临床文本自然语言处理系统构建研究——以 cTAKES 为例*

杨晨柳 胡佳慧 方 安 王 蕾 任慧玲

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 从系统架构、语料库构建、应用效果 3 方面阐述临床文本自然语言处理系统 cTAKES 构建方法，从设计基于开源框架的系统架构、开发模块化组件、构建临床语料库、注重创新以及针对中文特点建设系统 5 个方面提出对我国中文临床文本自然语言处理系统构建的建议。

〔关键词〕 临床文本；cTAKES；模块化；语料库；自然语言处理

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.12.011

Study on the Building of Clinical Text Natural Language Processing System—Taking cTAKES as an Example *YANG Chenliu, HU Jiahui, FANG An, WANG Lei, REN Huiling, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper elaborates on the method for building cTAKES, a clinical text natural language processing system, from the three aspects of system architecture, building of corpus and application effect, and puts forward recommendations on the building of Chinese clinical text natural language processing system from the five aspects, including design of system architecture based on open source framework, development of modular components, building of clinical corpus, attention given to innovation and building of the system in the light of the feature of Chinese language.

〔**Keywords**〕 clinical text; cTAKES; modularization; corpus; Natural Language Processing (NLP)

1 引言

临床文本是医疗活动过程中产生的一种重要信息资源，包含丰富的临床数据，对于临床科学研究、临床决策支持、护理质量评估等具有重要意义。在大数据时代如何有效利用这些临床数据成为健康医疗科学领域关注的重点之一。临床文本是临床医生实际诊疗细节的体现，其内容与医生在整个诊疗过程中的思考路径密切相关，难以基于统一、严格的表格形式来组织[1]，因此数据应用需要基于具体情况对其进行文本分析与信息提取。

自然语言处理（Natural Language Processing，NLP）是从文本资源中识别并提取信息的有效途径之一。目前 NLP 技术已成为临床文本处理的重要手段[2]，其在临床中的应用有赖于开放标准的支撑和

〔收稿日期〕 2018－09－27

〔作者简介〕 杨晨柳，实习研究员，发表论文 3 篇；通讯作者：胡佳慧，助理研究员。

〔基金项目〕 中国医学科学院医学与健康科技创新工程项目（项目编号：2017－I2M－3－014）；中国医学科学院中央级公益性科研院所基本科研业务费项目（项目编号：2018PT33005）。

信息系统的构建。临床文本自然语言处理系统（clinical Text Analysis and Knowledge Extraction System，cTAKES）[3]是大规模、全面化、模块化、可扩展的开源 NLP 系统，旨在处理和提取语义上可行的信息，支持异构数据的临床研究，且其良好的可扩展性能为要求严苛的临床研究及临床文本产生环境提供有利条件。本文旨在通过对 cTAKES 在系统架构、语料库构建及应用效果等方面的研究，为我国中文临床文本自然语言处理系统的构建提供参考借鉴。

2 cTAKES 构建方法

2.1 概述

cTAKES 是梅奥诊所（Mayo Clinic）为开放健康自然语言处理（Open Health Natural Language Processing，OHNLP）联盟开发的临床文本自然语言处理系统，主要用于从电子健康记录和临床自由文本中提取信息，识别临床命名实体的类型（如药物、疾病/病症、体征/症状、解剖部位等）及其文本范围、本体映射代码、主题、上下文等属性。cTAKES 采用模块化的流水线运行模式，基于字典查找算法组件，通过相应的概念标识符来识别概念和标签[4]，利用基于规则的机器学习技术，实现从临床文本叙述中提取有效信息。作为一种开放资源的自然语言处理系统，cTAKES 不断完善组件配置，为用户提供满足其需求的必备和可选组件，具有良好的灵活性与实用性。此外 cTAKES 还提供丰富的文本分析与信息抽取算法[5]，其实体识别与信息抽取组件在临床文本处理应用实践中可获得较好的准确率和 F 值等性能。

2.2 系统架构

2.2.1 运行模式 cTAKES 是由多个基于规则或机器学习技术的文本处理组件所构成的管道式系统[6]。在 cTAKES 处理流程中其后一个组件输入的文本信息依赖于前一个组件的输出内容。cTAKES 组件体系，见图 1。输入临床文本信息，经过文件预处理器处理后进入核心组件。核心组件的输出结果可以由多个组件处理，包括上下文标注、单词标准化和词性（Part - of - Speech，POS）标注组件。字典查询组件的输入来自标准化和浅层分析后的输出结果，但字典查询的输出并不严格要求标准化。通过分析组件，其输出流向临床文件组件或药物命名实体识别组件。对于药物命名实体识别组件，其输入是基于上下文标注组件的输出；对于副作用组件，其输入为药物命名实体识别组件的输出。成分句法分析、共指关系解析、关系抽取组件采用类似方法获取需要被处理的临床文本信息。

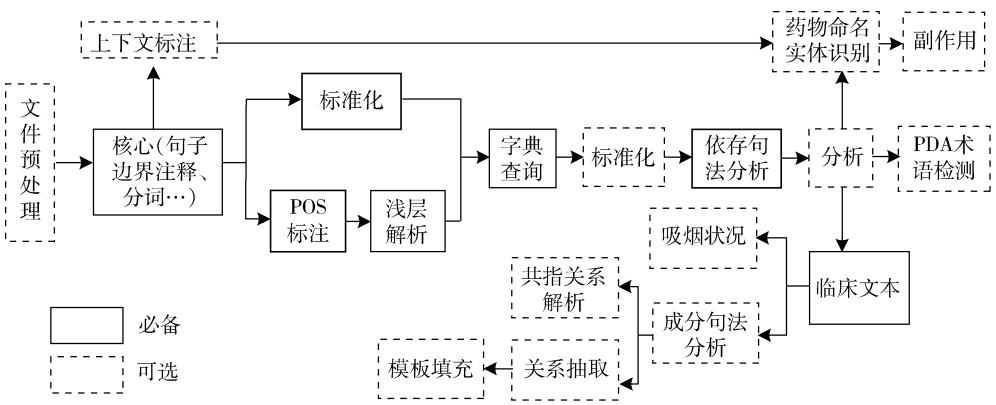

图 1 cTAKES 组件体系

2.2.2 模块化配置 cTAKES 采用模块化方式开发，支持对纯文本和临床文件体系结构（Clinical Document Architecture，CDA）文本的处理，用于命名实体识别及其相关属性标注，其主要由以下组件构成。（1）临床文本预处理。提供 CdaCasInitializer SECTION 注释器，如果 CDA 文档符合 DTD 格式，则将其转换为纯文本。在文本转换过程中注释器将段落标记插入到文本中，在未正确连接的单词中插

入连字符。（2）核心组件。包括分词器、句子边界注释器、替代语句探测器、标记生成器、文档分段描述注释器等。其中句子边界注释器基于 Apache OpenNLP 句子检测模块构成，采用机器学习的方法从大量临床文本训练语句中提取知识，识别句子边界。（3）上下文标注。使用基于统计的方法直接从文档中抽取高频词条，包括日期、分数、度量、题目、范围、数字、时间标注等内容，过滤无意义的词条，从而得到索引项。（4）POS 标注。也称词性标注，主要为句中每个词指派一个合适的词性，即确定单词是名词、动词、形容词或其他词性的过程。标注算法总体可归纳为 3 类：基于规则的标注算法、随机标注算法、混合标注算法。目前 cTAKES 主要使用基于规则的标注算法完成临床文本的词性标注，同时提供临床数据标注所需的模型。（5）浅层解析。也称动名词短语标注，对应 cTAKES 中的分块模块，主要用于标记名词及动词短语。浅层解析器可以创建用于训练临床数据的模型、使用训练后的模型进行临床文本标记以及调整某些分块的结束偏移量来确认分块的模型。（6）单词规范化。是 SPECIALIST 词汇工具的一个组件，为临床文本中的每个单词提供一个表示形式，依据词汇属性将单词进行规范化，具体包括字母大小写、单复数形态、拼写变化、标点符号、属性标记、停用词、变音符号、符号和连词，同时。规范化程序还可以实现同一单词与其不同描述字符之间的映射。（7）命名实体识别。基于字典的方法使每个命名实体从术语映射到概念。字典查询注释器是定制化的，可查找字典条目中单词与文档文本中单词的精确匹配项，也可通过查找字典中单词的排列顺序，实现单词规范形式的匹配。此外，cTAKES 还包含药物命名实体识别模块，又称药物注释器，主要用于识别药物命名实体和相关属性，如剂量、强度、途径等。（8）依存句法解析。主要用于提供句法信息，不同于深层解析器，该组件无需找到明确的简单从句、名字短语等，而是找寻单词之间的依存关系。（9）分析。主要用于检查和记录临床文本中注释的真实含义。如“糖尿病”可能意味着患者患有糖尿病。评估组件用于判断命名实体是否为否定、不确定或者条件性存在。分析组件由分析引擎聚合而成，包括条件属性、通用属性和主题属性注释器。

2.3 语料库构建

cTAKES 采用机器学习的方法，其模型的训练离不开大量的生物医学临床语料。PTB 和 GENIA 是目前广泛采用的两大语料资源。其中 PTB[7] 是 Penn Treebank 项目为语言结构标注产生的文本资源；GENIA[8] 是东京大学分子生物学的文献挖掘项目，其语料库是通过短语结构树标注产生的生物医学文献集合。cTAKES 语料库在 PTB 和 GENIA 语料库的基础上增加从梅奥诊所电子病历中采集的临床语料库生成的共 273 个临床注释，其中包含 100 650 个标签、7 299 个句子、61 份咨询、1 份出院总结、4 次教育访问记录、4 份全科检查、48 份专科检查、19 份多系统评估、43 份其他项目资源、1 份术前医学评估、3 项报告、3 项专业评估、5 项解雇总结、73 项随访、5 项治疗和 3 项实验[9]。语料库的构建有效提升 cTAKES 模块化组件的模型训练效果，显著提高命名实体识别及信息抽取的准确率。同时 cTAKES 功能模块的扩充使其用于机器训练的语料需求量不断增加，近年来包括波士顿儿童医院、科罗拉多大学、麻省理工大学、加利福尼亚大学等机构相继参与到 cTAKES 语料库建设中[10]。随着语料库的不断丰富与完善，cTAKES 将在更大范围内发挥其临床文本自然语言处理的作用。

2.4 应用效果

cTAKES 的构建采用非结构化信息管理架构（Unstructured Information Management Architecture，UIMA）和自然语言处理工具包 OpenNLP。其系统组件专门针对临床领域，包含丰富的语言和语义注释功能。作为单个组件，cTAKES 句子边界检测器的准确度和分词器的精度均可达 0.949，词性标注器精度可达 0.936，浅层解析器 F 值可达 0.924，命名实体识别器和系统级评估 F 值可达 0.715，重叠跨度为 0.824，其概念映射、否定、状态属性的准确度和重叠跨度分别为 0.957、0.943、0.859 和 0.580、0.939、0.839[9]。实际应用中，cTAKES 系统支持 Pretty Print、XML、HTML 3 种输入形式，见图 2。将英文电子病历文本样例录入系统，默认优质输出（Pretty Print）格式，结果页面可详细描述词性标注、命名实体识别、动名词短语标注、依存

关系分析等标注内容，见图 3。鉴于在临床文本处理方面表现出的良好性能，cTAKES 被尝试应用于英语以外的其他语言临床文本处理。其中为验证 cTAKES 是否可以准确地从德语中识别 UMLS 概念信息并且完成与 SNOMED – CT 的映射，M. Becker 等[5]利用 cTAKES 实现对德语临床文本的信息自动抽取。其采用流水线式处理模式，集成分块、分词、字典查找、标注、句子识别及词性识别组件。实验验证阶段，研究人员将 ShARe/CLEF eHealth 2013 分享的 199 条英文临床记录训练集及其德语翻译作为训练预料，对比发现德语的重复性识别精确度和 *F* 值均高于基于英文训练集所得到的性能。

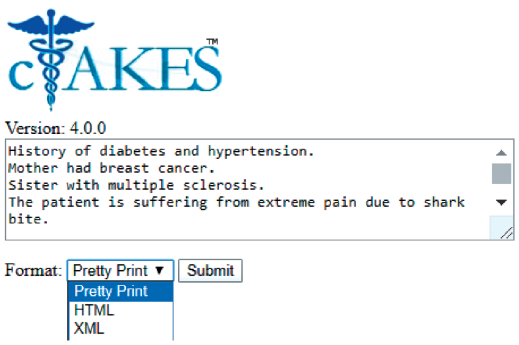

图 2 cTAKES 在线系统界面

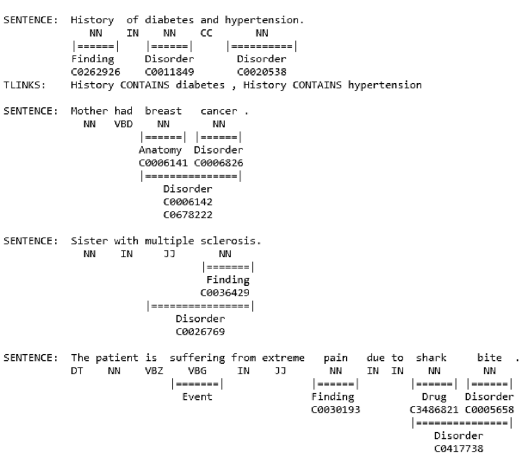

图 3 cTAKES 在线演示界面

3 中文临床文本处理系统构建建议

3.1 基于开源框架设计系统架构

在通用领域，Apache 基金会作为专门运作开源软件的非营利组织，为开发者提供许多经典的开源 NLP 工具集。其中 OpenNLP[11]是 Apache 的一个机器学习工具包，主要用于处理自然语言文本，支持大多数常用的 NLP 任务，如分词、分句、词性标注、命名实体识别、主块分析、语法解析等。目前 OpenNLP 已完成最大熵机器学习分类器的训练，可以从数据集中导出多样化的特征集[12]。UIMA[13]是用于分析非结构化内容（如文本、视频、音频）的组件和软件框架，为非结构化分析提供通用平台，能够实现多个分析组件的组织和集成工作。SPECIALIST 自然语言处理工具[14]由 The Lister Hill 美国国家生物医学通信中心的词汇系统小组开发，帮助应用程序开发人员进行生物医学领域的词汇规范化和文本分析，推动自然语言处理进程。cTAKES 系统在开发过程中主要基于已有的开源 NLP 工具软件，其主体结构建立在 UIMA 和 OpenNLP 已有的模块基础上；其词性标注器、浅层解析器等组件基于 OpenNLP 相关模块；其标准化模块使用 SPECIALIST 词汇工具组件实现对词汇属性的规范化处理。开源软件为 cTAKES 提供定制化的基础，同时避免版权等争议问题，为系统提供更良好的发展环境。

3.2 围绕模块化理念开发系统组件

模块化理念即在解决复杂问题时按照差分的方法对问题进行系统性分解，有序处理各模块问题。模块化设计的基本标准是分解、依赖、聚合，主要具备以下特点：定义封装的模块；定义新模块对其他模块的依赖；对其他模块引入的支持。自 2010 年 1.0 版本发布后，cTAKES 积极致力于 NLP 全过程中各模块的推理和工具的集成，结合实践需求，在原有系统框架基础上不断增加新的模块组件，根据待处理的临床文件类型，将不同组件组合成定制的注释组件流程。截至 4.0 最新版本，cTAKES 系统共包含 28 个模块单元[6]，涉及数据获取、数据预处理、句子边界注释、分词、词性标注、特征提取等多操作模块，用户可根据文本处理需求选择模块，完成功能组配。模块化功能建设及配置使 cTAKES 的灵活性更佳，用户可根据临床文件类型选择合适的流程方案，以获得更优的标注效果。

3.3 依据实际需求构建临床语料库

临床文本包含关于患者全部信息描述，其命名实体标注语料库的构建对医疗领域的知识挖掘具有重要意义。此外语料标注研究作为自然语言处理的重要组成部分一直备受关注，如 PTB 词性和句法标注语料库、生物医疗领域涵盖 10 万级标注的 GENIA 语料库、I2B2 特定任务标注语料库等[15]。cTAKES 用于机器学习的训练语料主要来源于开放可获取的标注文本资源（PTB 和 GENIA 语料库）以及梅奥诊所电子病历采集的注释语料。cTAKES 在 POS 标注和浅层解析组件训练时将 PTB 的注释扩展到临床领域，具体包含编号、数字、药物名称、缩写、药物相关属性、符号等[9]，不仅如此，cTAKES 还将 PTB、GENIA 和梅奥的语料结合，用于浅层解析器模型训练，结果得到更高的精确度及 *F* 值。经典语料库在机器学习中得到越来越多使用的同时，针对特点需求、专业领域的语料库也逐渐出现在人们的视野中，如 Wang 等[16]基于肝癌患者手术记录中抽取肿瘤相关信息构建含有 961 个和肿瘤相关的实体语料库；苏嘉等[15]基于中文电子病历的心血管疾病风险因素构建出国内首份心血管疾病风险因素中文标注语料库。

3.4 注重方法创新，提升应用性能

英文临床文本标识句子结束的标点符号存在歧义，因此仅根据标点符号不能准确判断句子是否结束，需要工具加以辅助。目前部分 NLP 工具仍未实现术语歧义的消除功能或使用启发式的方法来选择最佳注释。cTAKES 首次提出通过在其框架和流程中集成 YTEX[17]来消除歧义。YTEX 为计算概念之间的相似性提供一个框架，作为基于知识的词义消歧组件，YTEX 主要依赖统一医学语言系统（Unified Modeling Language System，UMLS）[18]的分类结构来改进特征排名，将临床 NLP、数据挖掘和特征工程等工具集成，语义相似性度量与机器学习算法相结合进行文档分类，最终确定单词的最佳概念。特别的是 YTEX 实现莱斯科方法（Lesk Method）[19]的自适应调整，通过将每个候选概念与其上下文概念之间的语义相关性相加来对模糊术语的候选概念进行评分，根据分值高低判断最佳注释，降低标注错误率。通过 cTAKES 对单个问题解决方法可以发现模块构建需要在现有 NLP 工具基础上积极创新探索，革新技术提升信息识别与抽取的准确率。

3.5 针对中文临床文本特点构建系统

中文临床文本标注系统的建设不仅要借鉴 cTAKES 整体建设思路，还应面向中文文本特点进行定制开发，从而实现系统建设目标。针对英文临床文本，cTAKES 主要采用分句、分词、词性标注、动名词短语标注、命名实体识别、依存关系句法分析、单词规范化、命名实体属性分析等一系列管道化流程处理，但由于中文文本与英文文本在用词及语法结构等方面存在差异，其面向英文文本开发的分词及单词规范化组件并不适用于中文临床文本。鉴于中文文本自身特点，基于字符串匹配、基于理解及基于统计的分词方法，已有部分国内机构相继推出 jieba、SnowNLP、ICTCLAS、FudanNLP、THULAC 等中文分词工具，为中文临床文本分词工具的建设提供借鉴。因此中文临床文本分词工具的设计和建设应充分了解分词需求，选择适用的分词方法及开源工具，不断完善系统功能。

4 结语

cTAKES 目前已在美国、印度、中国、德国、加拿大、英国、韩国、日本、法国、澳大利亚等多个国家的临床文本分析中得到应用。cTAKES 作为临床文本 NLP 系统，具有模块化功能配置及组件集成的特征，可以被灵活应用于各类医学临床文本处理；基于开源软件的开发思路为 cTAKES 系统避免版权争议，使其稳定性、安全性都极大程度得到提升；梅奥诊所电子病历标注语料库与经典标注语料库结合用于机器学习，配合算法使 cTAKES 得到更为理想的信息识别与提取准确率；将莱斯科方法应用于术语模糊度的评估中，通过在其框架和流程中集成 YTEX 来消除歧义。不同于多数通用 NLP 组件，cTAKES 重点关注临床文本的分析与知识提取，

经过不断探索与技术更新，其临床文本处理性能远超过生物医学科技文献。cTAKES 临床文本分析与知识提取系统的建设理念与方案为我国临床文本处理系统的构建提供良好的经验借鉴，其开源、包容、融合的理念也为临床文本自然语言处理提供新的方向。

参考文献

1 包小源，黄婉晶，张凯，等. 非结构化电子病历中信息抽取的定制化方法［J］. 北京大学学报（医学版），2018，50（2）：256－263.

2 Meystre SM，Savova GK，Kipper－Schuler KC，et al. Extracting Information from Textual Documents in the Electronic Health Record：a review of recent research［J］. IMIA Year book of Medical Informatics 2008，47（1）：128－144.

3 Khudairi，Sally. The Apache Software Foundation Announces Apache cTAKES v4.0［EB/OL］.［2018－11－27］. https：//globenewswire. com/news－release/2017/04/25/970806/0/en/The－Apache－Software－Foundation－Announces－Apache－cTAKES－v4－0. html.

4 Jovanović J，Bagheri E. Semantic Annotation in Biomedicine：the current landscape［J］. Journal of Biomedical Semantics，2017，8（1）：44.

5 Becker M，Böckmann B. Extraction of UMLS Concepts Using Apache cTAKES for German Language［J］. Stud Health Technol Inform，2016（223）：71－76.

6 James Masanz. cTAKES 4. 0 Component Use Guide［EB/OL］.［2018－11－27］. https：//cwiki. apache. org/confluence/display/CTAKES/cTAKES＋4. 0＋Component＋Use＋Guide.

7 Eleni Miltsakaki，Rashmi Prasad，Aravind Joshi，et al. The Penn Discourse Treebank［EB/OL］.［2018－11－27］. https：//alliance. seas. upenn. edu/～nlp/publications/pdf/miltsakaki2004. pdf.

8 S Kulick，A Bies，M Liberman，M Mandel，et al. Integrated Annotation for Biomedical Information Extraction［C］Boston：HLT/NAACL 2004 Workshop：Biolink，2004：61－68.

9 Savova G K，Masanz J J，Ogren P V，et al. Mayo Clinical Text Analysis and Knowledge Extraction System（cTAKES）：architecture，component evaluation and applications［J］. Journal of the American Medical Informatics Association Jamia，2010，17（5）：507.

10 Lars－Erik Bruce. Apache UIMA and Mayo cTAKES UIMA and How It Is Used in the Clinical Domain［EB/OL］.［2018－11－27］. https：//www. uio. no/studier/emner/matnat/ifi/INF5880/v12/undervisningsmateriale/seminar. pdf.

11 Pramod Chandrayan. A Guide To NLP Implementation Using OpenNLP：making machines speak［EB/OL］.［2018－11－27］. https：//codeburst. io/nlp－implementation－using－java－opennlp－guide－and－examples－80d86b02b5b5.

12 Sha R，Pereira F. Shallow Parsing with Conditional Random Fields［C］. Edmonton：NLT－NAACL 2003 workshop：2003，134－141.

13 David Ferrucci，Adam Lally. UIMA：an architectural approach to unstructured information processing in the corporate research environment［EB/OL］.［2018－11－27］. https：//pdfs. semanticscholar. org/9f8e/b04dbafdfda997ac5e06cd6c521f82bf4e4c. pdf.

14 Agah A. Medical Applications of Artificial Intelligence［M］. Boca Raton：CRC Press，Inc. 2013，387－388.

15 苏嘉，吴昊，杨锦锋，等. 基于中文电子病历的心血管疾病风险因素标注体系及语料库构建［J］. 自动化学报，2017，44（X）：1－7.

16 Hui W，Weide Z，Qiang Z，et al. Extracting important information from Chinese Operation Notes with natural language processing methods［J］. Journal of Biomedical Informatics，2014，（48）：130－136.

17 James Masanz. cTAKES 4.0－YTEX SentenceAnnotator［EB/OL］.［2017－11－27］. https：//cwiki. apache. org/confluence/display/CTAKES/cTAKES＋4.0＋－＋YTEX＋SentenceAnnotator.

18 Olivier Bodenreider. The UMLS and the Semantic Web［EB/OL］.［2018－11－27］. https：//www. w3. org/wiki/images/7/71/HCLSIG_BioRDF_Subgroup%24%24Meetings%24%242008－09－22_Conference_Call%24080922－BioRDF－UMLS－1. pdf.

19 Lesk M. Automatic Sense Disambiguation Using Machine Readable Dictionaries：how to tell a pine cone from an ice cream cone［C］. New York：Proceedings of the 5th Annual International Conference on Systems Documentation，1986：24－26.