面向医疗文本的实体及关系标注平台的构建及应用

张坤丽[1,2]，赵旭[1,2]，关同峰[1,2]，尚柏羽[1,2]，李羽蒙[1,2]，昝红英[1,2]

（1. 郑州大学信息工程学院，河南郑州 450001；2. 鹏城实验室，广东深圳 518055）

摘要：医疗文本数据是推行智慧医疗的重要数据基础，而医疗文本为半结构或非结构化数据，难以对其直接进行应用。对医疗文本中所包含的实体及实体关系进行标注是文本结构化的重要手段，也是命名实体识别、关系自动抽取研究的基础。传统的人工标注方法费力费时，已难以适应大数据发展的需求。该文以构建中文医学知识图谱的任务为驱动，构建了半自动化实体及关系标注平台，该平台融合多种算法，能够实现文本预标注、进度控制、质量把控和数据分析等多种功能。利用该平台，进行了医学知识图谱中实体和关系标注，结果表明该平台能够在文本资源建设中控制标注过程，保证标注质量，提高标注效率。同时该平台也应用于其他文本标注任务，表明该平台具有较好的任务移植性。

关键词：文本标注；标注平台；实体标注；关系标注；数据分析

Construction and Application of Entity and Relationship Labeling Platform for Medical Texts

ZHANG Kun-li[1,2], ZHAO Xu[1,2], GUAN Tongfeng[1,2], SHANG Bai yu[1,2], LI Yu-meng[1,2], ZAN Hong-ying[1,2]

(1. School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China;

2. The Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China)

Abstract： Medical text data is an important data foundation for the implementation of intelligent healthcare, and medical text is semi-structured or unstructured data, which is difficult to apply directly. Labeling the entity and entity relationships contained in medical texts is an important means of text structuring, and is also the basis for the study of named entity recognition and automatic relationship extraction. The traditional manual labeling method is a laborious and time-consuming task, and it has been difficult to develop big data. This paper is driven by the task of constructing Chinese medical knowledge graph, Constructed a semi-automated entity and relationship labeling platform that integrates multiple algorithms and can perform text pre-labeling, schedule control, quality control and data analysis. Based on this platform, the medical knowledge graph entity and relationship labeling are carried out. The results show that the labeling platform can control the labeling process in the construction of text resources, ensure the labeling quality, improve the labeling efficiency. At the same time, the platform is also applied to other text annotation tasks, indicating that the platform has better task portability.

Key words: text annotation ; labeling platform; entity annotation; relationship annotation; data analysis

基金项目：国家社科基金重大项目（18ZDA315）；河南省高等学校重点科研项目(20A520038)；河南省科技攻关项目（192102210260）；河南省科技攻关计划国际合作项目（172102410065）

1 引言

日益增长的医疗文本数据给整个行业的发展带来了巨大的机遇和挑战，绝大部分的医疗文本数据属于半结构化或者非结构化的数据，只有将半结构化或非结构化的数据转化为计算机可以处理的结构化数据，才能够对其进行一系列的科研应用，而对文本信息的标注正是对其进行结构化处理的基础[1]。通过文本标注得到的熟语料是一种非常重要的资源，是命名实体识别、关系自动抽取等相关研究的基础[2]。目前已标注完成的高质量语料仍旧十分缺乏，能够用于研究的语料更是屈指可数。文本标注资源的匮乏与当今海量的文本信息形成了鲜明的对比，资源的不足不利于对语言资源的深度研究。而文本标注任务是一项极其繁重枯燥的工作，传统的人工标注耗时耗力且成本巨大，令众多研究者望而却步，导致资源建设进展缓慢。

任务的进行是不可或缺的。除了标注质量之外，对于大型的标注任务，效率也是至关重要的，现有的标注工具[7-9]在效率方面表现不佳。另外，很多标注工具[6-7]对用户并不友好，在其他设备或者服务器上配置繁琐，用户体验不佳。针对以上几种标注工具普遍存在的问题，本文在构建中文医学知识图谱的过程中，面向医疗文本标注的具体任务，构建了融合多种自动识别及抽取算法、包含实体关系属性标注功能以及标注数据分析功能的可扩展半自动标注平台。本文所构建的标注平台主要有以下特点：

1）可视化操作界面，支持实体、实体及关系以及属性的标注，操作简单。

2）内置多种自动标注及抽取算法，辅助进行标注。

3）拥有三种不同形式的数据分析功能，同时支持生成标注对比报告。

4）进度控制，标注过程控制，保证标注质量。

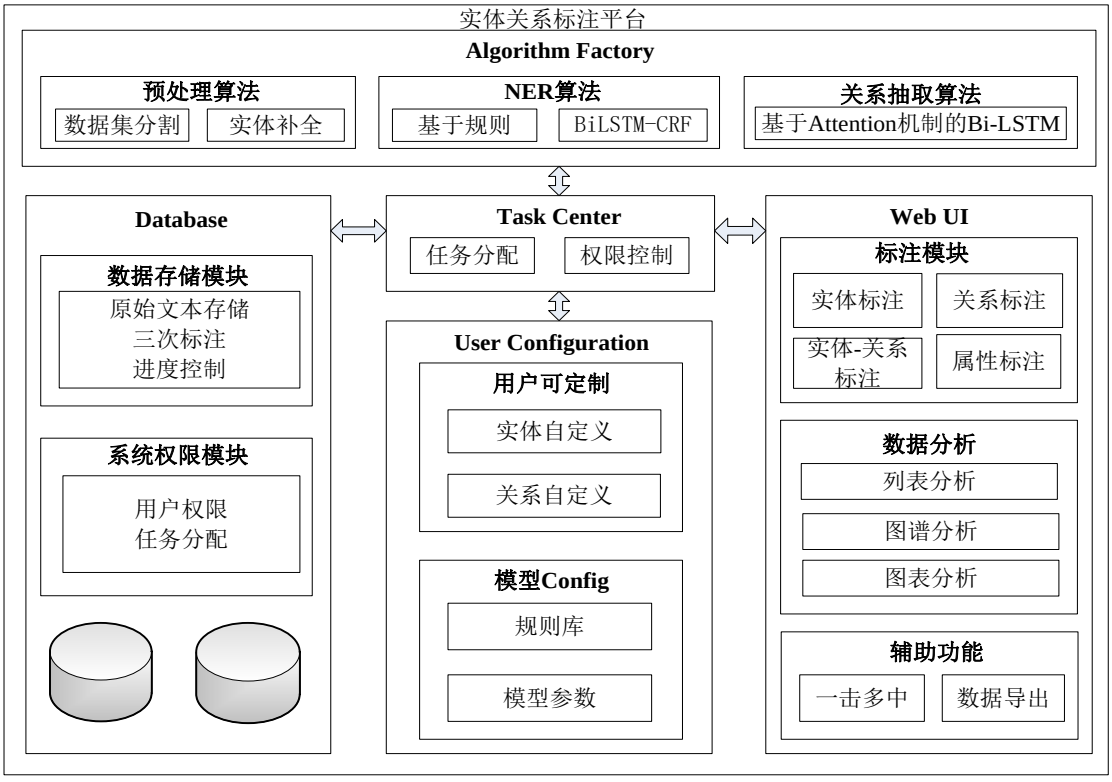

图 1 系统架构

从上世纪九十年代起，随着计算机技术的不断发展，各种辅助标注工具被应用于标注工作中。已有的标注工具[3-6]主要致力于为用户提供一个可视化的标注界面，缺乏对于标注质量的分析，而标注质量分析对于标注

5）良好的可定制性，不仅适用于医疗文本，更支持多种任务。

6）基于 python 的 web 框架开发，配置简单，可移植性强。

以该平台为基础，共完成了百种常见疾病和儿科学疾病的标注，并于 2019 年 8 月 2

日发布了中文医学知识图谱 2.0 版本[1]，目前已包含 11,076 种疾病，18,471 药物，14,794 症状，3,546 诊疗技术的结构化知识描述，描述医学知识的概念关系实例及属性三元组达 1,566,494。

2 总体架构及功能模块

本文所构建的实体及关系标注平台能够对要标注的数据进行预处理及半自动标注，同时提供标注任务分配，标注结果分析及质量把控等基本功能，以下将分别介绍系统架构及各主要功能模块。

2.1 系统架构

图 2 平台标注界面

本文所构建的基于 web 的面向医疗文本的实体及关系标注平台系统架构如图 1 所示，由任务中心（Task Center）、算法工厂(Algorithm Factory)、数据存储（Database）、用户配置（User Configuration）及 Web 界面（Web UI）五个模块组成。Task Center 模块是系统的控制中心，负责任务的分配以及权限的控制工作，并与各个模块进行交互。在 Algorithm Factory 模块中，系统预置了几种常用的算法来对数据进行预处理、命名实体识别（Name Entity Recognition, NER）及关系抽取，对数据进行预标注，以减少用户进行标注的工作量。用户可自行选择是否使用算法及使用哪种算法。在 Database 模块中，将所需标注的原始文件以及标注文件进行存储，并存储与系统权限相关的表信息。User Configuration 模块为用户提供个性化接口，目前开放了实体项及关系项自定义接口，以及自定义模型相关参数接口。Web UI 模块为用户提供可以进行交互的界面，通过简单的鼠标点击和拖动操作即可完成标注任务，同时支持对标注任务的数据分析和标注对比报告生成。算法模块将于第 3 节详细介绍，以下将详细介绍其余四个模块。

2.2 功能描述

2.2.1 任务中心模块

本模块主要负责平台的任务分配和权限控制。任务分配主要是对标注数据的增删改查以及对任务的建立和分配。平台支持 TXT 文件和符合平台格式的 JSON 文件，同一数据集可包含多个文件。任务的管理与数据相互独立，以确保标注文件的一致及安全，在创建任务的过程中可以同时选择多个数据集作为同一个任务进行标注，以及对该任务进行分组管理。平台用户根据权限的不同可以划分为超级管理员、任务管理员和普通用户三种类型。超级管理员拥有系统最高权限，任务管理员则由超级管理员指定分组，只能负责组内用户管理和任务管理。普通用户的管理以组来进行，不同组之间互不干扰，相互独立。只有相应任务管理员拥有对数据的操作功能，普通用户只有在管理员发布任务之后才能进行标注工作，且只有其所属组内数据可见。

[1] http://cmekg.pcl.ac.cn/

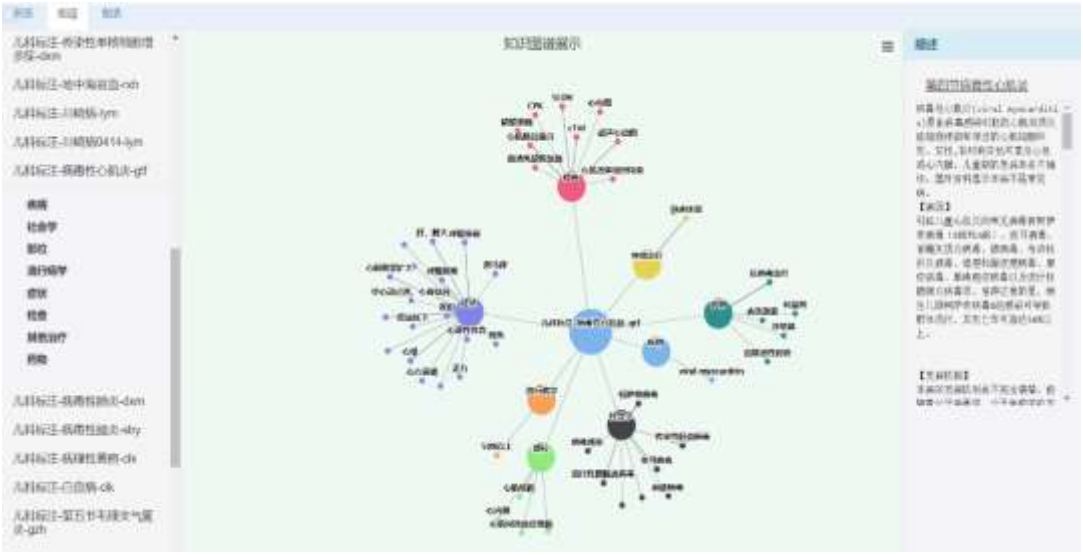

图 3 知识图谱形式

2.2.2 数据存储模块

数据存储模块主要负责对平台上用户所上传文件以及标注文件的实时保存，是整个平台赖以工作的基础模块。在数据库中，对上传的数据文件和任务文件进行分离存储，以防止数据的删除对任务的正常进行造成影响。平台默认的标注过程分为一标、二标和三标三轮标注过程，同时支持用户预定义两轮、三轮或者更多标注过程。平台对多轮标注过程的文件独立存储，并提供当前任务的实时进度显示，方便用户把握任务进度。除了进度控制，平台的权限控制功能也依赖于数据库中权限表的存储。

2.2.3 用户配置模块

用户配置模块开放了一系列进行个性化操作的接口，用户可以根据具体标注任务的需要来进行相应的个性化操作。在标注方面，用户可以选择使用平台默认的实体项和关系项配置文件，或者自行在平台界面上对配置文件进行修改操作，也可以选择从外部导入符合格式的配置文件。在算法工厂之中的模型，提供给用户自定义的功能，同时也可以通过配置文件的形式对模型进行个性化配置，对模型的介绍可见下一节。

2.2.4 Web 界面模块

本模块为用户提供一个可视化的操作界面，包含标注功能、数据分析功能以及其他辅助标注功能，以下为各个功能的详细介绍。

1） 标注功能

本文所构建的面向医疗文本的实体及关系标注平台，拥有实体标注、关系标注和属性标注的功能，平台标注界面如图2所示。对于在文本之中出现的实体，用户在点击实体按钮选择实体标签后选中对应文字即可完成实体标注。在完成实体标注之后，用户可以选择是否进行关系标注，并可随时在实体标注和关系标注两种标注模式之间来回切换。本文默认所定义的关系为（实体 1，实体 1 属性，实体 2，实体 2 属性，关系名称，关系属性）的六元组形式，用户也可以自定义关系表现形式。关系标注与实体标注类似，用户需要先选择对应的关系名称，然后点击实体 1 对应的实体，再点击实体 2 对应的实体即可完成标注。属性的标注则由用户选择是否开启，关闭属性标注时属性对应值置为空。若用户选择开启属性标注，则在关系标注的过程中，实体 1、实体 2 以及关系完成标注之后会弹出是否进行属性标注对话框，选是后用户选择属性对应文字即可完成属性的标注，选否则该属性置为空。

2）数据分析功能

在对数据进行处理的过程中，不仅对数据的标注是重要的，对标注数据的即时性分析也是必不可少的一项工作。结合上一部分标注功能的实现，平台提供标注数据的即时分析功能和标注对比报告的生成，方便用户在标注的同时把握标注质量。平台提供三种不同的数据分析方式，分别是列表形式、知识图谱形式和图表形式。列表分析方式的结

果提供 Excel 格式文件的导出。除了以上三种分析方式，平台也可以生成两个标注文件的详细对比报告，详细信息在 4.3 节展开介绍。

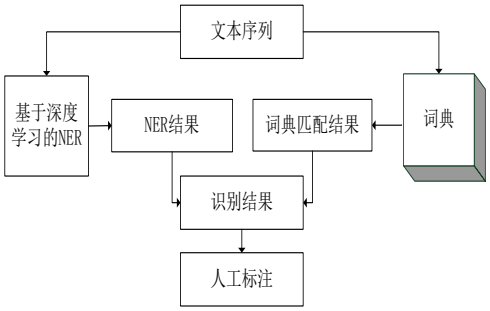

图 4 识别过程

3）辅助功能

平台目前暂时提供两种辅助功能，分别是一击多中和数据导出。平台通过对已标注数据中的实体进行记录，在标注的过程中用户只要标注一个实体，其他相同的实体将自行标注上去，可以提高标注效率。而在数据导出功能中，用户可以对关系进行检索（包含按关系名称、按文件名称等方式）后进行导出操作，导出的关系将以 Excel 文件的形式提供给用户进行下载。

3 算法工厂

本节主要对平台中预置的多种算法进行详细的说明，主要介绍预处理算法和 NER 算法，关系抽取算法暂时交由用户根据自身任务需要自行导入，本节不再进行具体介绍。

3.1 预处理算法

这部分主要对待标注数据分割及其中出现的实体进行补全处理。

1） 待标注数据集分割

待标注数据集的分割可分为按句和按篇章两种不同的分割方式，用户可以根据需要自行选择采用何种分割方式。

2）实体名称补全

对于文本较长的标注任务，若标注规范以一种单一实体为中心且该实体在仅在文中开头部分出现亦或在文中不曾出现，这种情况将会导致文本中许多重要的关系因为没有实体 1 的存在而不能完成标注或者只能跨很长的距离进行标注。针对这种在标注任务中多次出现的问题，系统可在文本中每句开头增添以@结尾的实体项（例如慢性心房颤动@xxxxxxxxxx）。用户可以自定义要在句子开头增添的实体项，以满足对于实体标注的需求。

3.2 NER 算法

命名实体识别任务主要是要识别出文本中出现的专有名称和有意义的数量短语并加以归类。所谓的命名实体主要包括实体(组织名、人名、地名)、时间表达式((货币值、百分数)等[10]。本文所针对的医疗文本命名实体识别的主要任务则需自动识别出其中出现的疾病、药物、症状等相关的命名实体。平台根据构建知识图谱过程中对医疗文本进行标注的需要，内置以下算法来进行预标工作，以减少人工标注的工作量，提高标注效率。

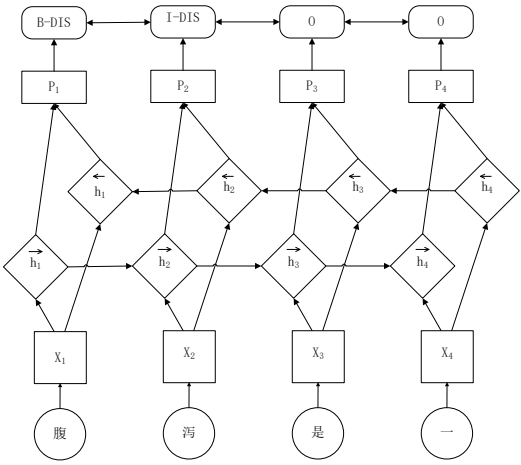

图 5 BiLSTM-CRF 模型

1）识别过程

实体识别的过程如图4所示，通过词典对输入的文本序列进行匹配，得到初步识别结果，再与基于深度学习的方法的结果融合之后交由人工进行标注，最终完成整个标注过程。

2） 基于规则的算法

基于规则的命名实体识别方法中的规则对文本的领域具有较强的依赖性，需要根据其所属领域，在专家的帮助下，来人工进行规则或者词典的建立[11]。

本文所采用的词典是在医学专家的建议下，结合了 ICD-10、《Mesh 医学词表》、《ICD9-CM 手术编码》、药品的解剖学、治疗学及化学分类系统（Anatomical

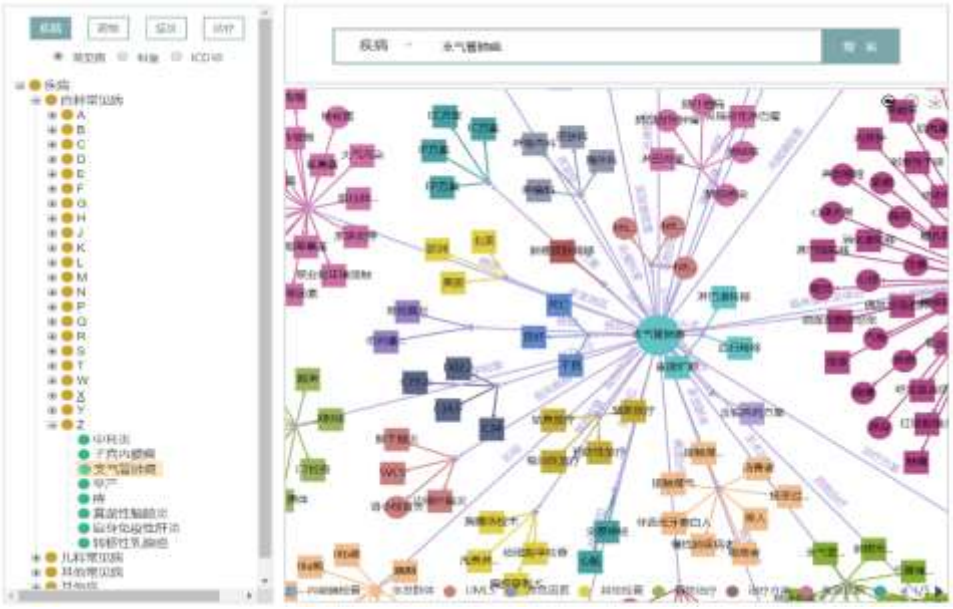

图 6 中文医学知识图谱 2.0

Therapeutic Chemical，ATC）、中文症状库 OpenKG.CN、《2018 年医保诊疗目录》以及大夫们所认可的丁香园网上的一些信息综合而成，词典各类型实体统计如表 1 所示。

表 1 词典实体数量统计

实体名称	数量
疾病	65, 528
部位	1, 347
症状	22, 327
检查	3, 369
手术治疗	5, 040
药物治疗	18, 441

3）基于深度学习的算法

近年来，基于深度学习的命名实体识别方法开始应用[12]，BiLSTM-CRF 也应用到了医疗领域的命名实体识别研究中来[13]，Dong C 等首次将基于字符的 BiLSTM-CRF 应用到中文命名实体识别任务中来[14]。本文使用基于字的 BiLSTM-CRF 来对未标注数据先进行命名实体识别，再交由人工进行处理。使用 BIO 标注集，以句子为单位，字向量默认是采用了简单的随机初始化方式，通过 BiLSTM 层和 CRF 层得到初步的命名实体识别结果。

4 应用

4.1 基于标注平台的中文医学知识图谱构建

在中文医学知识图谱构建的过程中，实体关系标注平台的构建是和标注任务同步进行的，在标注规范逐步完善的过程中，标注平台的构建也逐步进行。

实体资源库基于 MeSH[15]（Medical Subject Headings）主题词表，并且融合 ICD-10[16]（International Classification of Diseases）、ATC（Anatomical Therapeutic Chemical）等医学术语为资源库中的实体。最终形成的标注规范将实体分为 12 大类，分别为疾病、部位、症状、药物、检查、其他治疗、手术治疗、药物治疗、流行病学、预后、其他和社会学，并使用不同的参考标准界定每一类实体涵盖的范围，详细的标注规范另作他文讨论。

实体之间的关系包括：语义、疾病-部位、疾病-症状、疾病-检查、疾病-疾病、疾病-其他治疗、疾病-手术治疗、疾病-药物治疗、疾病-流行病学、疾病-预后、疾病-其他、疾病-社会学等 12 类型关系。

在实际的标注过程中，采用标注平台的自定义实体及关系项的功能，以配置文件形式将上述实体及关系项导入到平台之中，同时搜集并完善了平台内置词典，在不断的标注过程中对模型进行不断训练，结合一击多中的辅助功能，大大加快了整个标注任务的进行。

在构建实体及关系标注平台的过程中，

我们逐步完善了以疾病为核心的中文医学知识图谱命名实体和关系标注体系及规范，发布了融合儿科学疾病的中文医学知识图谱 2.0 版本（如图 6）。

4.2 标注结果及分析

为了保证标注的质量以及对标注进度的把控，根据对常见疾病的统计结果以及相应医学专家的建议，在构建中文医学知识图谱的过程中，选择了百种常见疾病及儿科疾病来进行标注，由本平台来完成任务的上传与分配及标注工作。

整个的标注过程历时半年多，由包括两名医学专家的多人参与标注工作。最终完成了 106 种疾病和 504 种儿科疾病的标注工作，共计完成标注 600 余万字，47078 种实体概念，69043 个实体关系六元组。去重后的实体数量如表 2，关系数量如表 3：

表 2 实体数量统计

实体名称	数量
疾病	5,643
部位	670
症状	4,590
检查	2,808
手术治疗	683
药物治疗	2,454
其他治疗	1,283
社会学	3,580
流行病学	854
预后	165
其他	745

表 3 关系数量统计

关系名称	数量
疾病-疾病	5,334
疾病-部位	966
疾病症状	7,729
疾病-检查	4,401
疾病-手术治疗	858
疾病-药物治疗	4,194
疾病-其他治疗	1,645
疾病-社会学	4,567
疾病-流行病学	1,042
疾病-预后	183
疾病-其他	858
同义词	753

4.3 质量把控

标注的质量把控对于语料库的构建十分重要，为了保证标注过程的准确性和一致性，本平台采用了多轮标注的形式。一标 、二标可并行标注，不一致的问题讨论后由第三人标注。

为了更加全面的对标注结果进行分析，平台可以生成两次标注的对比报告（如图 7），生成后的报告提供 pdf 格式进行下载。报告的内容主要包括以下两个方面：

1）总体分析：以 File1（三标文件）作为金标准，以表格的形式给出了两个文件之中所有实体类别的准确率、召回率和 F1 值，计算公式如式（1）-（3）所示。

2）内容对比：该部分详细的展示了标注文件的所有文本内容，并对其中的实体标签以不同类型颜色进行高亮显示，蓝色代表只在文件 1 中出现，红色代表只在文件 2 中出现，绿色实体则在两个文件中同时出现。

$$P = \frac{File_1 和 File_2 一致的标注结果总数}{File_2 的标注总数} \quad (1)$$

$$R = \frac{File_1 和 File_2 一致的标注结果总数}{File_1 的标注总数} \quad (2)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (3)$$

由于标注规范的更新和平台初期的不完善，在构建的过程中遇到的问题和相应解决方法大致分为以下两类：

图 7 标注对比报告


Annotation Comparison Report

1 Overall Statistics

2 Detail Content Comparison

1）实体项和关系项名称的变更。由于早期的标注规范并不够十分完善，在逐步的标注过程中，通过对标注过程中所遇到的问题进行分析和总结，加上医学专家的解答和建议，后期的实体及关系项已经与前期具有一定差异。针对此类问题的存在，平台提供可以将某一种实体名或者关系名进行统一替换为新名称的接口，方便在标注规范变更之后保持语料标注结果的一致性。

2）三元组关系到六元组关系的转变。在早期的标注规范之中，采用三元组<实体 1-实体 2-关系>的形式来对关系进行描述，但由于三元组不足以囊括一种关系所包含的信息，新的标注规范将额外的信息作为属性来进行标注。这个属性一般指实体的修饰、解释说明和条件限制等信息。因此，之后扩充为四元组关系<实体 1-实体 2-关系-关系属性>，增加了关系的属性描述。但四元组并不能涵盖描述关系中单个实体的信息，最终扩充为<实体 1-实体 1 属性-实体 2-实体 2 属性-关系-关系属性>的六元组关系。平台在保证六元组关系标注的同时保持对之前三元组关系和四元组关系的兼容，并新增属性标注开关，用户可自行决定是否启用。

4.4 其他应用

本文所介绍标注平台不仅可应用于医疗文本上的标注，由于其具有可扩展性，还可应用于其他领域文本的标注任务。第一，已经应用于其他垂直领域的实体及实体关系标注，如军事和法律，目前已完成军事语料的实体标注任务。第二，已经应用于小说文本标注任务，目前正在进行小说人物名称以及人物关系的标注任务。第三，可以应用于文本分类任务，为文本打上所属标签，适用于情感分析或者多标签文本分类任务。在使用平台进行标注的过程中，用户可以自由定义实体及关系类型，自由选择标注形式，以完成特定领域的标注任务。

5 总结和展望

本文以构建中文医学知识图谱任务为驱动，构建了文本实体及实体关系标注平台。分别从标注平台概述、算法概述以及实践效果分析三个方面对平台的整体架构进行了详细的介绍，对在平台的构建过程中所遇到的问题进行了分析和总结。截至 2019 年 3 月，在标注平台所标注语料的基础上，成功的构建了中文医学知识图谱 CMeKG1.0 版本，并且在 2019 年 7 月底又完成了五百多种儿科学疾病的标注工作，于 2019 年 8 月 2 日发布了中文医学知识图谱融合儿科学疾病的 2.0 版本。从目前的情况来看，标注平台还有相当大的完善空间，未来将致力于整合更多效果更好的算法到平台之中，并提供数据的离线自主学习功能，打造一个不仅适用于医学文本，具有多领域普遍适用性的实体及关系标注平台。

参考文献

[1] 李昊迪. 医学领域知识抽取方法研究[D]. 哈尔滨工业大学, 2018.

[2] Todd J, Richards B, Vanstone B J, et al. Text Mining and Automation for Processing of Patient Referrals[J]. Applied Clinical Informatics, 2018, 09(01):232-237.

[3] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. Gate: an architecture for development of robust hlt applications. In ACL. pages 168–175.

[4] Thomas Morton and Jeremy LaCivita. 2003. Wordfreak: an open tool for linguistic annotation. In NAACL: Demo. pages 17–18.

[5] Stephan Druskat, Lennart Bierkandt, Volker Gast,
Christoph Rzymski, and Florian Zipser. 2014. Atomic: An open-source software platform for multi-level corpus annotation. In Proceedings of the 12th Konferenz zur Verarbeitung nat ̈urlicher
Sprache.

[6] Wei-Te Chen and Will Styler. 2013. Anafora: a webbased general purpose annotation tool. In NAACL. volume 2013, page 14.

[7] Philip V Ogren. 2006. Knowtator: a prot ́eg ́e plugin for annotated corpus construction. In NAACL: Demo. Association for Computational Linguistics, pages 273–275.

[8] Pontus Stenetorp, Sampo Pyysalo, Goran Topi ́c,
Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In EACL: Demo. pages 102–107.

[9] Erdmann M, Maedche A, Schnurr H P, et al. From manual to semi-automatic semantic annotation: about ontology-based text annotation tools[C]// P Buitelaar & K Hasida. 2000.

[10] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3):329-340.

[11] 赵哲焕, 杨志豪, 孙聪,等. 生物医学文献中的蛋白质关系抽取研究[J]. 中文信息学报, 2018, 32(07):87-95.

[12] Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Computer Science, 2015.

[13] Kai X, Zhou Z, Hao T, et al. A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition[J]. 2017.

[14] Dong C, Zhang J, Zong C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[J]. 2016.

[15] Lipscomb C E . Medical Subject Headings (MeSH).[J]. Bull Med Libr Assoc, 2000, 88(3):265-266.

[16] Sundararajan V , Henderson T , Perry C , et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality[J]. Journal of Clinical Epidemiology, 2004, 57(12):0-1294.

张坤丽（1977—），博士研究生，讲师，主要研究领域为自然语言处理、医学信息处理。
E-mail：ieklzhang@zzu.edu.cn

赵旭（1995—），通信作者，硕士研究生，主要研究领域为自然语言处理。
E-mail：zhaox917@163.com

昝红英（1966—），博士，教授，主要研究领域为自然语言处理、语言资源构建。
E-mail：iehyzan@zzu.edu.cn