• 医学信息组织与利用 •

# 国外生物医学文本语料库分类及特点研究*

晏归来 安新颖 范少萍 周永称

（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〕 通过梳理国外 31 个生物医学文本语料库标注内容，根据语料库标注实体类型，参照 UMLS 语义类型将其划分为 6 大类。总结语料库在语义类型、数据源等方面特点，阐述生物医学文本语料库构建流程及关键步骤，以期为我国生物医学文本语料库相关研究奠定基础。

〔关键词〕 生物医学文本语料库；语义类型；语义关系

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2018.10.017

**Study on the Categories and Characteristics of Overseas Biomedical Text Corpuses** *YAN Guilai, AN Xinying, FAN Shaoping, ZHOU Yongcheng, Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China*

〔**Abstract**〕 The paper divides the corpus into six categories by analyzing annotated contents of the 31 overseas biomedical text corpuses and referring to UMLS semantic type according to the annotated entity types of the corpuses. It summarizes characteristics of the corpus in the aspects like semantic type and data source, expatiates on the building process and major steps of biomedical text corpus in the hope of laying down the foundation based on which related studies on China's biomedical text corpuses will be carried out.

〔**Keywords**〕 biomedical text corpus; semantic type; semantic relations

## 1 引言

随着大数据时代的来临，各类生物医学文本资源呈爆发式增长，大量生物医学知识以半结构化或非结构化的形式存储于各类文本资源中。有效利用各类文本资源中蕴含的知识对生物医学研究具有重要意义。传统的文献分析方法难以处理如此庞大的数据，无法从海量文本中直接提取研究人员感兴趣的事实信息，因此开发生物医学文本挖掘工具具有重要意义。

语料库是一定量计算机可读文本的集合，取样的文本在最大程度上代表一种语言或变体[1]。在生物医学领域，语料库是开发新文本挖掘算法或提高算法精确性的关键要素，以支持命名实体或关系抽取研究[2]。目前国际主流的生物医学语料库测评会

〔修回日期〕 2018－09－13

〔作者简介〕 晏归来，硕士研究生；通讯作者：安新颖，副研究员，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）。

议，如 TREC Genomics Track、BioNLP 等，组建自身语料库以测试算法有效性，加强算法的可比性。一些专门针对语料库质量测评的会议，如 BioCreative 等，更是关注语料库的规模与质量。鉴于语料库在机器学习算法测试中的重要作用，本文较为详细地梳理国际生物医学自然语言处理研究中最常使用的 31 个生物医学文本语料库，对其标注的语义内容及语料库特点进行深入剖析，以期帮助研究人员理解语料库内容，选择适宜训练语料集。

## 2 生物医学文本语料库分类

### 2.1 语料库概述

通过查阅文献、调研国际生物医学文本挖掘算法测评会议及学术会议发布的开源语料库共 31 个：GENIA[3]、GENETAG[4]、PhenoCHF[5]、AIMed Corpora[6]、Bioinfer[7]、LLL[8]、Bioscope[9]、EU - ADR[10]、DDI[11]、CRAFT[12]、CHEMDNER[13]、Chemdner patents CDP corpus[14]、CellFinder[15]、NCBI Disease[16]、SNPCorpus[17]、The SPECIES and ORGANISMS Resource[18]、Biotext[19]、BioText NC Semantics Dataset[20]、ADE corpus[21]、Corpus for Disease Names and、Adverse Effects[22]、ChemProt corpus[23]、Chemdner patents GPRO corpus[24]、CDR corpus[25]、CEMP Corpus[26]、MutationFinder Corpora[27]、Nagel Corpus[28]、Protein Residue Full Text Corpus[29]、Protein Residue Relations Silver Corpus[30]、PICorpus[31]、The Anaphora Corpus[32]、TestSuite Corpora[33]。根据语料库标注实体与关系类型，本文参照一体化医学语言系统（Unified Medical Language System，UMLS）语义类型对上述 31 个语料库进行分类汇总，可划分为 6 个大类，见图 1。一是疾病，语义类型为“Disorders”（DISO），主要包括 NCBI Disease、Biotext 语料库。二是基因/蛋白质，语义类型为“Genes & Molecular Sequences”（GENE），以及 T116（Amino Acid，Peptide，or Protein）和 T114（Nucleic Acid，Nucleoside，or Nucleotide），主要包括 GENETAG、Bioinfer 等语料库。三是序列/突变，语义类型为 T045（Genetic Function），主要包括 MutationFinder 和 SNP Corpus。四是化合物/药物，语义类型为“Chemicals & Drugs”（CHEM），除上述 T116 和 T114 外，其代表性语料库主要有 DDI、CHEMDNER。五是物种，语义类型为“Living beings”（LIVB），如 Species 语料库。六是细胞/解剖部位，语义类型为“Anatomy”（ANAT），如 PhenoCHF、GENIA 等。除上述根据语料库标注语义类型进行划分外，根据语料库是否标注语义关系还可分为包含语义关系的语料库和不包括语义关系的语料库。AIMed Corpora、Bioscope、EU - ADR、DDI、Chemdner Patents CDP Corpus、BioText NC Semantics Dataset、Protein Residue Relations Silver Corpus、PICorpus、The Anaphora Corpus、TestSuite Corpora 标注实体之间的语义关系，其中 Bioscope 和 The Anaphora Corpus 重点剖析生物医学研究领域不同实体之间的语义环境、句法构成及语义关系。

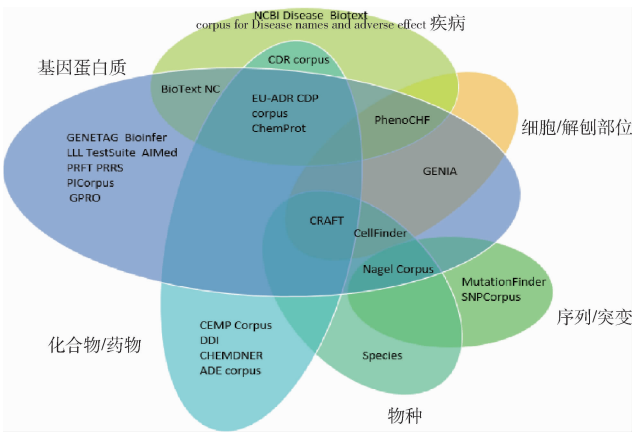


图 1 生物医学文本语料库分类

### 2.2 基因/蛋白质语料库

基因与蛋白质之间的相互作用关系是生物学研究的重要内容，也是解决大量医学难题的关键信息。基因、蛋白质是语料库研究的主要内容，上述语料库中共有 18 个语料库对基因或蛋白质进行标注。通过挖掘基因、蛋白质及其相互作用关系将极大地有利于疾病诊断、药物设计，促进相关生物医学研究的进展。基因和蛋白质相关测评任务（Gene and Protein Related Object Task，GPRO）利用基于网

络的文本挖掘工具——PubTator 重新标注两个现有的基因语料库：Te BioCreative II GN 和 Te Citation GIA test collection 语料库，在这两个语料库所标注人类基因的基础上添加基因家族和蛋白质结构域标注内容[24]。

### 2.3 疾病语料库

应用生物医学领域文本挖掘、分析等方法准确、高效地识别出关键的疾病和化学药物信息，经过规范化的分析处理来直观、可视化地展示结果，从而极大方便相关研究人员发现有价值的疾病、药物以及基因之间的关系[34]。NCBI 疾病语料库对疾病名称及其概念进行充分标注，语料库公开发布的 6 892 个疾病实体能被映射到 790 个疾病概念，相应的概念参考《医学主题词表》（Medical Subject Headings，MeSH）或在线人类孟德尔遗传（Online Mendelian Inheritance in Man，OMIM）概念标识，语料库采用人工标注的方式取得较高的标注人员间一致性（Inter – annotator Agreement，IAA）[16]。

### 2.4 序列/突变语料库

近年来从生物医学文献中挖掘基因突变或变异的位点受到广泛关注，在变异或突变位点上研究疾病致病机制具有至关重要的作用。MutationFinder 语料库标注 508 篇 Medline 摘要中的突变信息，该语料库被用于基于正则表达的同名工具评估任务[27]。单核苷酸多态性（Single Nucleotide Polymorphisms，SNP）实体标准化和标注较其他实体更为困难，SNP Corpus 共涉及 527 个 SNP 实体，该语料库构建的局限性包括以下几点：参考的标注资源仅包括 ENSEMA 和 EntrezGene、缺少完整的参考序列、缺少将 SNP 转化为命名实体的计算机辅助工具、测序错误导致的实质性差异[17]。

### 2.5 化合物/药物语料库

从海量生物医学研究文献中获得药物以及相互作用关系是目前生物信息研究热点之一。主要包括药物命名识别、药物靶标发现、药物疗效评价和药物不良反应现象等[35]。DDI 语料库是该研究领域最具代表性的语料库之一，目的是为文本中药物和药物相互关系研究提供金标准的标注语料。标注文本来源主要为 Medline 以及 DrugBank。DDI 使用 MetaMap 进行句子切分与药物命名实体识别，再由两位标注人员进行审核与标注，语料库药理性质相关概念参考 WHO ATC、Drug@ FDA、EMA、CIMA、PubChem、MeSH、WHO ATCvet、DrugBank 等多个药物相关信息资源[11]。

### 2.6 物种语料库

在文本中命名生物实体的识别文本中物种和其他种群名称是较为核心和相对困难的任务。为覆盖更多的物种类型，Species 语料库筛选包含 8 种实体（细菌学、植物学、昆虫学、医学、真菌学、预防学、病毒学和动物学），共计 100 篇文献摘要进行标注。ORGANISMS 语料库包括来自许多种群相关的期刊摘要，基于 Medline 文献数据库有机体名称进行开发[18]。

### 2.7 细胞/解剖部位语料库

细胞和解剖学信息作为生物医学研究领域的重要组成部分，对其进行分析和挖掘长久以来受到领域内研究人员高度关注。PhenoCHF 是表型领域首个语料库，旨在编码详细的表型信息，语料库由领域内专家根据详细的表型概念以及疾病 – 表型关系手动标注完成[5]。Cellfinder 基于 CELDA 本体，整合多来源研究文献和微阵列数据衍生数据所构建的干细胞数据集，内容涵盖解剖结构、细胞成分、细胞系、细胞类型、基因/蛋白、物种[15]。

## 3 生物医学文本语料库特点

### 3.1 标注类型丰富多样

随着自然语言处理、机器学习、生物信息学等领域技术的逐渐发展，传统单一语义类型的语料库难以满足复杂文本挖掘算法的学习，标注多种语义类型和复杂关系的语料库应运而生。如 GENIA 标注包括化学、生物学、解剖学在内的多种实体类型，是生物医学文本挖掘领域内金标准语料库，不少语

料库的构建均参与其标准规范或是将其作为算法的测试集或训练集。CRAFT 主要标注细胞、蛋白质、序列、物种、基因、化合物 6 类实体，基本涵盖当前研究领域最常见的各种实体。

### 3.2 语义关系标注日益完善

近年来生物医学文本语料库越发重视不同语义类型之间关系的构建，生物医学语义关系抽取能揭示疾病、药物、蛋白质、基因等重要生物医学实体之间的语义关系（如治疗、诊断、靶向等关系），是构建领域知识图谱、本体与知识库，实现知识发现，完善临床决策支持系统的重要基础，以进一步助力智慧医疗与精准医学，具有重要现实意义[36]。上述语料库中 Bioscope 和 The Anaphora Corpus 着重于语义关系的研究，Bioscope 分析生物医学文本中否定关系、推断关系以及语言范围。此外 DDI 语料库主要标注不同药物、化合物之间的关系，AIMed Corpus 挖掘了蛋白质相互作用关系，ADE Corpus 揭示药物与不良反应、剂量之间的关系。

### 3.3 标注数据源多源化

资料库标准数据源可划分为句子、文摘、生物医学全文本、专利文本和电子健康档案（Electronic Health Records，EHR）等 5 种，不同类型语料库根据其构建目的各有特色。句子语料库规模较少但内容精确，标注句子的选择和人工标注可控性较强，内容可以做到十分精准，易形成领域内金标准语料库，如 GeneTag、Bioinfer、Bioscope 等。全文本语料库包含较多领域专业术语，对标注人员专业素养要求更高，人工标注的难度较大，但包含丰富的语义类型和语义关系，如 CRAFT、CellFinder 等。文摘语料库人工标注的难度小于全文本，语义类型、语义关系也较为丰富。此类语料库在规模上也较适合当前机器学习研究需求，是当前文本语料库构建最为主流的源数据，其代表性语料库包括 GENIA、EU－ADR、DDI 等。专利文本语义类型丰富，内容专指性高，标注难度大，是一种非常重要的技术资料，书写格式和表达方式较为固定，与普通文献相比专利文本用词规范严谨、歧义较少但语句长度普遍长于普通文本，领域知识专业性很强，标注难度较大[37]。专利文本语料库领域专指性强、语义类型丰富，代表性语料库有 CHEMDNER 系列语料库。电子健康档案主要包含疾病、表型相关语义类型，内容规范化程度较高。是最基本、最重要的医疗卫生基础资料，文本结构较为规范，包含丰富的专业术语和实体间关系，可通过机器标注结合人工标注的方法进行语料库构建，提高标注效率。电子健康档案正逐渐被更多的国外内学者作为语料库构建数据源[38]，但其可获得性不及上述 4 类数据源，代表语料库有 PhenoCHF。

### 3.4 相关研究

目前生物医学文本语料库在多个研究领域得到广泛应用，其中主要的研究方向有命名实体识别、关系抽取、知识发现等。Saber A Akhondi 等人在构建的 CEMP 专利文本语料库基础上训练 tmChem 算法，将 $F$ 值提升至 86.82%，取得 94.23% 的精确度[26]。在生物医学命名实体识别的基础上，Harsha Gurulingappa 等人利用 ADE 语料库训练疾病－药物不良反应相互作用关系自动抽取算法，取得 64% 的召回率和 75% 的精确度[21]。目前各国均在精准医学研究领域开展深入的挖掘和分析，精准医学实现的基础是实现多来源数据的整合与分析。在标准的语义网络，整合生物医学本体和多类型文本资源，利用生物医学文本语料库训练自然语言处理算法，能帮助科研人员从海量信息中高效准确地找到相关知识开展研究，助力临床医生通过诊断结果精准地判断疾病类型、寻找最佳治疗方案。

## 4 生物医学文本语料库构建方法及启示

### 4.1 概述

国外生物医学文本语料库构建工作起步早、方法成熟，其语料库标注体系、标注方法、标注工具、一致性检验及应用场景对我国语料库构建及自然语言处理研究领域具有重要指导意义。本文在调研 31 个生物医学文本语料库标注内容、数据源类型、语料库规模的基础之上分析语料库构建流程，

总结语料库构建关键步骤，以期对我国生物医学文本语料库相关研究有所启发。

### 4.2 构建流程

生物医学文本类型多样且数量庞大，语义类型和语义关系标注难度较大，涉及专业的生物医学领域知识且需要对语义类型和关系进行明确的定义和分类，本文通过调研国外 31 个生物医学文本语料库构建流程，总结语料库构建流程。见图 2。生物医学文本语料库构建主要分为 4 个步骤，即标注文本筛选、标注规范制定、语料库标注及一致性检验。首先，根据语料库标注内容选择主题相关的标注文本，可根据影响因子、发表年限在内的多种文本外部信息进行筛选；其次，在分析标注文本特点、语义类型及语义关系基础上参考生物医学领域主题词表或本体构建标注规范；再次，选择或开发合适的标注工具并进行语料库标注；最后，对生物医学文本语料库进行一致性检验，根据检验结果调整标注规范并再次进行语料库标注，以期提高标注一致性。

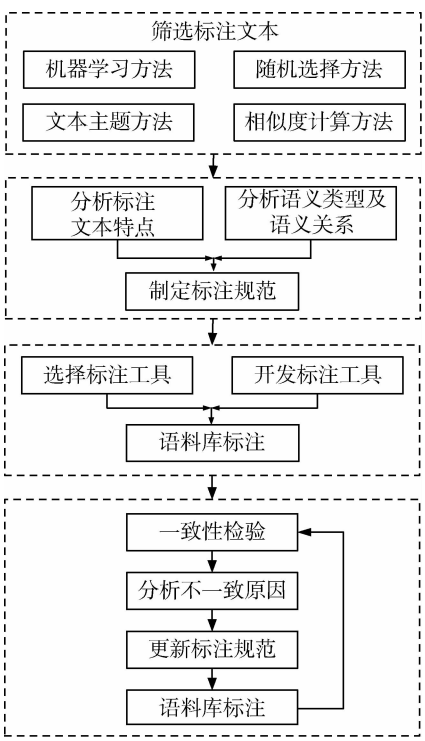


**图 2 生物医学文本语料库构建流程**

### 4.3 关键步骤

构建标注体系完整、语义类型及关系丰富、语料库体量较大且一致性较高的生物医学语料库难度较大，本文分析国外 31 个生物医学文本语料库，总结以下关键步骤。第一，筛选研究主题相关、语义类型及关系丰富的标注文本有助于提高语料库构建效率。常见的标注文本筛选方法有机器学习方法[39]、随机选择方法[11]、相似度计算方法[25]、文本主题方法[13]在内的多种方法。第二，构建标注规范参考体系在生物医学文本语料库构建中起到重要作用。通过选择或构建合适的领域词表、本体作为标注规范参考体系将极大地有利于准确定义蛋白质、基因、疾病、化合物等语义类型及各类语义关系，完善语料库标注规范。常见的词表或本体有 MeSH、GO、OMIM、UMLS 等。第三，制定可操作性、合理性标注规范是语料库标注不可缺少的环节。从修饰语、嵌入实体、同位语、标点符号、连接词等多方面撰写标注规则[40]；综合考量联合用药、靶向作用、否定关系等多种生物医学相互作用关系定义语义关系的相互作用域。第四，专业标注人员能对语料库标注起到事半功倍的作用。生物医学背景研究人员熟悉各类专业术语及生命活动调节机制，有助于提高语料库标注准确率和召回率。第五，标注一致性检测语料库建设质量。由于生物医学文本的复杂性与发展性以及各种标注方法的难点和缺陷，语料库标注的正确性与一致性较难保证。设计合理的语料库标注一致性检验方法，参考检验结果，在标注过程中不断调整标注规范及工具，以达到提高标注效率及一致性的目的。

### 4.4 对我国生物医学文本语料库构建的启示

生物医学文本语料库的构建对自然语言处理、知识图谱等研究具有重要意义，但其构建难度较大，涉及文本、标注工具筛选、标注规范制定及标注人员培训等方面。本文在调研国外语料库构建方法及流程的基础上总结其对我国生物医学文本语料库构建启示如下：采用机器标注（机器学习等）结合人工标注的方法提高标注效率、准确率及召回率；探索在全文及其他文本类型（专利、病历等）

上进行标注，以期获得更多、更专指的生物医学术语；构建更大规模、语义类型、语义关系更加丰富的语料库，对标国际金标准语料库，为数据挖掘提供更好的数据支撑；采取多轮预标注的方式逐步完善标注规范，正式标注时采用“背靠背”的标注形式，保证标注结果的一致性。

## 5 结语

生物医学文本语料库构建研究在近年来已取得较多成果，标注类型从单一类型逐渐扩展为多种类型并逐渐包含语义关系的标注。本文在调研 31 个生物医学文本语料库基础上根据语料库标注实体类型，参照 UMLS 语义类型对其进行分类，总结现有语料库的特点。本文的创新点有：充分调研国外 31 个生物医学文本语料库，参照 UMLS 语义类型进行划分，通过韦恩图直观展示各语料库所标注的语义类型，突出介绍各语义类型的代表性语料库；综合 31 个语料库构建方法，总结较为通用的生物医学文本语料库构建流程并梳理语料库构建的关键步骤。

由于生物医学文本语料库在医学文本挖掘与知识发现领域的重要作用，未来生物医学文本语料库研究可从以下方面做进一步探索，如加强语料库建设，整合更多生物医学研究领域词表或本体，借助众包、机构间协作等多种方式构建语义类型多样、语义关系丰富的生物医学文本语料库，为命名实体识别、关系抽取等研究打下坚实基础；探索不同数据源语料库构建方式，实现在全文本数据上进行语料库构建等。这些研究的开展将为不断扩展现有语料库标注类型与规模，提高相关机器学习算法的准确率与应用价值发挥重要作用。

## 参考文献


1 宋鸿彦，刘军，姚天昉，等. 汉语意见型主观性文本标注语料库的构建［J］. 中文信息学报，2009，23（2）：123－127.

2 Neves M. An Analysis on the Entity Annotations in Biological Corpora［J］. F1000 Research，2014（3）：96.

3 Kim JD，Ohta T，Tateisi Y，et al. GENIA Corpus－semantically Annotated Corpus for Bio－textmining［J］. Bioinformatics，2003，19（Suppl 1）：i180－i182.

4 Tanabe L，Xie N，Thom L H，et al. GENETAG：a tagged corpus for gene/protein named entity recognition［J］. BMC Bioinformatics，2005，6（1）：1.

5 Alnazzawi，Paul Thompson1，Riza Batista－Navarro，et al. Using Text Mining Techniques to Extract Phenotypic Information from the PhenoCHF Corpus［J］. BMC Medical Informatics and Decision Making，2015，15（Suppl 2）：S3.

6 Bunescu R，Ge R，Kate RJ，et al. Comparative Experiments on Learning Information Extractors for Proteins and Their Interactions［J］. Artif Intell Med，2005，33（2）：139－155.

7 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.

8 Nédellec C. Learning Language in Logic－genic Interaction Extraction Challenge［EB/OL］.［2018－10－11］. https：//www. cs. york. ac. uk/aig/lll/lll05/lll05－nedellec. pdf，80.

9 Vincze V，Szarvas G，Farkas R，et al. The BioScope Corpus：biomedical texts annotated for uncertainty，negation and their scopes［J］. BMC Bioinformatics，2008，9（11）：S9.

10 Erik M. van Mulligen，Annie Fourrier－Reglat，David Gurwitz，et al. The EU－ADR Corpus：annotated drugs，diseases，targets，and their relationships［J］. J Biomed Inform，2012，45（5）：879－884.

11 Herrero－Zazo M，Segura－Bedmar I，Martínez P，et al. The DDI Corpus：an annotated corpus with pharmacological substances and drug－drug interactions［J］. J Biomed Inform，2013，46（5）：914－20.

12 Bada M，Eckert M，Evans D，et al. Concept Annotation in the CRAFT Corpus［J］. BMC Bioinformatics，2012，13（1）：161.

13 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（1）：1－17.

14 Akhondi S A，Klenner A G，Tyrchan C，et al. Annotated Chemical Patent Corpus：a gold standard for text mining［J］. Plos One，2014，9（9）：e107477－e107477.

15 Neves M，Damaschun A，Kurtz A，et al. Annotating and Evaluating Text for Stem Cell research［C］. Istanbul：Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining（BioTxtM 2012）at Language Resources and Evaluation（LREC），2012：16－23.

16 Dog˘an RI，Lu Z. An Improved Corpus of Disease Mentions in Pubmed Citations［C］. Stroudsbury：Proceedings of the 2012 Workshop on Biomedical Natural Language Processing，

2012: 91 - 99.

17 Thomas PE, Klinger R, Furlong LI, et al. Challenges in the Association of Human Single Nucleotide Polymorphism Mentions with Unique Database Identifiers [J]. BMC Bioinformatics, 2011, 12 (Suppl 4): S4.

18 Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, et al. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text [J]. Plos One, 2013, 8 (6): e65390.

19 Barbara Rosario, Marti A. Hearst Classifying Semantic Relations in Bioscience Text [C]. Barcelona: the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), 2004.

20 Ariel Schwartz, Marti Hearst. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text [EB/OL]. [2018 - 10 - 11]. http: //psb. stanford. edce/psb - online/proceedings/psb 03/schwartz. pdf.

21 Gurulingappa H, Rajput A M, Roberts A, et al. Development of a Benchmark Corpus to Support the Automatic Extraction of Drug - Related Adverse Effects from Medical Case Reports [J]. J Biomed Inf, 2012 (45): 885 - 892.

22 Gurulingappa H, Klinger R, Hofmann - Apitius M, et al. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature [C]. Valetta: The 2ndWorkshop on Building and evaluating resources for biomedical text mining, 2012.

23 Taboureau O, Nielsen SK, Audouze K, et al. ChemProt: a disease chemical biology database [J]. Nucleic Acids Res, 2011 (39): D367 - 372.

24 Leaman R, Wei CH, Zou C, et al. Mining Chemical Patents with an Ensemble of Open Systems [EB/OL]. [2018 - 10 - 17]. https: //www. ncbi. nlm. nih. gov/pmc/antide. pmc48653271.

25 Li J, Sun Y, Johnson R J, et al. BioCreative V CDR Task Corpus: a resource for chemical disease relation extraction [EB/OL]. [2018 - 10 - 17]. https: //www. ncbi. nlm. nih. gov/pmc/ontides/pmc4860626.

26 Zhang Y, Xu J, Chen H, et al. Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning [EB/OL]. [2018 - 10 - 17]. http: //database. oxfordjournals. org/content/2016/baw049. full. pdf.

27 Caporaso JG, Baumgartner WA Jr, Randolph DA, et al. MutationFinder: a high - performance system for extracting point mutation mentions from text [J]. Bioinformatics, 2007, 23 (14): 1862 - 1865.

28 Kevin Nagel, Antonio Jimeno - Yepes, Dietrich Rebholz - Schuhmann. Annotation of Protein Residues Based on a Literature Analysis: cross - validation against UniProtKb [J]. BMC Bioinformatics, 2009, 10 (Suppl 8): S4.

29 Verspoor K, Mackinlay A, Cohn JD, et al. Detection of Protein Catalytic Sites in the Biomedical Literature [J]. Pac Symp Biocomput, 2012 (18): 433 - 444.

30 KE Ravikumar, Haibin Liu, Judith D Cohn, et al. Literature mining of protein - residue Associations with Graph Rules Learned Through Distant Supervision [J]. Journal of Biomedical Semantics, 2012, 3 (Suppl 3): S2.

31 Johnson H L, W A Baumgartner Jr, M Krallinger K B Cohen, et al. (2007) Corpus Refactoring: a Feasibility Study [J]. Journal of Biomedical Discovery and Collaboration. 2007 (2): 4.

32 Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman et al. Sortal Anaphora Resolution to Enhance Relation Extraction from Biomedical Literature [J]. BMC Bioinformatics, 2016 (17): 163.

33 Cohen KB, Tanabe L, Kinoshita S. A Resource for Constructing Customized Test Suites for Molecular Biology Entity Identification Systems [C]. Boston: Proceedinges of HLT - NAACL 2004 Workshop: Biolink 2004, 2004.

34 刘燕, 孙月萍, 郭臻, 等. 基于文本挖掘的高通量癌症基因组数据注释 [J] 中华医学图书情报杂志, 2016, 25 (12): 34 - 39.

35 胡双, 陆涛, 胡建华. 文本挖掘技术在药物研究中的应用 [J]. 医学信息学杂志, 2013, 34 (8): 49 - 53.

36 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.

37 赖鸿昌, 朱礼军, 徐硕. 面向专利的化合物和生物实体识别系统 [J]. 情报工程, 2015, 1 (4): 95 - 103.

38 杨锦锋, 关毅, 何彬, 等. 中文点击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.

39 Daniel Rubin, Caroline F Thorn, Teri E Klein, et al. A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge [J]. Journal of the American Medical Informatics Association, 2005, 12 (2): 121 - 129.

40 Zhiyong Lu, Michael Bada1, Philip V. Ogren, et al. Improving Biomedical Corpus Annotation Guidelines [EB/OL]. [2018 - 01 - 23]. https: //www. researchgate. net/publication/228695226_ Improving_ biomedical_ corpus_ annotation_ guidelines.