·医学信息学·

医学知识图谱自动构建研究

马 浩,戴国琳,刘新遥,万艳丽
(中国医学科学院医学信息研究所,北京 100020)

摘要:随着我国医疗信息化水平的发展,大量医疗数据迅速产生,对如此丰富的医学知识进行有效利用显得尤为重要。医学知识图谱提供了对医学知识进行合理组织、管理和利用的手段。本文首先介绍了知识图谱的定义及架构,之后从实体抽取、关系抽取和实体对齐三个方面梳理了医学知识图谱自动构建过程的常用技术,分析了该领域的研究现状及最新技术进展,总结了医学知识图谱在医疗语义搜索引擎、医疗问答系统及医疗决策支持系统方面的应用现状,并分析了医学知识图谱自动构建及应用方面面临的挑战,旨在为医学知识图谱的构建及应用提供借鉴。

关键词:医学知识图谱;自动构建技术;自然语言处理
中图分类号:TP18 **文献标识码**:B **DOI**:10.3969/j.issn.1006-1959.2022.04.003
文章编号:1006-1959(2022)04-0010-04

Automatic Construction of Medical Knowledge Graph

MA Hao,DAI Guo-lin,LIU Xin-yao,WAN Yan-li
(Institute of Medical Information,Chinese Academy of Medical Sciences,Beijing 100020,China)

Abstract:With the development of our country's medical informatization level, a large amount of medical data has been rapidly generated, and it is particularly important to effectively use such a wealth of medical knowledge. The medical knowledge graph provides a means to organize, manage and utilize medical knowledge. This article first introduces the definition and architecture of the knowledge graph, and then introduces the common techniques in the medical knowledge graph automatic construction steps including entity extraction, relationship extraction and entity alignment. The research status of knowledge graph automatic construction techniques is summarized, and the latest research progress is analyzed. Finally, the application status of medical knowledge graph in medical semantic search engines, medical question answering systems and medical decision support systems are introduced, the challenges faced in the automatic construction techniques and applications of medical knowledge graph are analyzed, in order to provides a reference for the construction and application of medical knowledge graph.

Key words:Medical knowledge graph;Automatic construction techniques;Natural language processing

随着我国医疗技术的发展和医疗领域信息化水平的提升,生物医学文献、电子病历等大量的数据迅速产生,这给医学的发展提供了重要的资源。如何合理、有效地利用海量医学数据成为了一项重要的研究课题。知识图谱最早是谷歌的一个知识库,它使用语义检索来提高谷歌搜索的质量[1]。知识图谱的基本组成是"实体-关系-实体"三元组和"实体-属性-属性值"对,其具有强大的语义处理能力,能够对医学知识进行合理的表示及利用,为医学的发展提供有力支持。医学知识图谱也是知识图谱应用的重要领域之一,目前医学领域经典的医学知识图谱有北京大学、郑州大学和鹏城实验室构建的中文医学知识图谱[2]、上海曙光医院构建的中医药知识图谱[3]、中国中医科学院构建的中医临床知识图谱[4]、中医养生知识图谱[5]等。本文主要对医学知识图谱的自动构建情况进行总结,以期为医学知识图谱的构建及应用提供借鉴。

1 医学知识图谱构建

构建医学知识图谱首先需要从非结构化、半结构化的数据源中,通过知识抽取和知识融合技术得到结构化的知识并将其存储于数据库中,形成的医学知识图谱可以支持构建医疗语义搜索引擎、医疗问答系统和医疗决策支持系统,具体构建流程见图1。

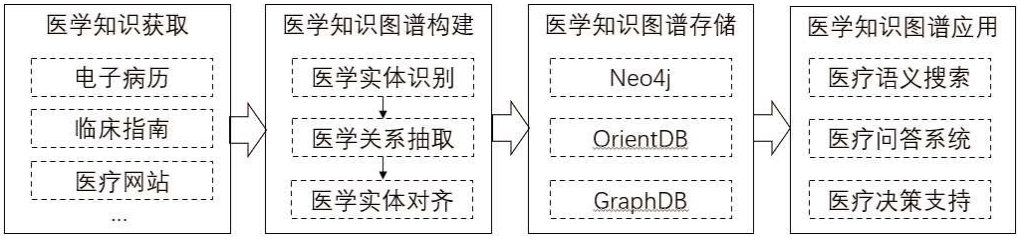

图1 医学知识图谱构建流程

基金项目:国家自然科学基金项目(编号:61971446)
作者简介:马浩(1998.1-),男,河南开封人,硕士研究生,主要从事医学知识图谱构建及应用研究
通讯作者:万艳丽(1978.2-),女,河南濮阳人,博士,副研究员,主要从事医疗卫生决策支持研究

1.1 命名实体识别 命名实体识别的概念在 1996 年的 MUC-6 会议上被提出，当时用来识别文本中的人名、机构名和地理位置[6]。在医学领域中，命名实体主要包括疾病名、药物名等。命名实体识别早期一般采用基于规则和词典的方法，此种方法可以取得较高的准确率，但召回率较低，规则构建的难度较大，迁移性较差。机器学习的方法一般把命名实体识别看作是序列标注任务，常用的模型有支持向量机（support vector machine，SVM）[7]、条件随机场（conditional random field，CRF）[8]等。机器学习的方法需要手工构建特征，构建过程费时费力，且这些特征往往不能扩展到其他任务。

深度学习的方法可以自动提取有效特征，不需要人工干预，很好的节省人力。目前在命名实体识别任务中最经典的深度学习方法是 BiLSTM-CRF 模型[9]。该模型的输入为经过预训练得到的词向量，通过前向和后向的 LSTM 层提取特征，最后经过 CRF 层得到标签序列。但是该模型也有一些缺陷，比如 BiLSTM-CRF 模型不能利用全局的上下文信息。对此，曾青霞等[10]在模型中加入注意力机制，在 CCKS2018 和 CoNLL 数据集中进行了实验，发现加入了注意力机制后模型的准确率有所提高。此外，深度学习的方法需要大规模的标注数据，在数据规模较小的情况下表现不佳。陈美杉等[11]提出了一种 KNN-BERT-BiLSTM-CRF 模型，通过迁移学习的方法对肝癌患者提问的文本进行命名实体识别，发现该方法取得了较高的 F1 值，并且只需要少量的标注语料。

1.2 关系抽取 实体关系抽取任务在 1998 年的 MUC-7[12]会议上第 1 次被提出，该会议给出了 3 种关系模板：Location_of、Employee_of 和 Product_of。在医学领域的关系主要有疾病导致症状、检查证实疾病等。关系抽取的方法早期有基于共现和基于规则的方法。基于共现的方法比较简单，召回率高，但精确度较低。基于规则的方法准确率较高，但召回率较低，可移植性差。基于机器学习的方法可以分为有监督、半监督和无监督的方法。有监督的方法具有较高的准确率，但是依赖于有标注的语料库，半监督和无监督的方法可以减少对人工标注数据的依赖。

基于深度学习的方法也可以进行医学实体间的关系抽取。张志昌等[13]使用双向的 GRU 联合注意力机制进行中文电子病历中的关系抽取。丁龙[14]提出一种基于注意力机制的 BiGRU-CNN 模型进行电子病历中的关系抽取，与其他模型相比，该方法取得了最高的 F1 值。李青青等[15]提出了一种基于 Attention 机制的主辅多任务模型抽取生物医学实体间的关系，该方法能够利用多个任务之间的相关信息，提升生物医学关系抽取的效果。

以上方法都是流水线的方法，即先抽取实体再抽取实体之间的关系，这种方法会存在错误传播的问题，并且无法充分利用两个任务之间的相关信息。牧杨子[16]使用 BiLSTM 模型进行中文电子病历的实体关系联合抽取，较好的完成了实体关系联合抽取任务。罗凌[17]提出一种新的标注策略来提取生物医学文本中的重叠关系，使用 Att-BiLSTM-CRF 模型对实体关系进行联合抽取，取得了优于流水线方法的结果。周侗[18]使用图卷积网络联合实体识别任务和关系抽取任务，进行中文电子病历的实体关系联合抽取，取得了很好的效果。

1.3 实体对齐 在医学知识图谱的构建过程中，医学实体“多词一义”的情况十分普遍，比如“帕金森症”还可表述为“帕金森障碍”“帕金森综合征”“PD”等。通过实体对齐工作可以对这些冗余的知识整合加工，提高知识的质量。实体对齐也可称为共指消解，其目标是发现多个知识库中指代现实世界中同一事物的实体，并将它们进行链接，从而可以进行多源知识的融合[19]。实体对齐可以通过基于属性相似度的成对实体对齐和考虑了实体间关系的集体实体对齐来实现。成对实体对齐常用方法有基于传统概率模型的方法和基于机器学习的方法等。集体实体对齐常用的方法有基于向量空间模型的方法、基于条件随机场模型的方法、基于相似性传播的方法等。

目前，基于知识表示学习的实体对齐方法是研究的热点。这种方法可以充分利用知识图谱中潜在的语义关系，有助于提高实体对齐的效果，具体的有基于翻译模型[20]的方法和基于图卷积神经网络[21]的方法。在医学领域，孙倩南[22]使用 TransE 算法对实体和关系进行嵌入，对不同数据源的呼吸科室医疗数据进行了实体对齐工作。滕飞等[23]在表示学习的基础上，根据医学知识的特点，加入词根集和规则用于医学实体对齐任务，提高了实体对齐的准确性。程瑞[24]通过图卷积网络对医疗知识图谱中的关系信息和结构信息进行建模，使用 TransE 对属性信息进行建模，最终将两者融合进行实体对齐，在 DBP15K 数据集上取得了较好的效果。

2 医学知识图谱应用

医学知识图谱能够对医学知识进行结构化表示并在此基础上进行查询与推理，目前主要应用于医疗语义搜索引擎、医疗问答系统、医疗决策支持系统等。

2.1 医疗语义搜索引擎 基于医学知识图谱的医疗语义搜索引擎可以准确地理解用户的搜索意图，提高用户的搜索体验，帮助用户快速找到自己感兴趣的内容。当用户进行查询时，语义搜索引擎可以将用

户查询的关键词映射到医学知识图谱中的概念之上，根据医学知识图谱中的概念层次结构进行推理，通过知识卡片的形式向用户返回相关的知识。目前谷歌、百度等搜索引擎都已经将知识图谱嵌入了搜索引擎。谷歌可以提供约400种健康状况的信息，当用户搜索疾病信息时，它可以通过信息卡片的形式展示疾病的特征。百度构建的知识图谱"知心"，可以用于支持用户对于医疗信息的搜索。受限于医学知识图谱的规模和质量，目前基于知识图谱的医疗语义搜索引擎的应用范围和效果仍有待进一步提高。

2.2 医疗问答系统 医疗问答系统是搜索系统的一种高级形式，可以通过自然语言来准确地回答用户的问题。对于用户提出的问题，基于知识图谱的医疗问答系统首先通过命名实体识别、关系抽取等自然语言处理技术对用户的问句进行语义解析，理解用户的问题，然后生成知识图谱的查询语句在知识图谱中进行查询，最后向用户返回答案。目前医疗问答系统的产品如北京慧医明智科技有限公司的"慧医大白"还有国外的"沃森医生"都可以提供基于医学知识图谱的医疗问答。也有不少研究者对医疗问答系统的构建进行了探索，如康莉[25]基于构建的心血管病知识图谱，采用深度学习的方法进行语义解析，最终实现了心血管疾病知识的问答系统。曹明宇等[26]构建了原发性肝癌的知识图谱，并基于此构建了原发性肝癌知识问答系统，可以对肝细胞癌相关问题进行回答。但是目前仍没有较为成熟的医疗问答系统出现，知识图谱的完整性、系统理解用户问题的准确性、推理的准确性及系统能回答问题的复杂性等方面都有待提高。

2.3 医疗决策支持系统 基于医疗知识图谱，可以构建医疗决策支持系统进行自动诊断，根据症状和化验结果给出诊断和治疗方案，帮助医生减少误诊的发生，提高医疗工作的质量。基于医学知识图谱的医疗决策支持系统主要通过推理引擎来完成决策支持过程。当用户输入症状和检查结果，推理引擎根据知识图谱和用户的输入给出诊断结果或接下来的治疗方案。目前百度的"灵医"、阿里巴巴的"Doctor You"、腾讯的"觅影"，都可以为医生提供临床决策支持服务。国外的"沃森医生"可以提供针对肿瘤疾病的决策支持，目前已经在部分医院得到应用。Gong F等[27]利用知识图谱实现了对患者的用药推荐并取得了良好的效果。郑少宇等[28]基于医学教材、诊疗指南等知识源构建了常见病的知识图谱，基于此开发了对于常见病的诊断辅助系统，可以在主要临床环节有效地进行决策辅助。目前医疗决策支持系统一般只能对医疗决策提供辅助，其提供决策的准确性还有待加强。

3 总结

知识图谱已成为当前研究的热点，但由于医疗大数据具有专业性强，结构复杂等特点，医学知识图谱的自动构建和应用依然面临很大的挑战。在医学知识抽取环节，抽取算法的准确率普遍不高，限制条件较多，可扩展性不强。医学实体对齐算法的计算复杂度较高，实体对齐方法缺乏训练数据，多语言的实体对齐也较为困难。在医学知识应用方面，由于现有医学知识推理能力的限制，医疗决策支持系统的准确性暂时还不能满足临床辅助决策要求。

总之，医学知识图谱能够促进医学数据的有效利用，进而促进医学的发展。我国医疗信息化水平的发展及海量医学数据的产生为医学知识图谱的发展提供了契机。相信在不久的将来，随着医学知识图谱构建的发展，其将在医疗领域发挥更大的作用。

参考文献：

[1]Wikipedia.Knowledge graph [EB/OL].https://en.wikipedia.org/wiki/Knowledge_Graph,2021-05-09/2021-08-25.

[2]奥德玛,杨云飞,穗志方,等.中文医学知识图谱CMeKG构建初探[J].中文信息学报,2019,33(10):1-9.

[3]阮彤,孙程琳,王昊奋,等.中医药知识图谱构建与应用[J].医学信息学杂志,2016,37(4):8-13.

[4]于彤,李敬华,朱玲,等.中医临床知识图谱的构建与应用[J].科技新时代,2017(4):51-54.

[5]于彤,李敬华,于琦,等.中医养生知识图谱的构建与应用[J].中国数字医学,2017,12(12):64-66.

[6]Grishman R.Message understanding conference-6: a brief history[C]//Proceedings of the 16th conference on Computational linguistics.Copenhagen:ACL,1996:466-471.

[7]Isozaki H,Kazawa H.Efficient support vector classifiers for named entity recognition [C]//Proceedings of the 19th international conference on Computational linguistics.Taipei:COLING,2002:1-7.

[8]Mccallum A,Li W.Early results for named entity recognition with conditional random fields, feature induction and Web-Enhanced lexicons[J].Computer Science Department Faculty Publication Series,2003(4):188-191.

[9]Huang ZH,Xu W,Yu K.Bidirectional LSTM-CRF models for sequence tagging[EB/OL].http://arxiv.org/abs/1508.01991,2015-08-09/2021-08-25.

[10]曾青霞,熊旺平,杜建强,等.结合自注意力的BiLSTM-CRF的电子病历命名实体识别 [J]. 计算机应用与软件,2021,38(3):159-162,242.

[11]陈美杉,夏晨曦.肝癌患者在线提问的命名实体识别研究:一种基于迁移学习的方法 [J]. 数据分析与知识发现,2019,3(12):61-69.

[12]Chinchor N,Marsch E.MUC -7 Information Extraction Task Definition[C]//Proceedings of a Seventh Message Understanding Conference (MUC-7).Fairfax:ACL,1998.

（下转第29页）

(上接第 12 页)

[13]张志昌,周侗,张瑞芳,等.融合双向 GRU 与注意力机制的医疗实体关系识别[J].计算机工程,2020,46(6):296-302.
[14]丁龙.面向电子病历的信息抽取技术研究[D].衡阳:南华大学,2020.
[15]李青青,杨志豪,罗凌,等.基于多任务学习的生物医学实体关系抽取[J].中文信息学报,2019,33(8):84-92.
[16]牧杨子.基于半监督学习的中文电子病历实体识别和实体关系抽取研究[D].海口:海南大学,2018.
[17]罗凌.生物医学文本挖掘若干关键技术研究[D].大连:大连理工大学,2019.
[18]周侗.面向中文电子病历的医疗实体及关系识别技术研究[D].兰州:西北师范大学,2020.
[19]田江伟,李俊锋,柳青.结合属性结构的图卷积实体对齐算法[J].计算机应用研究,2021,38(7):1979-1982,1992.
[20]Chen MH,Tian YT,Yang MH,et al.Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]//Proc of the 26th International Joint Conference on Artificial Intelligence.Melbourne:IJCAI,2017:1511-1517.
[21]Wang Z,Lv Q,Lan X,et al.Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks [C]//Proc of the 2018 Conference on Empirical Methods in Natural Language Processing.Brussels:ACL,2018:349-357.
[22]孙倩南.面向呼吸科室疾病的知识抽取与对齐[D].哈尔滨:哈尔滨工业大学,2019.
[23]滕飞,钟文,许强,等.一种基于表示学习的医学知识图谱实体对齐方法:中国,CN111309930A[P].2020-09-15.
[24]程瑞.面向中文医疗知识图谱的实体对齐方法研究及应用[D].北京:北京邮电大学,2020.
[25]康莉.基于知识图谱的心血管病问答系统的研究与实现[D].广州:华南理工大学,2020.
[26]曹明宇,李青青,杨志豪,等.基于知识图谱的原发性肝癌知识问答系统[J].中文信息学报,2019,33(6):88-93.
[27]Gong F,Wang M,Wang HF,et al.SMR: Medical knowledge graph embedding for safe medicine recommendation [EB/OL]. https://arxiv.org/abs/1710.05980,2020-11-29/2021-08-25.
[28]郑少宇,滕飞,马征,等.支持临床决策的医学知识图谱的构建与应用[J].重庆医学,2021,50(1):163-166.

收稿日期:2021-08-25;修回日期:2021-09-16
编辑/成森