任务将标注团队分为实体和实体关系标注团队。其中实体标注团队为中国医学科学院医学信息研究所团队，由 4 位具有医学背景的医学信息学相关专业研究生和 1 位研究人员组成。根据标注规范，采用多轮迭代的模式分组进行标注，每轮标注结果由高级标注员进行校对。实体关系标注团队由 3 位有关系标注经验的 CTD 标注人员组成。

2.4 范畴

所有的术语都要求对应到《医学主题词表》（Medical Subject Headings，MeSH）的具体概念标识并在文中标注出现位置信息。CID 关系标注通过 MeSH 概念标识对表示。（1）疾病/病症术语。考虑到药物副作用的特点，按照 MeSH 2015 疾病分支标注，包括体征和症状。（2）化学物质术语。选取 MeSH 2015 药物和化学物质分支。将“化学物质必须有明确的化学结构”作为标注的第 1 基本原则。（3）化学物质导致疾病关系。化学物质和疾病分别对应的 MeSH 概念标识对。当标注结果有分歧时，按照针对美国国立生物技术信息中心（National Center for Biotechnology Information，NCBI）疾病数据集设计的已有 NCBI 疾病标注规范[8]以及相关概念规范[9]标注疾病，按照构建 CHEMDNER 数据集的化学物质标注规范[10]标注化学物质。所有的关系抽取任务均基于实体标注结果。每条实体标注包括 1 个类型（如疾病或药物）和 1 对起始 - 终止位置，要求仅标注连续的字符串。实体标注要求同一类型的实体位置范围不能交叉或重叠，但对位置范围的覆盖不做限定，任务中不标注被覆盖的短实体标注。

2.5 文档选取

为在最短时间内提供高质量的标注语料，CDR 语料库主要从 CTD - Pfizer 中抽取文档。其中 500 篇训练集、500 篇发展集和大部分测试集（约 400 篇）均随机从 CTD - Pfizer 中选取。补充文档选取条件包括：（1）英文文章。（2）具有摘要信息。（3）2014 年或其以后出版。（4）具有至少 1 个 CID 关系。（5）与训练集和测试集文档在词分布上具有一定相似度。根据词分布相似度选取的具体方法[11]为：使用 PubMed Eutilities 工具（https://www.ncbi.nlm.nih.gov/books/NBK25501/）抽取训练集和发展集中任意 1 篇文档（DocT）的相关文档（DocR），使用第 1 - 3 项筛选条件筛选后计算每篇相关文档 DocR 与数据集中文

档的相似度总和，同 DocT 与数据集中其他文档的相似度总和比较，选择最接近的 DocR 并进一步判断其是否存在 CID 关系。该方法确保测试集与训练集、发展集的相关性和平衡性。

3　基于 PubTator 工具的半自动化标注方法

本研究中的语料标注采取在线标注方法，基于 NCBI 团队构建的 PubTator 工具进行半自动化标注。CDR 语料构建实施流程，见图 1。评测组织方设计评测任务后，按照评测要求，参考相关标注规范制定标注规范并收集待标注语料备用。随机挑选部分文档（50 篇）导入 PubTator 平台，分组开展预标注工作，两组预标注工作结束后由在线标注一致性检验工具用不同颜色标识标注结果，根据标注反馈结果进一步细化或修订标注规范。经过预标注后分组进行正式标注工作，每标注一批文档通过标注质量衡量指标（Inter Annotator Agreement，IAA）等计算标注一致程度。标注不一致的问题由高级标注员进行最终版本修订，导出标注结果。标注结果在发布前还需经过整体检验，主要通过统计分析方法纠正明显的标注错误，结果校验后由评测平台分训练集、发展集和测试集发布。在标注流程中涉及两种可视化半自动化辅助标注工具，其中预标注、正式标注都由 PubTator 工具在线完成，标注一致性校验基于由 Wei 开发的 PubTatorComparison 工具完成。PubTator 中预先内置疾病实体标注算法 DNorm 和药物实体识别算法 CHEMDNER。PubTatorComparison 工具对比 PubTator 导出的标注结果，以不同颜色分别展示。此外在整体校验步骤中，为查验是否存在明显的标注错误，加入 1 套程序对标注结果进行统计分析，重点对同一个实体指称且标注为不同概念 ID 的情况进行离群点检验分析。

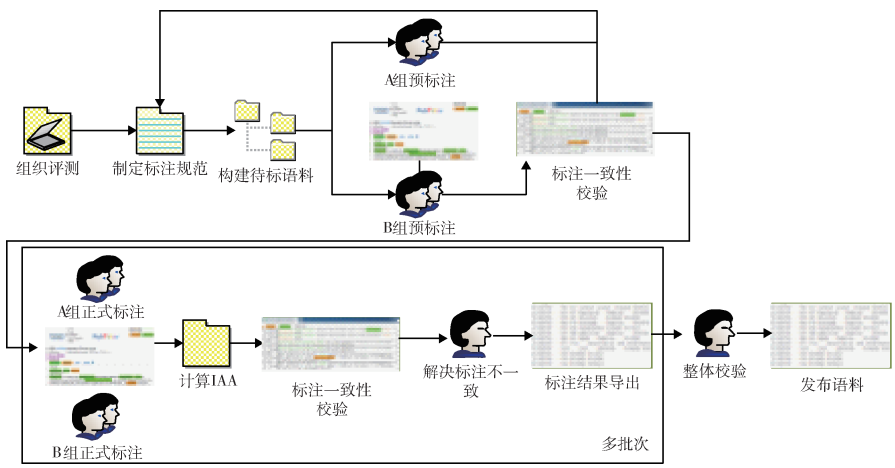

图 1　CDR 语料构建实施流程

4　药物疾病语义关系语料库标注结果和应用

4.1　标注结果

CDR 语料库统计信息，见表 1。从整体看，训练集、发展集和测试集的疾病实体规模、化学物质实体规模和实体关系规模相差不大，符合评测任务的预期。为检验语料的均衡性，本研究对数据集在样本空间的重叠问题进行分析。据统计疾病实体标注涵盖 2 920 个概念 ID，化学物质实体标注涵盖 2 144个概念 ID，关系标注涵盖 2 434 个 CID 关系。少量疾病实体、化学物质实体与 CID 关系至少出现在 1 个数据集中。各数据集中包含的唯一概念数量包括疾病实体概念数量、化学物质概念数量和 CID 关系数量可以准确反映各数据集的重叠情况。CDR 语料数据集样本（唯一概念）重叠分析，见表 2，疾病实体疾病和化学物质的重叠率约为 30% ~ 40%，而 CID 关系的重叠率较低（<15%）。可以看出样本重叠问题在 CDR 语料库中并不显著。标注一致性为反映数据质量的直接指标，考虑到关系标

注的特殊性，评测仅考察实体标注的标注一致性。CDR 标注一致性结果，见表 3。可见 CDR 语料库标注一致性较高，尤其是化学物质标注，一方面说明语料库的标注质量较高，另一方面也对前期标注规范制定工作做出肯定。

表 1 CDR 语料库统计信息概览

数据集	文献量（篇）	疾病		化学物质		CID	NCID
		指称	ID	指称	ID	关系	关系
训练集	500	4 182	1 965	5 203	1 467	1 038	4 394
发展集	500	4 244	1 865	5 347	1 507	1 012	4 249
测试集	500	4 424	1 988	5 385	1 435	1 066	4 343
合计	1 500	12 850	5 818	15 932	4 409	3 116	12 986

表 2 CDR 语料数据集样本（唯一概念）重叠分析

数据项	疾病	化学物质	CID
训练集	1 384	1 007	928
发展集	1 254	985	887
测试集	1 337	1 018	941
训练集∩发展集（%）	31.50	37.03	13.39
训练集∩测试集（%）	33.82	35.36	12.86
发展集∩测试集（%）	32.95	38.21	14.45
训练集∩发展集∩测试集（%）	22.04	25.54	6.48

注：重叠比率 = 重叠数/最大概念数。

表 3 CDR 标注一致性结果（*F* 值）

数据集	疾病	化学物质	计算所用文档数量
训练集	0.918 6	0.965 5	417
发展集	0.919 8	0.967 3	500
测试集	0.912 6	0.967 3	498
总数据集	0.916 8	0.966 8	1 415

4.2 应用

共有来自 12 个国家的 34 个团队参加 BioCreative V CDR 评测。在 CID 任务中评测的前两名均使用支持向量机（Support Vector Machine，SVM）分类器。多数评测文章认为语义关系抽取任务仍是颇具挑战性的任务，尤其是对于实体分别出现在不同句子中的“跨句语义关系”判别。评测举办后陆续有研究使用该评测语料开展相关关系抽取工作。据 Web of Science 统计截至 2017 年 12 月 31 日发表 15 篇文章（14 篇发表于《数据库》（*Database*）），其中 CID 关系抽取文章有 8 篇。按照基于统计机器学习和模式学习模型分类方式对该 8 篇文章进行概括，CID 关系抽取方法总结，见表 4。大部分方法基于机器学习模型，利用分类器或分类器组合解决 CID 关系抽取问题，部分工作基于模式学习。总结表明基于机器学习的模型是当前较为流行的关系抽取方法，很多工作基于多个层级的分类器，使用多维度特征包括语言学、知识库以及统计特征等。在评测后，新增部分结合神经网络模型的工作，或基于神经网络模型生成语义表示[14]，或基于神经网络模型进行分类[18]，且取得不错的效果。其中知识库特征所使用的知识库有一定差异，包括 MeSH、CTD、MEDI、SIDER 等。知识库特征提高模型的召回率的同时在一定程度上降低模型的准确率，这也侧面说明目前尚没有与文献语料准确对应的知识库，可见药物疾病关系语料库构建的必要性。

表 4 CID 关系抽取方法总结

分类方法	方法介绍
基于机器学习的模型	基于 SVM 的关系抽取模型[12]
	结合共指消解模块，基于 SVM 的关系抽取模型[13]
	结合基于词汇特征的模型、基于句法结构特征的树核模型和生成语义表示的神经网络模型[14]
	结合了句子层级的分类器与文档层级的分类器，并结合外部语料库（标注语料，未标语料）训练模型[15-17]
	结合了句子层级的分类器与文档层级的分类器。句子层级使用卷积神经网络模型，文档层级使用最大熵模型[18]
基于模式学习的模型	句子层级的模型。基于形式化模板的关系抽取方法[19]

根据 BioCreative V CDR 评测总结[11]，CID 关系取得的最佳 *F* 值 57.03% 达到较高的关系抽取水平。综合所有参赛队伍的评测结果，最佳 *F* 值为 62.80%，后期公开发表的所有研究成果的 *F* 值均没有超过 70%，说明 CID 关系抽取任务具有很大的挑战性，关系抽取模型仍有较大的提升空间。应用情况表明主流 CID 关系抽取方法为基于机器学习的模型，CDR 语料库为 CID 关系抽取算法改进提供有效的有监督学习语料。另外 CDR 语料库对出现在同一句话中的 CID 关系抽取支持度较高，但对跨句子的 CID 关系抽取支持度偏低，反映出目前关系抽取语料库面临的困境：基于篇章理解的关系抽取语料专业性较强，标注过程中主观性较高，较难形成统一的标准。

5 结论

本文详细介绍基于竞赛评测任务 BioCreative V CDR 的语料库构建工作，包括标注任务定义规范制定、文档选取、标注团队，重点介绍基于 PubTator 工具的半自动化标注方法以及标注结果和应用。主要的经验和启示如下：（1）在调研工作的基础上根据标注任务所要求的标注目标、数据规模、限期，选择合理的语料标注方式。（2）选定现有的标注体系作为参照，结合评测任务的实际需求制定标注体系，定义标注范围。（3）多轮预标注的方法有助于最终标注规范的制定。将数据集分成多批次标注，在每轮标注后对标注结果进行阶段性总结，有助于提高标注质量。（4）基于机器学习方法构建的半自动化标注方法可极大提高标注效率和质量。其中除半自动化标注工具外，还应有标注结果比较工具、全局校验工具等。（5）在文档选取阶段重视数据集的不平衡问题，应使测试集中的文档与训练集、发展集中的文档具有一定相关性和独立性，基于文档的词分布选取数据集中的文档。（6）数据集的标注质量可通过标注一致性指标（IAA，*F* 值等）进行评估，数据集的平衡性可通过数据样本重叠分析进行评估。

药物疾病语义关系抽取在业界仍是具有挑战性的任务，且尚没有与文献语料准确对应的知识库，药物疾病关系语料库构建具有较强的研究和应用意义。在语料库构建实践过程中，还存在有待改进的问题。例如标注人对其标注的信息具有不同的确信值，该确信值有助于区分标注信息是否经过人工推理，可在语料库构建过程中记录标注信息的确信值以供后期参考。另外文档选取对于整个数据集非常重要，有必要从整体出发，选定主题，按照主题分布对数据集进行统一规划。此外考虑到目前的半自动化标注工具中仅对实体进行预标注，没有对关系进行预标注，且没有有效措施降低关系标注的漏标问题，可以通过半自动化关系标注的辅助功能开发为标注者创造更为人性化的关系标注环境，努力降低主观性带来的影响。

参考文献

1 Neves M. An Analysis on the Entity Annotations in Biological Corpora [J]. F1000Res, 2014, (3): 96.

2 李芳，刘胜宇，刘峥. 生物医学语义关系抽取方法综述[J]. 图书馆论坛，2017，(6)：61-69.

3 Davis AP, Wiegers TC, Roberts PM, et al. A CTD-Pfizer Collaboration: manual curation of 88, 000 scientific articles text mined for drug-disease and drug-phenotype interactions [J/OL]. (2013-11-28) [2017-06-02]. https://doi.org/10.1093/database/bat080.

4 Li J, Sun Y, Johnson RJ, et al. BioCreative V CDR Task Corpus: a resource for chemical disease relation extraction [J/OL]. (2016 - 05 - 08) [2017 - 06 - 02]. https://doi.org/10.1093/database/baw068.

5 Rosario B, Marti AH. Classifying Semantic Relations in Bioscience Rext [C]. Barcelona: ACL'04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004: 430 - 437.

6 Xia F, Yetisgen - Yildiz M. Clinical Corpus Annotation: Challenges and strategies [C]. Istanbul: Proceedings of Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2012) of the International Conference on Language Resources and Evaluation (LREC), 2012: 32 - 39.

7 Uzuner Ö, Solti I, Xia F, et al. Community Annotation Experiment for Ground Truth Generation for the i2b2 Medication Challenge [J]. Journal of the American Medical Informatics Association, 2010, 17 (5): 519 - 523.

8 Doğan RI, Lu Z. An Improved Corpus of Disease Mentions in PubMed Citations [C]. Montreal: BioNLP 12 Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 2012: 91 - 99.

9 Doğan RI, Leaman R, Lu Z. NCBI Disease Corpus: a resource for disease name recognition and concept normalization [J]. Journal of Biomedical Informatics, 2014, (47): 1 - 10.

10 Krallinger M, Rabal O, Leitner F, et al. The CHEMDNER Corpus of Chemicals and Drugs and its Annotation Principles [J]. Journal of Cheminformatics, 2015, 7 (Suppl 1): S2.

11 Wei CH, Peng Y, Robert L, et al. Assessing the State of the Art in Biomedical Relation Extraction: overview of the BioCreative V chemical - disease relation (CDR) task [J/OL]. (2016 - 03 - 19) [2017 - 07 - 03]. https://doi.org/10.1093/database/baw032.

12 Pons E, Becker B, Akhondi SA, et al. Extraction of Chemical - induced Diseases Using Prior Knowledge and Textual Information [J/OL]. (2016 - 04 - 14) [2017 - 08 - 03]. https://doi.org/10.1093/database/baw046.

13 Le H, Tran M, Dang TH, et al. Sieve - based Coreference Resolution Enhances Semi - supervised Learning Model for Chemical - induced Disease Relation Extraction [J/OL]. (2016 - 07 - 26) [2017 - 08 - 03]. https://doi.org/10.1093/database/baw102.

14 Zhou H, Deng H, Chen L, et al. Exploiting Syntactic and Semantics Information for Chemical - disease Relation Extraction [J/OL]. (2016 - 04 - 14) [2017 - 07 - 30]. https://doi.org/10.1093/database/baw048.

15 Xu J, Wu Y, Zhang Y, et al. CD - REST: a system for extracting chemical - induced disease relation in literature [J/OL]. (2016 - 03 - 25) [2017 - 07 - 30]. https://doi.org/10.1093/database/baw036.

16 Li ZH, Yang ZH, Lin HF, et al. CIDExtractor: a chemical - induced disease relation extraction system for biomedical literature [C]. Shenzhen: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016: 994 - 1001.

17 Li H, Tang B, Chen Q, et al. HITSZ_ CDR: an end - to - end chemical and disease relation extraction system for BioCreative V [J/OL]. (2016 - 06 - 05) [2017 - 08 - 03]. https://doi.org/10.1093/database/baw077.

18 Gu J, Sun F, Qian L, et al. Chemical - induced Disease Relation Extraction Via Convolutional Neural Network [J/OL]. (2017 - 04 - 02) [2017 - 08 - 03]. https://doi.org/10.1093/database/bax024.

19 Lowe DM, O'Boyle NM, Sayle RA. Efficient Chemical - Disease Identification and Relationship Extraction Using Wikipedia to Improve Recall [J/OL]. (2016 - 04 - 08) [2017 - 08 - 03]. https://doi.org/10.1093/database/baw039.