糖尿病电子病历实体及关系标注语料库构建*

叶娅娟[1]，胡斌[1]，张坤丽[1,2†]，咎红英[1,2]

1.郑州大学　信息工程学院，河南　郑州
2.鹏城实验室，广东　深圳
1735623453@qq.com；2609829897@qq.com；ieklzhang@zzu.edu.cn；
iehyzan@zzu.edu.cn

摘要

电子病历是医疗信息的重要来源，包含大量与医疗相关的领域知识。本文从糖尿病电子病历文本入手，在调研了国内外已有的电子病历语料库的基础上，参考I2B2实体及关系分类，建立了糖尿病电子病历实体及实体关系分类体系，并制定了标注规范。利用实体及关系标注平台，进行了实体及关系预标注及多轮人工校对工作，形成了糖尿病电子病历实体及关系标注语料库(Diabetes Electronic Medical Record entity and related Corpus，DEMRC)。所构建的DEMRC包含8,899个实体、456个实体修饰及16,564个关系。对DEMRC进行一致性评价和分析，标注结果达到了较高的一致性。针对实体识别和实体关系抽取任务，分别采用基于迁移学习的BiLSTM-CRF模型和RoBERTa模型进行初步实验，并对语料库中的各类实体及关系进行评估，为后续糖尿病电子病历实体识别及关系抽取研究以及糖尿病知识图谱构建打下基础。

关键词： 糖尿病；电子病历；实体及关系标注体系；语料库构建

Construction of Corpus for Entity and Relation Annotation of Diabetes Electronic Medical Records

Yajuan Ye[1], Bin Hu[1], Kunli Zhang[1,2] and Hongying Zan[1,2]

1.School of Information Engineering, Zhengzhou University, Zhengzhou, Henan, China
2.Peng Cheng Laboratory, Shenzhen,Guangdong
1735623453@qq.com；2609829897@qq.com；ieklzhang@zzu.edu.cn；
iehyzan@zzu.edu.cn

Abstract

Electronic medical record (EMR) is an important source of medical information, which contains a lot of medical knowledge. In this paper, starting with the text of diabetes electronic medical records, on the basis of investigating the existing corpus of electronic medical records at home and abroad, and referring to the classification of I2B2 entities and relationships, the entity and entity relationship classification system of diabetes electronic medical records is established, and the labeling standards are established. Using entity and relation tagging platform, entity and relationship pre-tagging and multiple rounds of manual proofreading are carried out, and the entity and relational tagging corpus of diabetes electronic medical record (Diabetes Electronic Medical Record entity and related Corpus, DEMRC) is formed. The constructed DEMRC consists

国家重点研发计划（2017YFB1002101），国家社科基金重大项目（17ZDA138），国家自然科学基金（62006211），河南省科技攻关项目（192102210260），河南省医学科技攻关计划省部共建项目（SB201901021），河南省高等学校重点科研项目（19A520003，20A520038），教育部人文社科规划项目（20YJA740033），郑州市协同创新重大专项科技攻关项目（20XTZX11020）

†Corresponding author

of 8899 entities, 456 entity modifications and 16564 relationships. The consistency of DEMRC is evaluated and analyzed, and the labeling results achieve high consistency. For the task of entity identification and entity relationship extraction, the BILSTM-CRF model based on transfer learning and the Roberta model were used for preliminary experiments and various entities and relationships in the corpus were evaluated, which laid a foundation for the follow-up research on the Entity identification and relation extraction of diabetes electronic medical records and the construction of the diabetes knowledge graph.

Keywords: diabetes , electronic medical records , entity and relation annotation system , corpus construction

1 引言

随着电子信息的发展，医疗服务模式由被动接受转为主动获取，如何从海量医疗信息中获取有用的信息成为当下研究的热点。命名实体识别和实体关系抽取是信息抽取中的重要任务，对信息抽取技术的研究与应用都有重要的意义。对于中文电子病历，命名实体识别任务的识别目标主要包括检查、手术、疾病、症状、药物、部位等。实体关系抽取是从非结构化的文本中抽取结构信息，其主要任务目标不仅仅是抽取文本中的实体关系，更重要的是判断实体间关系的类型，如“患者糖尿病3年，给予二甲双胍治疗，血糖控制可”，实体关系抽取需要抽取出“糖尿病”与“二甲双胍”之间的关系为“治疗改善了疾病”。电子病历作为医疗信息的重要来源，研究电子病历实体及实体关系抽取可以极大的促进医疗行业的发展。

此外，随着我国居民生活方式的变化，糖尿病已成为流行病，且逐渐呈年轻化趋势，我国成年人患病率高达11.2%，其中2型糖尿病约占我国糖尿病患者总数的90%。近年来，国内外陆续发布并更新多项预防、治疗糖尿病及并发症、合并症指南及专家共识，促进了糖尿病防治工作的规范化。刘勇 (2020)以《中国2型糖尿病防治指南(2017年版)》内容为核心，结合电子病历、医学指南、医学词典等基础数据，构建了基于糖尿病防治的医学知识图谱，促进了糖尿病医学知识的共享、传播和利用。尽管糖尿病防治工作已经得到了自然语言领域众多学者的关注，但由于非结构化的医学文本具有实体类型复杂、实体关系密度高、对具体关系描述连续等特性，以及目前公开的语料库仍然比较缺乏，这些在一定程度上阻碍了其研究工作的进展。

临床电子病历是进行医学研究的重要数据来源，其真实性和严谨性被大众普遍认可。因此本文采集了糖尿病电子病历数据，制定了糖尿病电子病历实体及关系标注规范，采用多轮迭代的标注模式，构建了糖尿病电子病历实体及关系标注语料库，为电子病历命名实体识别和关系抽取以及糖尿病知识图谱的构建提供数据支撑。本文其余组织部分如下：第2节介绍相关研究，第3节概述糖尿病电子病历实体及关系标注体系，第4节是阐述语料库的构建过程并对构建结果进行统计和分析，第5节是进行初步实验对语料库进行评估，最后一节是结语。

2 相关研究

近年来，国内外众多学者聚焦于语料库构建的研究。目前构建语料库的方法有自动构建 (Brockett and Dolan, 2005; Dolan and Brockett, 2005)和人工构建 (Vincze et al., 2008; 周惠巍et al., 2017)两种。在医学领域，国外著名医学本体库有SNOMED、IBM Watson Health和CT (Stearns et al., 2001)等。2006年Meystre等人 (Meystre and Haug, 2006)构建了包含80种常见医疗术语命名实体标注语料，并标注了医疗问题的修饰词信息，该语料共含有160 份文档，文档类型包括病程记录、出院小结等信息。2008年Savova等人 (2010)构建了160份医疗文档规模的命名实体语料，对其中的疾病实体进行了标注，并首次对实体和实体关系的修饰信息进行了细致的分类，该语料包括住院记录、门诊记录以及出院小结3种类型。2009年Roberts等人 (2009)构建了2万份癌症患者病历的标注语料，并详细介绍了语料库的标注和构建方法，其研究主要用于对从患者病历中自动提取临床重要信息系统的开发和评估。2010年I2B2组织了关于抽取概念、概念的修饰及关系的评测任务 (Zlem et al., 2011)，发布了394份训练语料和477份测试语料。I2B2 2010语料标注体系覆盖了医疗问题、治疗和检查三种实体类型，实体之间的关系类型有3种，医疗问题的修饰共6种。该语料标注工作由医学专

家完成，采用F值进行评价。另外，Mizuki等人 (2013)组织了在日文电子病历上进行命名实体识别的TCIR-10MedNLP任务，发布50份病历标注语料，该语料库的标注体系覆盖了患者个人信息和治疗信息，由于标注语料是医生虚构的50份文档，且只标注了主诉和诊断，难以获得真实电子病历中更多的信息。2014年Névéol等人 (2014)采用机器标注和人工校对的方式构建了涉及15类实体的命名实体标注语料，共包含2500篇医学文章。2017年Campillos等人 (2018)构建了500份文档规模的法语命名实体及实体关系语料库，文档类型包含出院小结、程序报告、医生来信和处方，对11类实体和37类关系进行了标注。

相比较于英文及日文医疗数据，中文医疗语料库的构建起步较晚。Lei等人 (2014)收集了北京协和医院800份电子病历并由两位医生参考I2B2 2010标注规范进行制定，包含4类实体，把治疗细分为药物和过程，初始语料包括入院小结和出院记录各400份。Wang等人 (2014)采用复旦大学附属中山医院的115份肿瘤患者的手术记录作为语料，由3位医生参与制定标注体系，定义12种实体类型，共标注961个实体。有关医疗命名实体标注体系较多，但规模较小，覆盖的医疗概念实体少。2016年杨锦锋等人 (2016)结合中文电子病历中命名实体的特点，制定了中文电子病历命名实体和实体关系详细标注规范，涉及5种实体类型、6种关系类型，通过手工标注构建了标注体系完整、规模较大的中文电子病历标注语料库，语料库包含了992份病历文本。2018年，阿里云举办的天池大赛，提供的糖尿病数据集共包含84个特征、1个0-1标签的1000条妊娠糖尿病样本，通过数据挖掘和机器学习的方法预测出有高风险患妊娠糖尿病（Gestational Diabetes Mellitus,GDM）的患者。眷红英等人 (2020)将儿科经典教材作为初始语料，以儿科疾病为中心，制定了涵盖11类医学实体，45种子关系的命名实体和实体关系标注体系。Xia等人 (2009)对医疗领域语料库的标注模式进行了总结，主要包括领域专家标注、众包标注和团体标注三种标注模式。

经过调研，目前已经存在的电子病历实体和实体关系语料库存在一些问题：(1)电子病历标注语料的数据类型不完整，仅包含入院记录、首次病程、查房记录、出院小结和出院医嘱等多类文档中的其中几类；(2)标注语料来源于某个科室或者多类科室，缺乏针对糖尿病所构建的实体关系数据集；因此，本文筛选了完整的糖尿病电子病历数据进行标注，并完成了语料库的构建、分析与评估。

3 糖尿病电子病历实体及关系标注体系

3.1 标注体系

对糖尿病电子病历进行分析的过程中发现，病历文本中存在身体部位与其对应的症状被修饰词分割的情况，且修饰词与症状之间存在一对多的关系，如“全腹无压痛、反跳痛及肌紧张”，否定修饰把身体部位与症状隔开了，如果只标注后面的“压痛”、“反跳痛”和“肌紧张”则与原文表达意思有出入，因此本文添加了身体部位的标注，当身体部位与症状之间存在其他词时，需要对身体部位进行标注，身体部位与后面的多个症状建立位置关系，为了使得标注的信息更全面，修饰词“无”需要与“压痛”、“反跳痛”和“肌紧张”分别建立否定修饰关系。此外，电子病历数据中的时间也是很重要的一类实体，糖尿病是一种慢性病，病历文本中的时间对糖尿病的分析起着很重要的作用，本文把时间划分为时间段和时间点两个子类，并将时间与疾病和症状分别建立修饰关系。

通过对糖尿病电子病历的分析，并参考I2B2 2010评测数据 (Zlem et al., 2011)以及国内外已有的医学领域标注语料实体及关系分类，在医生的指导下，制定了糖尿病电子病历实体及关系标注体系。所制定的标注体系包含的实体主要有疾病、症状、检查、治疗、部位五大类，考虑到治疗方式对疾病的影响不同，将治疗分为药物治疗、手术治疗和其他治疗。

在实体分类的基础上，本文将关系划分为治疗-疾病、治疗-症状、检查-疾病、检查-症状、疾病-症状和部位-症状6类关系，具体的子分类如表 1所示。

修饰识别(Assertion Detection)是电子病历信息抽取过程特有的任务，指在给定病历文本中的疾病、症状等特定类别实体的情况下，从文本中识别出这些实体的修饰成分的过程。本文的处理方式是标注实体时，将修饰词标注为修饰，将修饰类型作为一种关系建立在修饰词与被修饰实体之间，此时，修饰识别就转化为了修饰关系抽取，方便进行语料库的评估以及后续研究。本文共包含疾病修饰、症状修饰和治疗修饰3类，其中症状的修饰划分为11类，疾病的修饰划分为9类，治疗划分为3类。表 2是具体的修饰分类说明。

实体1	标签	关系类型	标签	实体2	标签
治疗	dru/sur/oth	治疗改善了疾病	TrID	疾病	dis
		治疗恶化了疾病	TrWD		
		治疗导致了疾病	TrCD		
		治疗施加于疾病	TrAD		
		因为疾病而没有采取治疗	TrNAD		
治疗	dru/sur/oth	治疗改善了症状	TrIS	症状	sym
		治疗恶化了症状	TrWS		
		治疗导致了症状	TrCS		
		治疗施加于症状	TrAS		
		因为症状而没有采取治疗	TrNAS		
检查	ite	检查证实了疾病	TeRD	疾病	dis
		为了证实疾病而采取检查	TeCD		
检查	ite	因为症状而采取检查	TrATe	症状	sym
		检查证实了症状	TeRS		
疾病	dis	疾病导致症状	DIS	症状	sym
部位	bod	位置	POS	症状	sym

表 1. 实体关系分类

修饰词	**标签**	**修饰类型**	**标签**	**实体1**	**标签**
修饰	mod	否认	dis_neg/sym_neg	症状/疾病	sym/dis
		严重程度	sym_sev		
		可能的	dis_pro/sym_pro		
		非患者本人的	dis_none/sym_ none		
		待证实的	dis_unver/sym_unver		
		性质	sym_nat		
		频率	sym_fre		
		有条件的	dis_cond/sym_cond		
		当前的	dis/sym		
修饰	mod	否认	dru_neg/sur_neg/oth_neg	治疗	dru/sur/oth
		既往	dru_past/sur_past/oth_past		
		当前的	dru/sur/oth		
时间	tim	既往	dis_past	疾病	dis
		持续	dis_dua		
时间	tim	既往	sym_past	症状	sym
		持续	sym_dua		

表 2. 修饰关系分类

第二十届中国计算语言学大会论文集，第622页-第632页，呼和浩特，中国，2021年8月13日至15日。

3.2 标注准则

命名实体所遵循的标注规则：(1)不重叠标注，即同一字符串不能标注为两种不同的实体类型。(2)不嵌套标注，即一个实体不能在另一个实体的内部。(3)实体尽可能不包含标点符号（、，。：；）以及连接词（或、和、以及）。

实体修饰会存在一些符号，比如“？”是指“可能的”，因此为了保证标注信息的完整性。本文中实体修饰标注所遵循的标注原则：(1)遵循命名实体标注规则。(2)允许标注“？”、“+”及“-”等符号。

实体关系所遵循的标注原则：优先标注句内关系，若句内不存在实体关系，允许跨句标注。

3.3 特殊情况说明

(1)修饰词与实体之间关系数目的界定

在糖尿病病历文本中，修饰词与实体之间存在一对多或者多对一的情况，为了保证标注的信息更全面，本文的处理是一个实体允许标注多个修饰关系，一个修饰词也可与多个实体建立修饰关系。但如果一个实体存在两个类似的修饰词，则只标注其中一个。

①1天前无明显诱因出现双手麻木，为间断性。

该样例中，“双手麻木”为症状，“1天前”和“间断性”均为修饰词，“1天前”用来界定症状发生的时间，“间断性”用来表示症状的性质，二者均为“双手麻木”的修饰词，均需要标注。

②口渴、多饮、多尿7年，测空腹血糖12.0mmol/L，口服降糖药物二甲双胍片0.5 3次/日，未控制饮食及运动，空腹血糖波动在12- 13mmol/L。

该样例中的“7年”修饰“口渴”、“多饮”、“多尿”三个实体，表示症状持续的时间，标注时需要将“7年”分别与“口渴”、“多饮”、“多尿”建立修饰关系。

③体型肥胖，2型糖尿病可能性大，青年患者不排除1型糖尿病可能性大。

这种情况下，“不排除”与“可能性大”均为可能的，本文的处理是“不排除”不标，如果只有“不排除”作为修饰则保留。

(2)患者拒绝医生建议的标注界定

如果出现患者拒绝某种治疗，后续的结果与该治疗无关，因此，本文的处理是，病历文本中所提到的治疗实体无需标注。

④患者双肺多发磨玻璃密度结节，请呼吸科会诊，建议在CT引导下型肺穿刺活检术，患者拒绝，要求暂查血真菌感染等相关指标。

该样例中的“肺穿刺活检术”和“CT”则不需要标注。

(3)身体部位的作用范围及标注说明

部位是症状的重要信息，无法把二者严格分开，所以需要根据上下文，确定症状和部位是否合并。如果症状和部位直接相邻没有被标点符号（主要是指逗号、句号、冒号、顿号、分号，并且，圆括号，尖括号，引号这三个符号不起分隔作用）等隔开的话，就把症状和部位合并到一起标注为一个完整的症状，否则需要标注部位和症状。

⑤头颅无畸形、压痛、包块。

该样例中，首先是否定修饰词“无”分别与“畸形”、“压痛”、“包块”建立修饰关系，“头颅”表示身体部位，其作用范围除了“畸形”之外，还包括“压痛”、“包块”，该情况下，本文的处理是将部位“头颅”与后面的三个症状分别建立位置关系。

4 语料库构建过程及构建结果

4.1 数据预处理

本文构建语料库的初始语料来自于临床电子病历，最初获得的电子病历数据主要包括病程记录、入院记录、患者病情评估、手术记录、其他记录和知情文件六种类型。对该批数据进行分析，保留后缀为“入院记录”、“病程记录”、“其他记录”的文件，其中一位病人的病历文件中包含“入院记录”和“病程记录”各一份，“其他记录”四份。

对“入院记录”、“病程记录”、“其他记录”这三类数据进行内容和格式分析，其中“入院记录”无需进行处理，对“病程记录”、“其他记录”这两类数据的处理方式如下：

(1)“病程记录”包含首次病程记录和查房记录，也包含一些噪音数据，需要将“病程记录”拆分为“首次病程记录”和“查房记录”，并过滤掉一些标记信息，此外，原始数据中还会存在一些格式及标点符号的问题，为了方便后续标注，对病历文本进行了统一的格式处理。

(2)“出院记录”包含“出院小结”及“出院医嘱”，都包含在“其他记录”中，每份电子病历中有四份“其他记录”，“出院医嘱”较为典型的是包含用药指导关键词。“出院小结”较为典型的是包含基本信息、入院日期、出院日期、诊疗经过等信息。仅将“其他记录”中的“出院小结”和“出院医嘱”保留进行标注，另外两份“其他记录”则删除。

对原始电子病历数据分析发现，存在缺失的情况，因此需要进行数据的筛选，筛选出完整的电子病历数据。共筛选出糖尿病电子病历文本800份，其中包括“入院记录”、“病程记录”各200份，“其他记录”共400份。

在标注之前需要将筛选好的电子病历数据进行去隐私化，病历文本中包含患者和医生的隐私信息，患者的隐私信息主要包括姓名、证件号码、家庭住址、工作单位，医生的隐私信息主要是姓名，此外，病历文本中出现的医疗机构名称也是隐私信息。本文主要参考文献 (Zhao et al., 2018)的方法进行数据去隐私化。

4.2 标注平台及标注过程

为了提升标注人员的标注效率，对张坤丽等人 (2020)开发的实体关系标注平台进行二次开发部署，使之适用电子病历实体及关系的标注，如图 1所示。在标注平台上标注人员可以选中不同的电子病历实体及关系进行标注，其中不同类别的实体使用不同的色块表示，方便区分，两个实体之间的关系显示在二者之间。平台提供了文件管理功能，可以查看一标、二标等完成进度情况，方便进行多轮标注工作，此外，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。

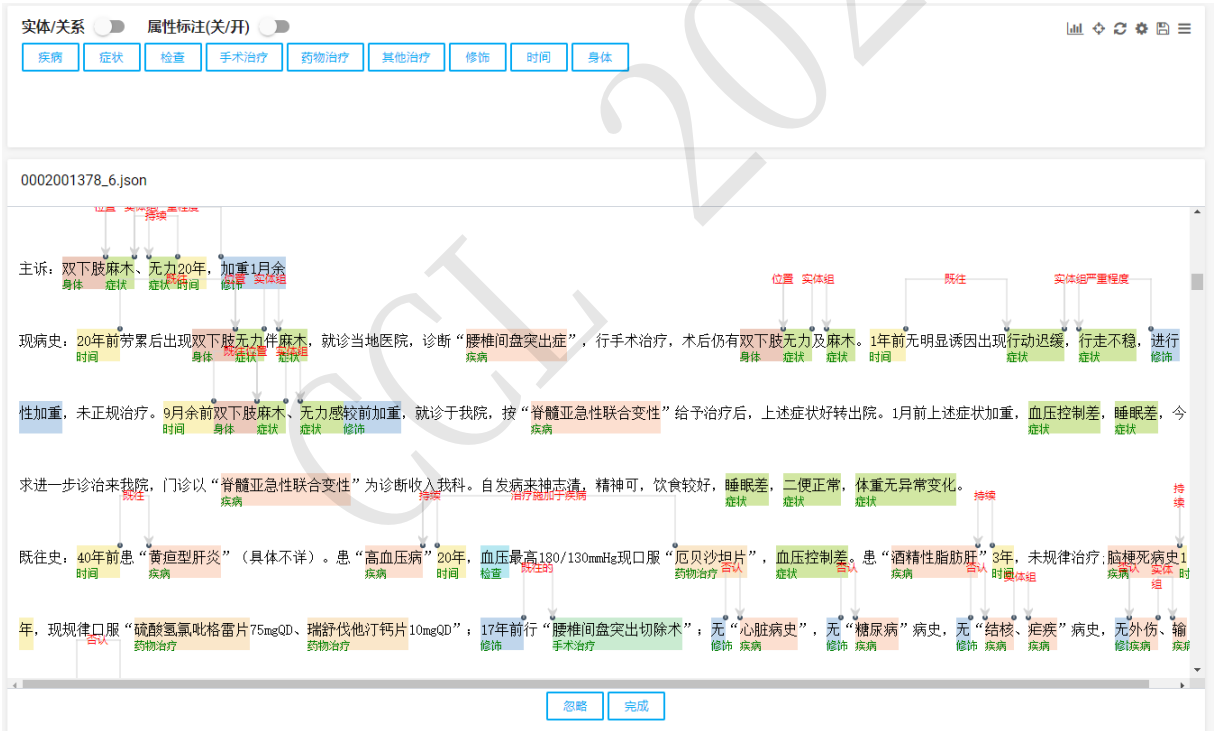

图 1. 标注平台

本文所进行的实体关系的标注建立在实体标注的基础之上，考虑到实体标注的质量会影响到关系标注的质量，因此，本文的标注分为两个过程，先进行实体的标注，实体标注完成之后，再进行实体关系的标注，并将每个过程分为预标注和正式标注两个阶段。

在预标注阶段，组织标注人员学习标注规范，同时组织标注人员进行预标注，使得标注人员熟悉电子病历实体及关系标注规范，并收集在实际标注过程中出现的问题。两轮预标注后，经过与医学专家讨论，进一步对标注规范进行完善，使标注规范更贴近本次研究任务，为正式

标注的顺利进行打下基础。

在正式标注阶段，标注过程采取多轮迭代模式，即每一个电子病历文本由两名标注人员负责。一标人员完成标注任务后，记录下存在疑问的地方，接着由二标人员在此基础上进行标注并且记录下不一致和不确定的地方，与医学专家讨论后获得统一的解决方案。讨论之后再由一标者负责修改语料，形成最后的三标文件。在这个阶段，会继续根据标注人员标注时的反馈意见修改标注规范，使标注规范更加适用于电子病历文本。具体的标注过程如图 2所示。为了对标注结果有更直观的展示，原始语料与标注之后的语料对比如图 3所示。

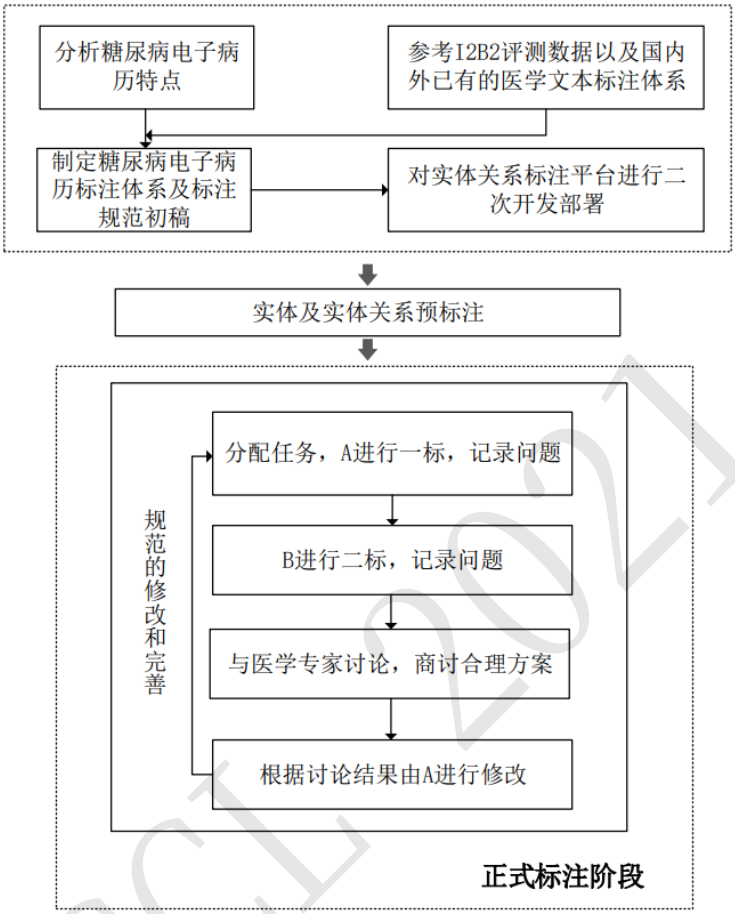

图 2. 实体及实体关系标注流程图

4.3 语料库统计与分析

经过多轮迭代标注，对标注好的数据进行统计，所构建的DEMRC中实体及实体关系数目分别如图 4、图 5所示，从实体数目来看，症状将近占据总实体数目的37.90%，其次是检查与疾病分别占实体总数的21.90%和11.10%，从关系数目来看，“检查-症状”和“药物治疗-疾病”这两类关系最多，这是因为临床电子病历数据中存在大量否定修饰的症状，大多出现在检查部分，此外，糖尿病存在一些并发症。

语料库构建一般选用Kappa值 (Carletta, 1996)和F1值 (Carletta, 1996)作为标注一致性的评价指标。在人工构建实体或关系标注语料库构建的研究中，通常使用F1值计算标注一致性，具体做法是，将一标人员(A1)的标注结果作为标准答案，计算二标人员(B1)标注结果的准确率(P)、召回率(R)以及F1值。在将独立标注语料与最终语料比较时，将最终语料视作A1，即为标准答案，独立标注语料视为B1。

采用上述方法，对本文构建的DEMRC进行一致性计算，其中，命名实体识别一致率达到了0.8562，实体关系一致率达到了0.9416，文献 (Artstein et al., 2008)指出，标注一致性达到0.8以上时，可以认为语料的一致性是可信赖的。

```
"text": "主诉：发现血糖升高10余年，视物模糊、手足麻木7月，呕吐、纳差2天。",
"spo_list": [
    {
        "Combined": false,
        "predicate": "时间|持续|症状",
        "subject": "10余年",
        "subject_type": "时间",
        "object": {
            "@value": "血糖升高"
        },
        "object_type": {
            "@value": "症状"
        }
    },
    {
        "Combined": false,
        "predicate": "时间|持续|症状",
        "subject": "7月",
        "subject_type": "时间",
        "object": {
            "@value": "手足麻木"
        },
        "object_type": {
            "@value": "症状"
        }
    },
    {
        "Combined": false,
        "predicate": "时间|持续|症状",
        "subject": "2天",
        "subject_type": "时间",
        "object": {
            "@value": "纳差"
        },
        "object_type": {
            "@value": "症状"
        }
    }
]
```


图 3. 标注结果对比图

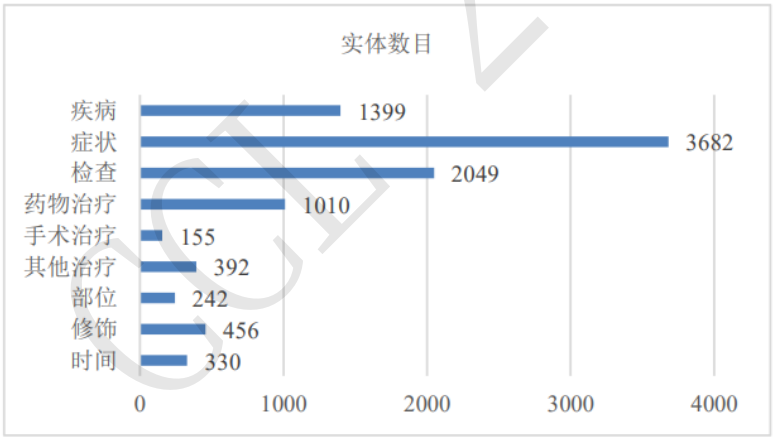

图 4. DEMRC实体数目统计

5 实体及关系抽取初步实验

为了分析语料库对模型性能的影响，对本文构建的DEMRC分别进行了实体识别及关系抽取初步实验，并根据实验结果对所构建的语料库进行详细评估。

针对实体识别任务，使用序列标注模型T-BiLSTM-CRF (Zhang et al., 2020)进行实验，对标注好的数据进行分句和去重处理，将10,484条实体标注数据按照8:1:1随机进行划分，其中，8,389作为训练集，1,048条作为验证集，1,047条作为测试集，并将数据转换为序列格式。整个网络初始的学习率为3e-3，字向量维度为100，批量的大小为32，Dropout比率为0.5，采用Adam优化算法进行40轮训练。表 3为实体识别结果。

通过表 3可以看出，时间的识别属于通用领域的命名实体识别任务，识别效果最好，其次

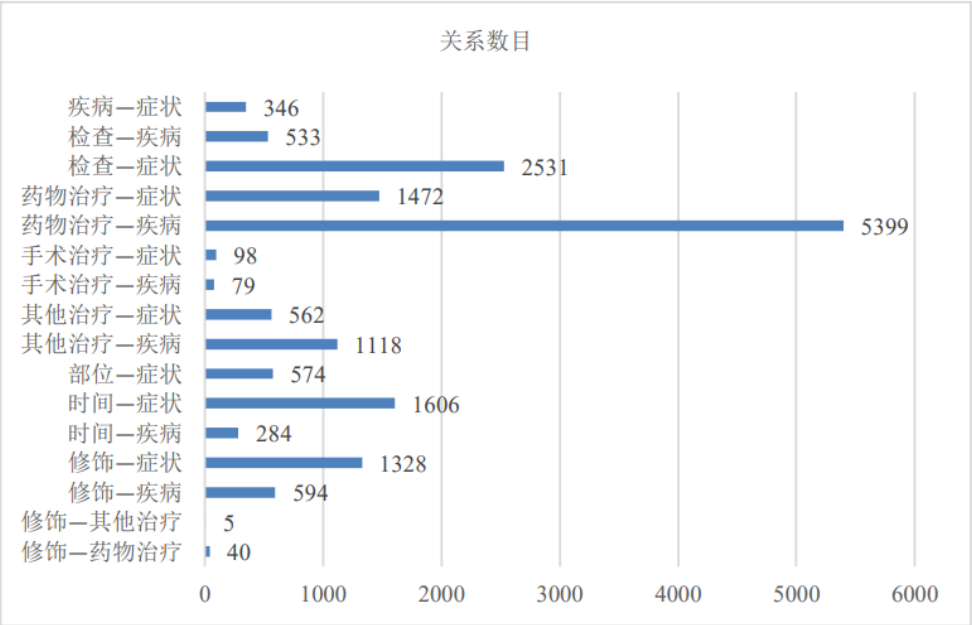

图 5. DEMRC实体关系数目统计

实体类型	P(%)	**R(%)**	**F1(%)**
疾病	76.03	72.34	74.14
症状	83.44	80.47	81.93
修饰	81.71	82.50	82.11
时间	89.78	89.78	89.78
药物治疗	86.58	88.31	87.44
检查	82.93	86.85	84.85
部位	83.48	85.42	84.44
手术治疗	71.43	83.33	76.92
其他治疗	86.58	70.51	77.72
总体	82.47	82.28	82.37

表 3. T-BiLSTM-CRF实体识别结果

是药物治疗、检查和身体部位的识别，因为这三类实体的实体边界相对确定，而疾病、手术治疗和其他治疗的识别性能较低，这与我们在标注语料时情况相似，手术治疗和其他治疗的数量较少，标注时容易漏标，而有一些疾病与症状不太容易区分而被误标。

针对实体关系抽取任务，本文使用RoBERTa (Liu et al., 2019)模型进行实验，主要是采用RoBERTa作为预训练模型，进行头实体、尾实体及关系预测。整个网络初始的学习率为3e-5，字向量维度为310，批量的大小为16，Dropout比率为0.5，进行30轮训练。对标注好的数据进行预处理，将得到的6,380条实体关系标注语料数据按照8:1:1的比例划分，5,104条作为训练集，638条作为验证集，638条作为测试集，表 4为实体关系实验结果。

从表 4可以看出，修饰-疾病、时间-疾病、药物治疗-疾病、时间-症状以及修饰与症状之间的关系抽取效果较好，针对修饰-疾病和修饰-症状，是因为疾病和症状中存在大量的否定修饰，时间-疾病和时间-症状之间的修饰关系中，仅包含既往和持续两类修饰关系，因此这类关系较好识别，对于药物治疗-疾病之间的关系，从图 5的数据统计可以看出，语料中此类关系数据量较大，模型可以得到很好的训练，故识别效果较好，这些和人工标注时的情况一致。

针对关系抽取效果较低的关系类型，对其进行分析发现，主要存在以下原因，一方面是标注数据的问题，在原始语料中该类实体关系存在较少，模型不能得到很好的训练，从而导致关系抽取效果较差。另一方面，关系抽取不仅仅是预测出实体对之间存在关系，还需要预测出关

实体关系类型	P(%)	**R(%)**	**F1(%)**
时间-症状	62.25	66.40	64.26
修饰-症状	67.06	72.35	69.60
药物治疗-症状	10.32	29.55	15.29
检查-症状	52.91	61.29	56.79
检查-疾病	27.69	32.73	30.00
修饰-疾病	83.98	82.66	83.31
部位-症状	42.97	43.19	43.08
其他治疗-症状	32.71	40.23	36.08
时间-疾病	75.0	79.12	77.01
药物治疗-疾病	84.92	63.14	72.43
其他治疗-疾病	62.39	39.67	48.50
修饰-药物治疗	40.0	66.67	50.0
手术治疗-疾病	20.0	33.33	25.0
疾病-症状	9.091	57.14	15.69
总体	60.21	62.70	61.43

表 4. RoBERTa实体关系抽取结果

系是什么，对于数据量少，但关系子类较多的实体对，由于子类关系之间会存在一定的相似性，从而会增加模型从多个候选关系中抽取出正确关系的难度。

根据实体识别及关系抽取初步实验，可以根据实验结果对人工标注过程进行调整，比如对于识别效果不好的几类关系，分析其特点，对标注人员着重进行培训，以此来提高标注的质量。通过以上的实验结果可以看出，数据量的大小会影响模型的性能，未来将会考虑如何提升少样本数据的关系抽取效果。

6 结论

本文在对糖尿病电子病历特点进行分析的基础上，并参考I2B2 2010的类型定义，制定了标注规范，遵循这一规范，建立了糖尿病电子病历标注规程以及标注一致性控制方案。本文重点介绍了糖尿病电子病历实体及关系标注体系和语料库的构建过程，经过对糖尿病电子病历标注方案的不断完善及多轮标注，该语料库已完成8,899个实体、456个实体修饰及16,564个关系的标注。在此基础上，对该语料库进行一系列的数据统计和标注一致性分析。并利用T-BiLSTM-CRF模型和RoBERTa模型分别对糖尿病实体及实体关系语料库进行评估，为今后的电子病历信息抽取研究以及糖尿病知识图谱构建打下基础。

参考文献

Artstein, Ron, Poesio, and Massimo. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

C. Brockett and W. B. Dolan. 2005. Support vector machines for paraphrase identification and corpus construction.

L. Campillos, L Deléger, C. Grouin, T. Hamon, and A Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources & Evaluation*, 52(2):1–31.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*.

W. B. Dolan and C. Brockett. 2005. Automatically constructing a corpus of sentential paraphrases.

X. Fei and M. Yetisgen-Yildiz. 2009. Clinical corpus annotation: Challenges and strategies. *yildiz*.

J. Lei, B. Tang, X. Lu, K. Gao, J. Min, and X. Hua. 2014. Research and applications: A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association Jamia*, 21(5):808.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Stéphane Meystre and Peter J. Haug. 2006. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599.

M. Morita, Y. Kano, and T. Ohkuma. 2013. Overview of the ntcir-10 mednlp task.

A Névéol and S. Rosset. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sunghwan, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association Jamia*, 17(5):507.

M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. 2001. Snomed clinical terms: overview of the development process and project status. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, page 662.

V. Vincze, G. Szarvas, R. Farkas, G MRa, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *Bmc Bioinformatics*, 9(Suppl 11):S9–S9.

H. Wang, W. Zhang, Q. Zeng, Z. Li, K. Feng, and L. Liu. 2014. Extracting important information from chinese operation notes with natural language processing methods. *Journal of Biomedical Informatics*, 48(C):130–136.

K. Zhang, D. Yue, and L. Zhuang. 2020. Improving chinese clinical named entity recognition based on bilstm-crf by cross-domain transfer. In *HPCCT & BDAI 2020: 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*.

Y. S. Zhao, K. L. Zhang, H. C. Ma, and K. Li. 2018. Leveraging text skeleton for de-identification of electronic medical records. *Bmc Medical Informatics & Decision Making*, 18(Suppl 1):18.

Uzuner Zlem, B. R. South, S. Shen, and Scott L Duvall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association Jamia*, (5):552.

齐梦霁刘勇. 2020. 基于糖尿病防治的医学知识图谱构建的研究. 医学信息, v.33;No.521(18):19–22.

周惠巍, 杨欢, 徐俊利, 张静, and 亢世勇. 2017. 中文模糊限制信息范围语料库的研究与构建. 中文信息学报, 31(3).

张坤丽, 赵旭, 关同峰, 尚柏羽, 李羽蒙, and 昝红英. 2020. 面向医疗文本的实体及关系标注平台的构建及应用. 中文信息学报, (6):36–44.

昝红英, 刘涛, 牛常勇, 赵悦淑, 张坤丽, and 穗志方. 2020. 面向儿科疾病的命名实体及实体关系标注语料库构建及应用. 中文信息学报, (5):19–26.

杨锦锋, 关毅, 何彬, 曲春燕, 于秋滨, 刘雅欣, and 赵永杰. 2016. 中文电子病历命名实体和实体关系语料库构建. 软件学报, 27(11):2725–2746.