面向医学文本处理的医学实体标注规范

张欢[1,2]，**宗源**[1,2]，**常宝宝**[1,2]，**穗志方**[1,2]，
昝红英[2,3]，**张坤丽**[2,3]
（1. 北京大学计算语言学教育部重点实验室，北京100871；
2. 鹏城实验室，广东深圳518055
3. 郑州大学信息工程学院，河南郑州450001）

摘要

随着智慧医疗的普及，利用自然语言处理技术识别医学信息的需求日益增长。目前，针对医学实体而言，医学共享语料库仍处于空白状态，这对医学文本信息处理各项任务的进展造成了巨大阻力。如何判断不同的医学实体类别？如何界定不同实体间的涵盖范围？这些问题导致缺乏类似通用场景的大规模规范标注的医学文本数据。针对上述问题，该文参考了UMLS中定义的语义类型，提出面向医学文本信息处理的医学实体标注规范，涵盖了疾病、临床表现、医疗程序等9种医学实体，以及基于规范构建医学实体标注语料库。该文综述了标注规范的实体体系、标注细则、混淆处理、语料标注以及医学实体自动标注基线实验等相关问题，希望能为医学实体语料库的构建提供可参考的标注规范，以及为医学实体识别提供语料支持。

关键词： 智慧医疗 ；医学实体 ；标注规范 ；标注语料

Medical Entity Annotation Standard for Medical Text Processing

ZHANG Huan[1,2]，**ZONG Yuan**[1,2]，**CHANG Baobao**[1,2]，
SUI Zhifang[1,2]，**ZAN Hongying**[2,3]，**ZHANG Kunli**[2,3]
（1. Key Laboratory of Computational Linguistics，Ministry of Education，
Peking University，Beijing 100871，China;
2. Peng Cheng Laboratory，Shenzhen，Guangdong 518055，China;
3. School of Information Engineering，Zhengzhou University，
Zhengzhou，Henan 450001，China）

Abstract

With the popularization of smart healthcare, the demand of applying natural language processing technology to identify medical information is increasing day by day. At present, there is no unified annotation standard for medical named entities in China, and the medical shared corpus is still in a blank state, which causes great resistance to the progress of medical text information processing tasks. How to judge different categories of medical entities? How to define the coverage of different entities? These problems lead to the lack of a similar mass of general scenario standard of medical text data. In view of the above problems, We referred to the semantic types defined in UMLS and proposed a unified medical entity annotation standard for medical text processing, covering 9 kinds of medical entities such as disease, symptom, medical procedure and so on，and constructed medical entity annotated corpus based on standards. This paper summarizes related issues such as the entity system, annotation principles, obfuscation processing, corpus annotation process, and medical entity automatic labeling baseline experiments, hoping to provide reference for medical entity corpus build annotating standard, as well as the medical support the corpus entity recognition.