计算机科学
COMPUTER SCIENCE

医学知识图谱研究与应用综述

蒋川宇, 韩翔宇, 杨文蕊, 吕博涵, 黄小欧, 谢夏, 谷阳

引用本文

蒋川宇, 韩翔宇, 杨文蕊, 吕博涵, 黄小欧, 谢夏, 谷阳医学知识图谱研究与应用综述[J]. 计算机科学, 2023, 50(3): 83-93.

JIANG Chuanyu, HAN Xiangyu, YANG Wenrui, LYU Bohan, HUANG Xiaoou, XIE Xia, GU Yang. Survey of Medical Knowledge Graph Research and Application [J]. Computer Science, 2023, 50(3): 83-93.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700241

医学知识图谱研究与应用综述

蒋川宇　韩翔宇　杨文蕊　吕博涵　黄小欧　谢　夏　谷　阳
海南大学计算机科学与技术学院　海口 570228
(cyhhyg@hainanu.edu.cn)

摘　要　医学数据数字化推进过程中,如何选择合适的技术来对医学数据进行高效处理和准确分析,是当今医学领域普遍面临的问题。利用具有优秀联想与推理能力的知识图谱技术来对医学数据进行处理与分析,能更好地实现智慧医疗、辅助诊断等应用。医学知识图谱的完整构建过程包括知识抽取、知识融合和知识推理。其中知识抽取可细分为实体抽取、关系抽取和属性抽取,知识融合则主要包括实体对齐和实体消歧。首先,对现今医学知识图谱的构建技术和实际应用进行归纳整理,针对每一具体构建过程阐明技术发展脉络。在此基础上,对相关技术进行介绍并说明其优点和局限性。其次,介绍几个已成熟运用的医学知识图谱。最后,根据知识图谱在医学领域的技术与应用现状,给出未来知识图谱可进行的技术兼应用性的研究方向。

关键词:医学;大数据;知识图谱;数据处理;知识图谱构建技术
中图法分类号　TP181

Survey of Medical Knowledge Graph Research and Application

JIANG Chuanyu,HAN Xiangyu,YANG Wenrui,LYU Bohan,HUANG Xiaoou,XIE Xia and GU Yang
School of Computer Science and Technology,Hainan University,Haikou 570288,China

Abstract　In the process of digitisation of medical data,choosing the right technology for efficient processing and accurate analysis of medical data is a common problem faced by the medical field today. The use of knowledge graph technology with the excellent association and reasoning capabilities to process and analyse medical data can better enable applications such as wise information technology of medicine and aided diagnoses. The complete process of constructing a medical knowledge graph includes knowledge extraction,knowledge fusion and knowledge reasoning. Knowledge extraction can be subdivided into entity extraction,relationship extraction and attribute extraction,while knowledge fusion mainly includes entity alignment and entity disambiguation. Firstly,the construction technologies and practical applications of medical knowledge graphs are summarised,and the development of the technologies is clarified for each specific construction process. On this basis,the relevant techniques are introduced,and their advantages and limitations are explained. Secondly,introducing several medical knowledge graphs that are being successfully applied. Finally,based on the current state of technology and applications of knowledge graphs in the medical field,future research directions for knowledge graphs in technology and applications are given.

Keywords　Medicine,Big data,Knowledge graph,Data processing,Knowledge graph construction technology

早期医学数据处理方法有数据降维、符号学习和表示学习等。数据降维是将高维空间数据映射到低维空间进行表示处理,以减少处理高维数据带来的计算成本。符号学习在机器学习中被用来从已知确定的数据中学习规则,再利用学习到的规则对未知数据进行预测[1-2]。表示学习是把数据表示为稠密低维的向量,使得在构建分类器或其他预测器时能更容易提取到有用的信息[3-4]。

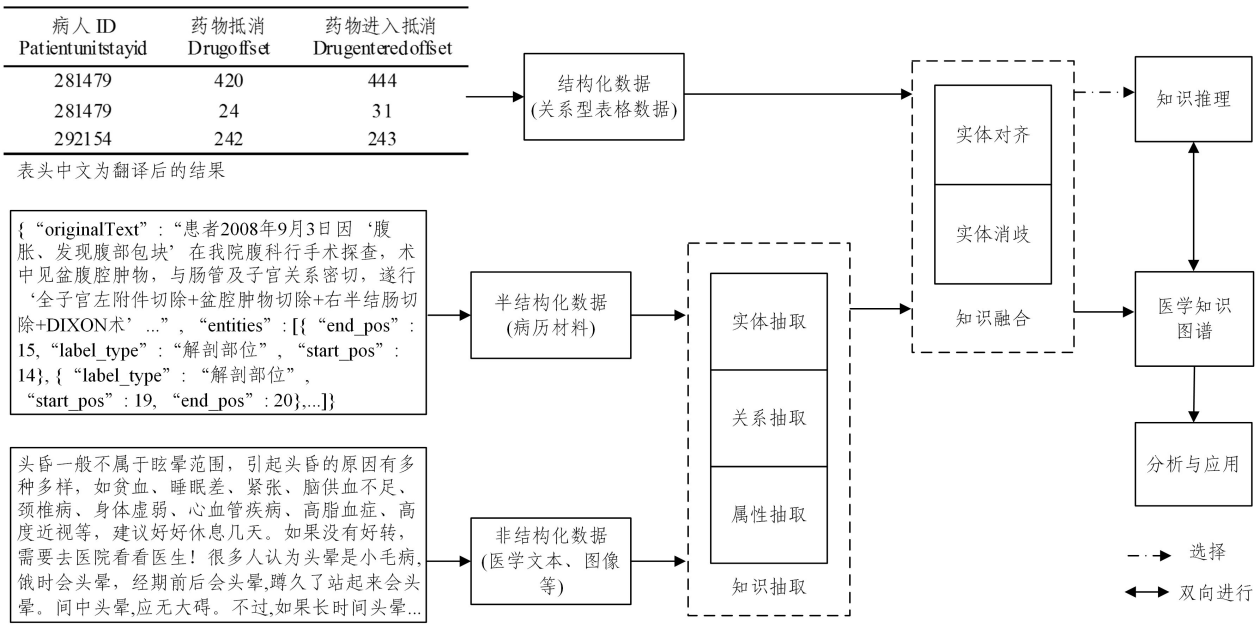
随着医学数据数字化进程的推进,神经网络、深度学习[5]等新方法效果显著。如 Zhao 等[6]通过构建基于卷积神经网络[7](Convolutional Neural Networks,CNN)的框架,更好地预测了主要组织相容性复合体与肽配体之间的结合亲和力。但现今在对海量医学数据进行处理时,仍存在利用率低[8]、数据处理困难[9]等问题。

知识图谱,一种有向图结构的知识库,涵盖了实体、概念及其相互间的语义关系,由谷歌于 2012 年提出[10]并运用到搜索引擎中。通过在医学领域引入知识图谱技术,可帮助解决目前医学领域的数据处理和实际应用需求。对比常规知识图谱,医学知识图谱实体鲜明、属性明确、数据量大且迭代更新快,同时还有关系复杂和逻辑性强的特点。医学知识图谱的构建如图 1 所示,其由知识抽取、知识融合和知识推理组成,其中比较重要的是知识抽取与知识融合两个过程。

到稿日期:2022-07-25　返修日期:2022-12-19
基金项目:医学大数据特殊病种诊疗模型与技术研究(ZDKJ2021042)
This work was supported by the Hainan Province Science and Technology Special Fund(ZDKJ2021042).
通信作者:谷阳(guyangl@163.com)

病人 ID Patientunitstayid	药物抵消 Drugoffset	药物进入抵消 Drugenteredoffset
281479	420	444
281479	24	31
292154	242	243

表头中文为翻译后的结果



注：结构化数据的网址为 https://eicu-crd.mit.edu/；半结构化数据的网址为 https://tianchi.aliyun.com/dataset/92085；非结构化数据的网址为 https://tianchi.aliyun.com/dataset/135962

图 1 医学知识图谱构建流程图

Fig. 1 Flowchart of medical knowledge graph construction

为帮助研究人员更好地在医学领域开展关于知识图谱技术的研究，本文对医学知识图谱完整构建过程中的相关技术进行综述。

1 医学知识图谱的知识抽取

1.1 实体抽取

实体抽取，又被称为命名实体识别（Named Entity Recognition，NER），主要是在杂乱的医学文本中抽取实体。由于医学数据具有复杂性，实际抽取过程通常存在抽取困难、准确度低等问题。医学知识图谱的实体抽取方法可分为[11]：基于词典和规则的实体抽取方法、基于概率统计的机器学习方法和基于深度学习的实体抽取方法。

（1）基于词典和规则的实体抽取方法。Coden 等[12]提出 CDKRM（Cancer Disease Knowledge Representation Model）来对癌症特征及其关系进行表示，其捕获了非结构化病理报告中有价值的癌症相关信息。通过将表示模型与其设计的信息提取系统 MedTAS/P 相结合，可完成非结构化自由文本病理报告中的实体抽取等任务，进而实现文本病理报告的自动填充。Savova 等[13]设计的 cTAKES（Clinical Text Analysis and Knowledge Extraction System），能对临床文本中的实体进行抽取。该系统由多个组件顺序组成，可灵活组装其他组件，自由度高。基于词典和规则的实体抽取方法依赖于专业领域词典和领域专家设计的规则，对数据量大、复杂性强的数据难以取得好的抽取结果。

（2）基于概率统计的机器学习方法。主流模型有：隐马尔可夫模型[14]（Hidden Markov Model，HMM）、最大熵马尔可夫模型[15]（Maximum Entropy Markov Model，MEMM）和条件随机场[16]（Conditional Random Field，CRF）等。Zhang 等[17]给出了生物医学领域中的一组特征，并提出了一种缩写识别方法和两种级联命名实体识别方法。这不仅为生物医学领域提供了丰富的特征集，同时也是级联命名实体识别的一次尝试。Wang 等[18]将命名实体识别任务转换成分类任务，进而提出基于 CRF 的临床命名实体识别。该方法取得了不错的识别效果，但仅在内科病历数据上进行了实验，具有局限性。Jonnalagadda 等[19]在对临床叙述抽取中首次引入了分布式语义，进而提出基于分布式语义的概念抽取方法，显著提高了概念抽取性能。基于概率统计的机器学习方法比基于词典和规则的实体抽取方法抽取准确度更高、效率更好。

（3）基于深度学习的实体抽取方法。在医学领域用于实体抽取的深度学习技术主要有 CNN、循环神经网络[20]（Recurrent Neural Network，RNN）等。Sun 等[21]将 Bi-LSTM-CRF 和汉字的部首与词源特征相结合，提出了端到端的中文医学命名实体识别方法，取得了不错的效果。Li 等[22]将 CNN-BLSTM-CRF 模型用于生物医学语料上的实体抽取，不仅提高了抽取性能，而且避免了对人工特征的依赖。Li 等[23]提出了 BiLSTM-Att-CRF 模型用于中文电子病历的临床命名实体识别，取得了良好的效果。该模型可更多地捕获上下文的有用信息，从而避免了远距离导致的信息丢失。Ji 等[24]构建了 Attention-BiLSTM-CRF 模型来对中文电子病历进行实体抽取，实现了利用文档级信息来缓解标记不一致性。之后 Ji 等[25]提出了一种基于多个神经网络模型的协同合作方法，该方法将两个 BiLSTM-CRF 模型与 CNN 模型进行结合，具有良好的泛化能力。基于深度学习的实体抽取方法利用了神经网络的并行性与信息综合能力等，进一步提高了抽取准确度。实体抽取方法具体如表 1 所列。

表 1 实体抽取方法

Table 1 Entity extraction methods

阶段	方法/模型	优点	不足	实验数据	应用	*Precision*	*Recall*	*F*1
基于词典规则的实体识别抽取	MedTAS/P[12]	能在非结构化的病理报告中填充癌症疾病知识表示模型	会受错误信息干扰；对不遵循通用语法规则的语言效果不佳	结肠癌病理报告	文本病症病理报告中的特征提取	0.78	0.82	0.80
	cTAKES[13]	构建了一个词汇相对丰富的查找字典，能较为快速地完成抽取任务	需要维护查找 NER 所需的字典；无法识别复杂的同义词实体	Penn TreeBank[26]等	临床文本分析与知识抽取	0.80	0.65	0.72
基于概率统计的机器学习方法	一种增强基于 HMM 的命名实体识别器的方法[17]	集成了丰富的生物医学领域特征，取得了更好的识别效果；首次在生物医学领域上采用了级联命名实体识别	对部分复杂结构无法进行有效处理	GENIA[27]	生物医学领域的命名实体识别	0.68	0.65	0.67
	基于 CRF 的临床命名实体识别方法[18]	可正确识别未出现在训练集中的实体	实验数据单一	电子病历语料库[1)]	中文电子病历的命名实体识别	0.91	0.87	0.89
	基于分布式语义的概念抽取[19]	提高了概念抽取准确度	分布式语义带来的问题未得到解决	i2b2/VA[28]	临床文本中的概念抽取	0.79	0.83	0.81
基于深度学习的方法	结合词根和词根特征的 Bi-LSTM-CRF[21]	模型不依赖专业领域词典和人工制定的规则与特征，迁移性强	效果提升不明显，侧重字符层面的特征提取	CCKS 2017[2)]	中文电子病历的命名实体识别	0.88	0.90	0.89
	CNN-BLSTM-CRF[22]	词向量结合 CNN 避免了人工特征，模型识别准确度高、泛化能力好	准确度易受词向量好坏的影响	Biocreative II[3)]	生物医学领域的命名实体识别	0.89	0.89	0.89
				JNLPBA2004[4)]		0.80	0.70	0.74
	BiLSTM-Att-CRF[23]	模型更多地捕获了上下文的有用信息，从而避免了远距离导致的信息丢失	识别新词的能力不足	CCKS 2017	中文电子病历的命名实体识别	0.90	0.90	0.90
				CCKS 2018[5)]		0.87	0.84	0.85
	Attention-BiLSTM-CRF[24]	引入注意力机制一定程度上缓解了标记不一致问题；实体自动更正算法减少了实体边界划分错误，再结合药品词典和后处理规则进一步提高了准确度	不太适用于前沿医学上的实体抽取	CCKS2018	中文电子病历的命名实体识别	0.91	0.90	0.91
	CNN＋BiLSTM-CRF(2)[25]	模型具有好的泛化能力	模型较复杂	CHIP 2018[6)]	中文电子病历的命名实体识别	0.86	0.86	0.86

1.2 关系抽取

关系抽取作为自动化构建知识图谱过程中的关键技术，是对未知关系事实进行提取并添加到知识图谱中，这些关系可将离散的医学实体联系起来，进而解决医学实体间语义链接问题[29-30]。医学领域中的关系抽取技术的发展分为早期基于规则的方法、基于传统机器学习的方法和基于深度学习的方法 3 个阶段。

(1)基于规则的方法。Zhu 等[31]建立了基于规则的知识抽取及可视化平台来对中医古籍中的疾病“崩漏”进行知识抽取，为中医疾病知识的逻辑化描述提供了方法。El-Halees 等[32]设计了信息抽取系统用于在医疗记录中抽取知识，首先将非结构化医学文本转换为结构化数据，接着使用关联规则揭示医疗数据间的关系。该系统可协助医务人员检测医疗数据间的隐藏关系，从而帮助制定更合理的患者护理方案。基于规则的关系抽取方法在词汇匮乏时抽取效率高，但是其严重依赖特定领域规则，导致整体关系抽取效果一般，且难以进行迁移。

(2)基于传统机器学习的方法。Nikfarjam 等[33]设计了机器学习与图推理机制相结合的系统，以提取临床记录中事件和时间表达间的关系。Zhao 等[34]在生物医学文献中的关系抽取上使用了支持向量机(Support Vector Machine，SVM)，并在特征表示上增加了语义谓词特征，该方法将准确率提高了 5%～10%。类似地，Roberts 等[35]设计了基于统计机器学习的关系抽取系统，同样使用 SVM 进行关系抽取。该系统可抽取肿瘤叙述语料库上的临床关系，促进了临床文本上信息挖掘的发展。基于传统机器学习的方法关系抽取准确度较高，但抽取过程易受训练数据和实体抽取结果的影响，这给实际应用带来了局限性，也影响了知识图谱的下游应用。

(3)基于深度学习的方法。Zheng 等[36]提出了新的标记方案，将实体和关系的联合抽取转变为标注问题，并基于此研究了端到端的模型来抽取实体和关系。该模型效果优于大多数流水线和联合学习方法，但对于重叠关系的识别存在不足。此外，标记方案与 CTD-BLSTM 模型相结合，可实现中医领域上的关系抽取[37]。Gao 等[38]提出了融合关系发现词和深度学习的医疗诊疗关系抽取模型，通过聚类和 TF-IDF 算法

1) http://www.easyemr.com.cn/activitys/activity-list.aspx?activityid=32

2) http://www.sigkg.cn/ccks2017/

3) https://biocreative.bioinformatics.udel.edu/tasks/biocreative-ii/

4) http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004

5) http://www.sigkg.cn/ccks2018/

6) http://icrc.hitsz.edu.cn/chip2018/Task.html

来抽取关系发现词，经向量表示后输入到深度学习模型 BIGRU-2ATT 中作为额外特征。该模型提高了抽取性能，但需人工参与，不适用于大规模语料上的抽取。基于深度学习的方法能更好地捕获待抽取关系的特征，从而提高关系抽取准确度，但使用深度学习技术往往需要大量标注数据进行训练，这增大了领域专家的工作量。此外，由于深度学习技术自身的可解释性存在不足，这也造成了相应抽取方法的可解释不强。关系抽取方法具体如表 2 所列。

表 2 关系抽取方法

Table 2 Relation extraction methods

阶段	方法/模型	优点	不足	实验数据	应用
基于规则的方法	基于正则表达式的知识抽取方法[31]	实现了中医古籍文献上的知识抽取	自动化程度较低；抽取性能有待提高	崩漏相关中医古籍文献的小型数据库	中医古籍文献中的知识抽取
	基于关联规则的方法[32]	能较好地提取关联规则	FP-growth 算法带来的内存开销等问题未得到解决	医疗报告	医疗报告中的知识抽取
基于传统机器学习的方法	SVM+图推理[33]	在临床记录上取得了良好的时间关系提取效果	过于依赖训练数据；SVM 和图推理的结合策略有待提高	n2c2 2012[39]	临床记录中提取时间关系
	基于 SVM 的关系抽取方法[34]	信息增益方法的使用和语义谓词特征的加入提高了抽取准确度	抽取的关系较少	医学文献摘要	生物医学文献中的实体关系
	基于 SVM 的关系抽取系统[35]	有良好的临床关系抽取准确度，处理速度快	准确度依赖于实体识别效果	C77	肿瘤语料库中的关系抽取
基于深度学习的方法	基于新标记方案的实体和关系联合抽取[36]	具有较好的抽取效果	针对重叠关系上的识别存在不足	NYT[40]	中医领域中的关系抽取
	融合关系发现词与深度学习的关系抽取模型[38]	将关系发现词作为 BiGRU-2ATT 模型的额外特征提高了抽取性能	需人工参与；影响模型性能的因素过多	医疗诊疗关系样本	医疗文本上的关系抽取

1.3 属性抽取

属性抽取，是在非结构化信息中抽取实体的属性信息，分为基于规则的属性抽取方法和基于机器学习的属性抽取方法。

(1)基于规则的属性抽取。Kersloot 等[41]提出了 DIRECT(Disease Information and Relationship ExtraCtion Tool)，它由 cTAKES，SNOMED CT 概念过滤器和概念关系检测算法等组成。该程序可完成对临床叙述的编码，从而节省临床医生等相关人员的时间。Mykowiecka 等[42]在对波兰语临床文本的信息提取上，开发了基于规则的信息抽取系统。系统在属性抽取上取得了不错的结果，可应用于患者数据收集和文档质量提升。基于规则的属性抽取，由于属性值结构具有不确定性，造成规则制定较为困难。此外不同医学知识图谱构建中对属性定义存在偏差，基于规则的方法难以适应，迁移性不强。

(2)基于机器学习的属性抽取。Jiang[43]提出了基于 LSTM 的多实例多标签的属性抽取算法，为避免有监督的方法需手工标注数据和无监督的方法无法适用于特定领域的问题，该算法使用远程监督的方法来生成样本。算法训练的模型能实现对目标实体属性值的抽取，实际应用中也完成了传染病相关的知识图谱构建。Shi 等[44]提出了一种新的联合深度学习方法来提取临床实体和属性，该方法避免了基于管道方法的错误传播问题，取得了更好的识别效果。Xu 等[45]提出一种基于序列标记的方法来检测不同概念的属性，推动了临床自然语言处理系统的发展。Du 等[46]提出一种基于 ALBERT 的属性抽取框架，相比传统深度学习模型进一步提高了抽取准确度，在实际应用场景中也取得了不错的表现。基于机器学习的属性抽取更好地利用了上下文信息，从而提高了属性抽取性能，但应用上仍存在局限性。属性抽取方法具体如表 3 所列。

表 3 属性抽取方法

Table 3 Attribute extraction methods

阶段	方法/模型	优点	不足	实验数据	*Precision*	*Recall*	*F*1
基于词典规则的实体识别抽取	DIRECT[41]	可高精度地检测肿瘤学概念和属性关系	算法不能判别检测到的关系在临床上是否正确	98 份肺癌患者的治疗进度记录	1.00	0.75	0.86
	基于规则的信息信息抽取系统[42]	解决了诸多语言问题，取得了不错的实验结果	构建相对困难	乳房 X 光检查报告、糖尿病患者的医学记录	0.99	0.96	0.98
基于机器学习的属性抽取	基于 LSTM 的多实例多标签的属性抽取算法[43]	远程监督的方法解决了人力标注成本高的问题，可适用于特定领域的属性抽取；使用词向量作为特征输入降低了整个任务的复杂度	远程监督带来的错误标注问题未得到解决	TAC Knowledge Base Population (KBP) 2015[1)]	0.56	0.41	0.47
	联合深度学习方法[44]	避免了基于管道方法的错误传播问题	错误识别率较高	中文电子病历	0.90	0.88	0.89
	基于 Bi-LSTM-CRFs 的序列标记方法[45]	相比级联方法取得了更好的性能	特定类型的属性难以被检测	ShARe 语料库[47]等	0.81	0.83	0.82
	基于 ALBERT 的属性抽取框架[46]	重视属性关联所带来的信息	适应范围待扩大	病理诊断报告	—	—	—

1) https://tac.nist.gov/2015/KBP/index.html

2 医学知识图谱的知识融合与知识推理

2.1 实体对齐

实体对齐旨在找到不同医学知识图谱中表征现实同一事物的实体，是医学知识融合的关键技术之一。医学知识图谱的实体对齐，可分为传统的实体对齐方法和基于表示学习的实体对齐方法。

(1)传统的实体对齐方法。Jean-Mary 等[48]提出 ASMOV(Automated Semantic Matching of Ontologies with Verification)算法用于本体匹配，提升了本体匹配的效果。该算法主要包括相似性计算和语义验证两个部分，其中相似性计算部分是对词汇、实体集等的相似度进行计算，语义验证部分则是在预对齐结果上再进行语义验证。Jiménez-Ruiz 等[49]设计了基于逻辑规则的可扩展本体匹配系统 LogMap，能适应大规模的本体匹配，但过于依赖从本体中提取的词汇特征。传统的实体对齐方法侧重于字符层面上的实体对齐，使用领域词典和人为设计的规则，可适用常规实体对齐任务，但对于跨领域和复杂的实体对齐任务仍存在不足。

(2)基于表示学习的实体对齐方法。Ma 等[50]在中文医疗实体对齐上提出 SiBERT 方法，增强了对齐实体的特征表示，提高了对齐准确度和计算效率，但仍存在应用上的局限性。Wang 等[51]基于混合图注意力网络提出了生物医学本体匹配框架 BioHAN，使用 BioBERT 进行向量嵌入，然后将嵌入转移至双曲空间以捕获层次特征。在此基础上，进行多跳相邻聚合来获得良好的本体特征表示，取得了较好的本体匹配效果。Hao 等[52]提出的 MEDTO 模型同样使用 BioBERT 进行向量嵌入。在此基础上，使用双曲图卷积神经网络[53](Hyperbolic Graph Convolutional Neural Network，HGCN)和异构图层分别捕获层次结构信息和非层级结构信息，进而完成本体匹配任务。基于表示学习的实体对齐方法结合了多视图信息，能够更好地捕获医学对齐实体的语义特征。方法能取得不错的对齐结果，部分模型也具有迁移性，但也存在过分依赖标签数据和训练开销大等问题。

2.2 实体消歧

实体消歧，又被称为实体链接，用于解决数据多源造成的同名异义问题。目前医学上的实体消歧方法主要是将实体向量化后通过机器学习相关技术进行消歧。

Wu 等[54]开发了临床缩写识别和消歧框架，以在临床叙述上提高自然语言系统的缩写处理能力。该框架通过词义清单和基于配置文件的词义消歧(Word Sense Disambiguation，WSD)方法[55]实现了缩写的词义消歧，取得了不错的效果。其不足之处在于词义清单的构建需要人工参与，不适用于大规模实体消歧任务。Zhu 等[56]在生物医学的实体链接上首先应用潜在类型建模思想，提出了基于神经网络的实体链接模型 LATTE。模型进一步提高了实体链接性能，表明了引入潜在类型进行建模的有用性。Mondal 等[57]采用三元组网络[58](Triplet Network)来排名知识库中的候选实体，并给出了一种不需要人为制定规则的候选实体生成方案，推动了医学实体链接的发展。此外，还有 Xu 等[59]提出的 unMERL 和 Abdurxit 等[60]提出的基于实体间和实体内注意力的实体消歧方法等。目前的实体消歧方法，通过引入外部知识和机器学习技术取得了良好的实体消歧效果，但针对复杂的情形仍存在问题，且外部知识的引入也带来了局限性。相关的知识融合方法具体如表 4 所列。

表 4 知识融合方法

Table 4 Knowledge fusion methods

知识融合方法	方法/模型	优点	不足	实验数据
实体对齐	ASMOV[48]	将本体多方面的信息用来匹配，并使用语义验证来提升匹配效果	算法相对复杂	OAEI 2008[1)]
	LogMap[49]	可适用于大规模的本体匹配	过于依赖从本体中提取的词汇特征和属性三元组	NCI[2)]等
	SiBERT[50]	具有好的实体表征；对齐速度快	适用范围有限；对英文的向量嵌入存在不足	HIT-CIR Tongyici Cilin (Extended)[3)]等
	BioHAN[51]	具有好的本体匹配效果	仅考虑本体中的子类关系，限制了图表示学习能力；匹配效率有待改进	OAEI 2021[4)]
	MEDTO[52]	对层次和非层次信息的捕获，进一步提高了本体匹配效果	信息融合策略可进一步提高	MIMIC-III[61-62]等
实体消歧	CATD 框架[54]	可帮助提高现有自然语言处理系统在临床缩写识别和消歧上的性能	词义清单的构建需人工参与	合成衍生物数据库[63]
	LATTE[56]	对潜在类型的建模提高了实体链接效果	向量嵌入方法可进一步优化	MedMentions[64]等
	基于三元组网络的方法[57]	避免了人为规则设计	在基于注意力的疾病相似性计算上仍有待提高	NCBI 疾病语料库[65]
	unMERL[59]	无监督的模型避免了人工参与，具有良好的可扩展性；上下文信息的提取和语义相关性的使用，降低了上下文噪声和信息缺乏对链接效果的影响	实体识别方法有待改进	互动百科[5)]
	文献[60]	实体间和实体内注意力的集成，更好地捕获了实体和其链接实体间的信息；推理速度快	不适用于一对多的链接预测	ADR-TAC 2017[66]、NCBI 疾病语料库

1) http://oaei.ontologymatching.org/2008/

2) https://www.cancer.gov/about-nci

3) http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

4) http://oaei.ontologymatching.org/2021/

5) https://www.baike.com/sitecategory-10.html

2.3 知识推理

知识推理是在初步构建完成的知识图谱上，进一步通过逻辑推理来得到新的知识。医学上，知识推理能帮助进行辅助诊断[67]、挖掘医学信息等。医学知识图谱上的推理方法可分为基于规则的推理方法和基于表示学习的推理方法。

(1)基于规则的推理方法。Bousquet 等[68]构建了数据挖掘工具 PharmaMiner，根据 DAML＋OIL 描述逻辑来编写 MedDRA 代码进行推理，改进了药物警戒上的信号检测。该方法在结果上产生了更多的关联，但这些关联不能被具体确定，仅能增加查询的响应数量。Chen 等[69]提出了结合案例推理与规则推理的方法，通过在案例推理上自定义推理规则，来对应急响应事件知识图谱中的隐含知识进行更好的挖掘。该方法能为突发事件相关人员提供辅助决策，但不具有时序特征。基于规则的逻辑推理，优点在于模型设计简单、高效，在小规模知识图谱上能取得较好的推理结果。但由于规则需要人工构建，不适用于动态变化的知识图谱，迁移性不强。

(2)基于表示学习的推理方法。基于表示学习的医学推理是将知识图谱中的实体、关系等映射到向量空间中进行推理的方法。Chen 等[70]在临床领域提出了基于 LSTM 序列增量学习的时序知识图谱链接预测模型，以端到端的方式提取时序特征，从而实现链接预测。该模型引入了时序信息，较传统链接预测模型提高了准确度，但缺乏对图结构信息的利用。Lan 等[71]提出基于路径的推理方法，以解决知识图谱推理上存在的实体和路径稀疏问题。该方法采用预训练 BERT 模型编码实体和路径的文本语义信息，再结合注意力机制的使用，提高了推理性能。Biswas 等[72]将基于 ComplEx[73]嵌入的张量分解方法用于合并症预测，其中 ComplEx 嵌入方法能对非对称和对称关系进行向量嵌入，但不具有传递性。Gao 等[74]设计了 PTMKG-WE 框架，采用 Word2vec[75]进行词嵌入，并给出聚类和均值两种方法来改进关系的特征嵌入，可用来预测医学知识图谱中的三元组。基于表示学习的医学推理能表征医学知识图谱上的多样化关系，从而取得较好的推理效果，但对多跳关系处理和在稀疏知识图谱上的推理上仍存在不足。知识推理方法具体如表 5 所列。

表 5 知识推理方法

Table 5 Knowledge reasoning methods

类别	方法/模型	优点	不足	实验数据	应用
基于规则的推理	PharmaMiner[68]	改进了信号检测算法	能产生更多关联，但不能确定是信号还是噪声	FPVD[1)]	药物警戒中的信号检测
	基于案例和规则的推理[69]	具有可解释性；挖掘了图谱中的隐含知识，推理出了实际需要的资源调度方案	缺乏对时序特征的考虑；推理准确度待提高	—	医疗应急响应
基于表示学习的推理	基于 LSTM 序列增量学习[70]	注重时序特征，提高了推理准确度	缺乏对图结构信息的利用	电子病历	临床领域时序知识图谱上的链接预测
	基于路径的知识推理方法[71]	缓解了推理上的实体和关系稀疏问题	语义特征表示方法有待改进	CSKG	医学知识图谱补全
	基于 ComplEx 嵌入的张量分解方法[72]	推理准确度高	对复杂关系的推理存在不足	HPO[76]等	合并症预测
	PTMKG-WE[74]	改进了关系特征表示	预测准确度有待提高	NDF-RTD[2)]等	医学知识图谱上的三元组预测

为帮助医学知识图谱的研究者更好地开展相关研究，我们整理了部分收集到的医学数据集，具体如表 6 所列。

表 6 公共医学数据集

Table 6 Common datasets for medicine

类型	名称	描述
文本数据集	CMeEE[77]	中文医学实体抽取数据集，数据来源于医学教科书和临床实践等，包含 504 种常见的儿科疾病、7085 个身体部位、12907 种临床症状和 4354 个医疗程序等九大类医学实体
	CMeIE[78]	中文医学信息抽取数据集，数据来源于医学教科书和临床实践等，包含近 7.5 万三元组数据、2.8 万疾病语句和 53 种关系类型
	Yidu-S4K[3)]	医渡云结构化 4K 数据集，由医渡云医学人工根据真实病历编辑得到，可用于医疗命名实体识别和医疗实体及属性抽取
	TCMID[79]	中医药综合数据库，收集了各种来源的中医药相关信息，包含了 47000 个处方、8159 种药材、25210 个化合物、6828 种药物、3791 种疾病和 17521 个相关目标
	MIMIC-III	可免费使用的大型数据库，包括 2001—2012 年间贝斯以色列女执事医疗中心重症监护室住院的 4 万多名患者的相关健康数据，主要是人口统计学、床边的生命体征测量等信息
	DrugBank[80]	生物信息学和化学信息学数据库，目前最新版本(5.1.10)包含 15324 种药物条目，其中有 2734 种小分子药物、1572 种生物技术药物、134 中营养品和 6716 种实验药物

1) https://sante.gouv.fr/soins-et-maladies/medicaments/

2) https://bioportal.bioontology.org/ontologies/NDF-RT

3) http://openkg.cn/dataset/yidu-s4k

（续表）

类型	名称	描述
图像数据集	MedPix[1)]	在线医学图像、教学案例和临床主题数据库，集成了图像和文本元数据，包括12000多个病例场景、9000个主题和近59000张图像
	VIA Group Public Databases[2)]	VIA集团公共数据库，包含ECLAP全肺CT图像和处理药物反应两个公共数据库，内容是DICOM格式的肺部CT图像和放射科医生的异常记录
	LIDC-IDRI[3)]	LIDC-IDRI肺结节数据库，由胸部医学图像文件和对应诊断结果的病变标注组成，共收录了1018个研究实例
	DRIVE[4)]	DRIVE数据库，来自荷兰的糖尿病视网膜病变筛查项目，筛查人群包括400名年龄在25～90岁的糖尿病患者
	EchoNet-Dynamic[81]	EchoNet-Dynamic数据库，包括10030个标记的超声心动图视频和相应专家所做的注释
语音数据集	TORGO[82]	TORGO数据库，数据来自患有脑瘫（CP）或肌萎缩侧索硬化症（ALS）的人群，包括对齐的声学数据和测量的3D发音特征
	VoxCeleb[83-85]	人类语音视听数据集，数据来源于在YouTube的采访视频中提取的人类语言片段，共计有超过2000 h的音频和视频
	MedDialog CN[86]	MedDialog中文数据库，来自医生和病人之间的对话，包含110万个对话和400万个话语

3 医学知识图谱应用案例

近年来，随着医疗卫生信息化的迅速普及与B2C[87]医疗模式的推进，产生了大量的医疗数据。知识图谱作为一种从海量数据中抽取结构化知识的手段，为医疗系统中动态、海量、异构的数据处理提供了一种有效的方式。通过大数据和知识图谱实现医学数据聚合，进一步构建大数据＋互联网的智慧医疗系统来实现辅助医疗，是未来医疗的发展趋势。下面介绍几个成功的医学知识图谱应用案例。

3.1 COVID-19知识图谱

Yang等[88]通过融合COVID-19科学文献知识图谱、西药治疗知识图谱和中药治疗知识图谱，生成了COVID-19知识图谱。为保证知识图谱的质量，采用人工抽样的方式进行检查。构建出的知识图谱可帮助医生进行快速的诊断和治疗，也为相关研究提供了数据支撑。但半自动的构建方式不易于拓展，同时图谱本身也存在体量和细粒度偏小等问题。为此，可通过引入无监督、主动学习等方法帮助进行知识图谱的构建，以实现自动化构建，并减少对人力资源的利用。

3.2 中医临床知识图谱

Yu等[89]为解决中医临床知识在不同的组织结构和信息系统中出现的“数据孤岛”问题，构建了中医临床知识图谱，主要分为3个步骤：顶层本体设计、语义网络构建和知识图谱内容填充。中医临床知识图谱一定程度上解决了“数据孤岛”问题，有力地展现了整个中医临床领域的知识框架，可帮助用户进行辅助决策、掌握中医药的复杂知识体系。但整个知识图谱过于重视内容的填充，忽略了对冗余信息的处理以及图谱上的知识推理，这会导致知识图谱中出现知识冗余和存在过多的隐含信息。为此，可引入知识融合技术去冗余，引入知识推理和数据挖掘技术进行数据挖掘。

3.3 乙肝知识图谱

Yin等[90]使用收集到的乙型肝炎电子病历和寻医问药网站上爬取的乙肝医学补充知识构建了乙肝知识图谱，用于乙肝智能问答系统。对于患者的提问，先采用Bi-LSTM＋CRF模型进行实体识别，通过模板匹配来在知识图谱中进行检索查询，以实现智能问答。基于知识图谱的智能问答系统能更好地实现检索查询，为医生就诊提供帮助，提高了就诊速度。但整个知识图谱存在体量过小和问题模板覆盖面不全等问题。针对问题模板覆盖面不全的问题，可尝试使用基于概率的级联检索方法来进行问题答案确定，从而避免设计模板。

3.4 生物学见解知识图谱（BIKG）

Geleta等[91]构建了生物学见解知识图谱（Biological Insights Knowledge Graph，BIKG），图谱包含了来自Hetionet[92]，Opentargets[93]等数据源上的数据。在图谱的构建过程中，命名实体识别和关系抽取分别使用了Termite Tagger和LINK软件。BIKG集成了内部和外部的相关数据，可应用机器学习算法获得新的见解，同时在组织层面上能缩短数据准备时间、提高质量等。但目前提出的BIKG适应用例有限，且仅供阿斯利康内部使用。在未来的工作中，Geleta等也将进一步使BIKG能适应更多新用例，同时提高其对图神经网络、可解释的人工智能等机器学习技术的适用性。

3.5 BIOS生物医学知识图谱

Yu等[94]开发了一个大型的开放生物医学知识图谱BIOS（Biomedical Informatics Ontology System），这是利用文本挖掘、深度学习等技术构建出的超大规模医学知识图谱，其术语发现、语义分析等构建过程都是由模型自动进行实现。目前最新版本的BIOS（2022V2.2）达到了2693万概念数量和5415万术语数量，并开放下载。BIOS相比UMLS能更好地为国内医学大数据分析提供平台基础，可应用在生物医学领域的自然语言处理、基于医学知识图谱的智能问答等。

结束语 本文对医学知识图谱的完整构建过程进行了综述，对知识抽取、知识融合和知识推理的具体方法进行了较为详细的陈述。为帮助研究者更好地开展相关研究，列出了部分医学数据集，并介绍了几个知识图谱应用案例。

在现有的医学知识图谱技术背景下，可深入研究的方向很多，结合笔者调研，我们给出以下3个医学知识图谱未来可深入研究或应用的方向。

1) https://medpix.nlm.nih.gov/home

2) http://www.via.cornell.edu/databases/

3) https://www.cancerimagingarchive.net/

4) https://drive.grandchallenge.org/

(1)注重医学知识图谱构建过程中各技术间的联合训练。现有知识图谱构建过程中,已出现利用实体关系联合抽取的方法来同时提高实体、关系抽取的准确度。由于医学知识的逻辑性强,故知识推理的准确度会更依赖实体和关系的抽取结果。因此,可尝试将构建过程进行关联,即将实体、关系抽取与知识推理相联合,以提高实体、关系抽取的准确度。

(2)尝试弱监督、无监督的医学知识图谱构建。现有医学知识图谱构建技术大多需要人工事先标注部分数据,这带来了不确定性,也不利于知识图谱的扩大。故可在知识图谱的构建中进一步引入弱监督、无监督的方法,以解决人工标注数据所带来的问题。

(3)多领域医学知识图谱融合。现有的医学知识图谱较多的是针对特定医学领域,缺少包含多病种、组织的知识图谱。但由于特定医学领域的专业性强,故直接构建多领域医学知识图谱存在诸多困难。为此可融合多个特定领域医学知识图谱,间接实现多领域医学知识图谱构建。这可在现有融合技术上做进一步研究。

参 考 文 献

[1] SALAM A, SCHWITTER R, ORGUN M A. Probabilistic rule learning systems: a survey [J]. ACM Computing Surveys, 2021, 54(4): 79: 1-16.

[2] RAEDT L D, THON I. Probabilistic Rule Learning[C]// International Conference on Inductive Logic Programming. 2010: 47-58.

[3] LIU Z Y, SUN M S, LIN Y K, et al. Knowledge representation learning: a review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.

[4] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[5] DENG L, YU D. Deep learning: methods and applications [J]. Foundations and Trends in Signal Processing, 2013, 7(3/4): 197-387.

[6] ZHAO T, CHENG L, ZANG T, et al. Peptide-major histocompatibility complex class i binding prediction based on deep learning with novel feature [J]. Front Genet, 2019, 10: 1191-1198.

[7] ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[8] REDVERS N, BLONDIN B. Traditional indigenous medicine in north america: a scoping review [J/OL]. PLoS One, 2020, 15(8). https://doi.org/10.1371/journal.pone.0237531.

[9] SUN W, CAI Z, LI Y, et al. Data processing and text mining technologies on electronic medical records: a review [J]. Journal of Healthcare Engineering, 2018, 2018(5): 1-9.

[10] SINGHAL A. Introducing the Knowledge Graph: Things, Not Strings[EB/OL]. (2012-05-16) [2022-05-02]. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

[11] HOU M W, WEI R, LU L, et al. Research review of knowledge graph and its application in medical domain [J]. Journal of Computer Research and Development, 2018, 55(12): 2587-2599.

[12] CODEN A, SAVOVA G, SOMINSKY I, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model [J]. Journal of Biomedical Informatics, 2009, 42(5): 937-949.

[13] 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, 2010, 17(5): 507-513.

[14] ZHOU G, SU J. Named Entity Recognition Using an HMM-Based Chunk Tagger[C]// Annual Meeting on Association for Computational Linguistics. 2002: 473-480.

[15] MCCALLUM A, FREITAG D, PEREIRA F. Maximum Entropy Markov Models for Information Extraction and Segmentation [C]// International Conference on Machine Learning. 2000: 591-598.

[16] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// International Conference on Machine Learning. 2001: 282-289.

[17] ZHANG J, SHEN D, ZHOU G, et al. Enhancing hmm-based biomedical named entity recognition by studying special phenomena[J]. Journal of Biomedical Informatics, 2004, 37(6): 411-422.

[18] WANG J, PENG Y, LIU B, et al. Extracting Clinical Entities and Their Assertions from Chinese Electronic Medical Records Based on Machine Learning[C]// International Conference on Materials Engineering, Manufacturing Technology and Control. 2016: 1503-1508.

[19] JONNALAGADDA S, COHEN T, WU S, et al. Enhancing clinical concept extraction with distributional semantics [J]. Journal of Biomedical Informatics, 2012, 45(1): 129-140.

[20] LIU J W, SONG Z Y. Overview of recurrent neural networks [J]. Control and Decision, 2022, 37(11): 2753-2768.

[21] SUN X, MAN Y. Enhance Chinese Medical Name Entity Recognition with Etymon Features[C]// International Conference on Computer, Communication and Network Technology. 2018: 490-494.

[22] LI L S, GUO Y K. Biomedical named entity recognition with cnn-blstm-crf [J]. Journal of Chinese Information Processing, 2018, 32(1): 116-122.

[23] LI L Q, ZHAO J, HOU L, et al. An attention-based deep learning model for clinical named entity recognition of chinese electronic medical records [J/OL]. BMC Medical Informatics and Decision Making, 2019, 19(Suppl 5). https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0933-6.

[24] JI B, LIU R, LI S, et al. A hybrid approach for named entity recognition in chinese electronic medical record [J/OL]. BMC Me-

dical Informatics and Decision Making, 2019, 19(Suppl 2). https://doi.org/10.1186/s12911-019-0767-2.

[25] JI B, LI S, YU J, et al. Research on chinese medical named entity recognition based on collaborative cooperation of multiple neural network models [J/OL]. Journal of Biomedical Informatics, 2020, 104. https://doi.org/10.1016/j.jbi.2020.103395.

[26] MARCUS M P, MARCINKIEWICZ M A, SANTORINI B. Building a large annotated corpus of English: the penn treebank [J]. Computational Linguistics, 1993, 19(2): 313-330.

[27] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name [J]. Machine Learning, 1999, 34: 211-231.

[28] UZUNERÖ, SOUTH B R, SHEN S, et al. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

[29] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[30] SUN Z Y, E H H, SONG M N, et al. The method of medical knowledge graphs construction based on big data technology [J]. Computer Engineering & Software, 2020, 41(1): 13-17.

[31] ZHU L, ZHU Y, YANG F. Knowledge extraction research for semantic expression of diseases inchinese medicine [J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2016, 18(8): 1241-1250.

[32] EL-HALEES A, ELHAJ M. Extracting Information from Medical Reports[C]// Palestinian International Conference on Information and Communication Technology. 2021: 1-5.

[33] NIKFARJAM A, EMADZADEH E, GONZALEZ G. Towards generating a patient's timeline: extracting temporal relationships from clinical notes [J]. Journal of Biomedical Informatics, 2013, 46(Supplement): S40-S47.

[34] ZHAO X, LIN S, HUANG Z. Extraction of Semantic Relations from Medical Literature Based on Semantic Predicates and SVM [C]// International Conference on Health Information Science. 2018: 17-24.

[35] ROBERTS A, GAIZAUSKAS R, HEPPLE M, et al. Mining clinical relationships from patient narratives [J/OL]. BMC Bioinformatics, 2008, 9(Suppl 11). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S11-S3.

[36] ZHENG S, WANG F, BAO H, et al. Joint Extraction of Entities and Relations Based on A Novel Tagging Scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1227-1236.

[37] WU Y, ZHU X, ZHU Y. An improved approach to the construction of chinese medical knowledge graph based on ctd-blstm model [J]. IEEE Access, 2021, 9: 74969-74976.

[38] GAO F, YANG J X, GU J G. Extraction of diagnosis and treatment relationship based on fusion relation discovery words and deep learning [J]. Computer Applications and Software, 2021, 38(12): 168-173.

[39] SUN W, RUMSHISKY A, UZUNER O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge [J]. Journal of the American Medical Informatics Associatio, 2013, 20(5): 806-813.

[40] RIEDEL S, YAO L, MCCALLUM A. Modeling Relations and Their Mentions Without Labeled Text [C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2010: 148-163.

[41] KERSLOOT M G, LAU F, ABU-HANNA A, et al. Automated snomed ct concept and attribute relationship detection through a web-based implementation of ctakes [J/OL]. Journal of Biomedical Semantics, 2019, 10(1). https://doi.org/10.1186/s13326-019-0207-3.

[42] MYKOWIECKA A, MARCINIAK M, KUPŚĆ A. Rule-based information extraction from patients' clinical data [J]. Journal of Biomedical Informatics, 2009, 42(5): 923-936.

[43] JIANG H J. Slot filling via deep learning[D]. Hangzhou: Zhejiang University, 2017.

[44] SHI X, YI Y, XIONG Y, et al. Extracting entities with attributes in clinical text via joint deep learning [J]. Journal of the American Medical Informatics Association, 2019, 26(12): 1584-1591.

[45] XU J, LI Z, WEI Q, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text [J/OL]. BMC Medical Informatics and Decision Making, 2019, 19(Suppl 5). https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0937-2.

[46] DU M, WANG W, WANG S, et al. A Unified Framework for Attribute Extraction in Electronic Medical Records[C]// 2020 3rd International Conference on Algorithms, Computing and Artificial. 2021: 1-7.

[47] ELHADAD N, PRADHAN S, GORMAN S L, et al. SemEval-2015 Task 14: Analysis of Clinical Text[C]// Association for Computational Linguistics. 2015: 303-310.

[48] JEAN-MARY Y R, SHIRONOSHITA E P, KABUKA M R. Ontology matching with semantic verification [J]. Journal of Web Semantics, 2009, 7(3): 235-251.

[49] JIMÉNEZ-RUIZ E, GRAU B C. LogMap: Logic-Based and Scalable Ontology Matching[C]// International Semantic Web Conference. 2011: 273-288.

[50] MA Z, ZHAO L, LI J, et al. SiBERT: a siamese-based bert network for chinese medical entities alignment [J]. Methods, 2022, 205: 133-139.

[51] WANG P, HU Y. Matching biomedical ontologies via a hybrid graph attention network [J/OL]. Frontiers in Genetics, 2022, 13. https://www.frontiersin.org/articles/10.3389/fgene.2022.893409.

[52] HAO J, LEI C, EFTHYMIOU V, et al. MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks [C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2946-2954.

[53] CHAMI I, YING R, RÉ C, et al. Hyperbolic Graph Convolutional Neural Networks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 4868-4879.

[54] WU Y, DENNY J C, ROSENBLOOM S T, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (card) [J]. Journal of American Medical Informatics Association, 2017, 24(e1): e79-e86.

[55] XU H, STETSON P D, FRIEDMAN C. Combining Corpus-Derived Sense Profiles with Estimated Frequency Information to Disambiguate Clinical Abbreviations[C]// AMIA Annual Symposium Proceedings. 2012: 1004-1013.

[56] ZHU M, CELIKKAYA B, BHATIA P, et al. LATTE: Latent Type Modeling for Biomedical Entity Linking [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9757-9764.

[57] MONDAL I, PURKAYASTHA S, SARKAR S, et al. Medical Entity Linking Using Triplet Network[C]// Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019: 95-100.

[58] HOFFER E, AILON N. Deep Metric Learning Using Triplet Network[C]// International Workshop on Similarity-Based Pattern Recognition. 2015: 84-92.

[59] XU J, GAN L, CHENG M, et al. Unsupervised medical entity recognition and linking in chinese online medical text [J/OL]. Journal of Healthcare Engineering, 2018. https://doi. org/10. 1155/2018/2548537.

[60] ABDURXIT M, TOHTI T, HAMDULLA A. An efficient method for biomedical entity linking based on inter-and intra-entity attention [J/OL]. Applied Sciences, 2022, 12(6). https://doi. org/10. 3390/app12063191.

[61] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III, a freely accessible critical care database [J]. Scientific Data, 2016, 3(1): 1-9.

[62] GOLDBERGER A L, AMARAL L A N, GLASS L, et al. PhysioBank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals [J]. Circulation, 2000, 101(23): e215-e220.

[63] RODEN D M, PULLEY J M, BASFORD M A, et al. Development of a large-scale de-identified dna biobank to enable personalized medicine [J]. Clinical Pharmacology & Therapeutics, 2008, 84(3): 362-369.

[64] MOHAN S, LI D. Medmentions: a large biomedical corpus annotated with umls concepts [J]. arXiv: 1902, 09476, 2019.

[65] DOGAN R I, LEAMAN R, LU Z. NCBI disease corpus: a resource for disease name recognition and concept normalization [J]. Journal of Biomedical Informatics, 2014, 47: 1-10.

[66] ROBERTS K, DEMNER-FUSHMAN D. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track [C/OL]// Text Analysis Conference. 2017. https://tac. nist. gov/publications/2017/additional. papers/TAC2017. KBP_Event_Nugget_overview. proceedings. pdf.

[67] MOHAMMADHASSANZADEH H, WOENSEL W V, ABIDI S R, et al. Semantics-based plausible reasoning to extend the knowledge coverage of medical knowledge bases for improved clinical decision support [J]. BioData Mining, 2017, 10(1): 1-31.

[68] BOUSQUET C, HENEGAR C, LOUËT A L, et al. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach [J]. International Journal of Medical Informatics, 2005, 74(7/8): 563-571.

[69] CHEN Y X, YANG C C, GE T Y, et al. Research on medical emergency response mechanism based on knowledge reasoning [J]. Journal of Chinese Computer Systems, 2022, 43(3): 638-643.

[70] CHEN D H, YIN S N, LE J J, et al. A link prediction model for clinical temporal knowledge graph [J]. Journal of Computer Research and Development, 2017, 54(12): 2687-2697.

[71] LAN Y, HE S, LIU K, et al. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion [J/OL]. BMC Medical Informatics and Decision Making, 2021, 21(Suppl 9). https://bmcmedinformdecismak. biomedcentral. com/articles/10. 1186/s12911-021-01622-7.

[72] BISWAS S, MITRA P, RAO K S. Relation prediction of co-morbid diseases using knowledge graph completion [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 18(2): 708-717.

[73] TROUILLON T, WELBL J, RIEDEL S, et al. Complex Embeddings for Simple Link Prediction[C]// International Conference on Machine Learning. 2016: 2071-2080.

[74] GAO M, LU G, CHEN F. Medical knowledge graph completion based on word embeddings [J/OL]. Information, 2022, 13(4). https://doi. org/10. 3390/info13040205.

[75] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. arXiv: 1301. 3781, 2013.

[76] KÖHLER S, DOELKEN S C, MUNGALL C J, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data [J]. Nucleic Acids Research, 2014, 42(D1): D966-D974.

[77] ZAN H, LI W, ZHANG K, et al. Building a Pediatric Medical Corpus: Word Segmentation and Named Entity Annotation [C]// Chinese Lexical Semantics Workshop. 2020: 652-664.

[78] GUAN T, ZAN H, ZHOU X, et al. CMeIE: Construction and Evaluation of Chinese Medical Information Extraction Dataset [C]// Natural Language Processing and Chinese Computing. 2020: 270-282.

[79] XUE R, FANG Z, ZHANG M, et al. TCMID: traditional chinese medicine integrative database for herb molecular mechanism analysis[J]. Nucleic Acids Research, 2013, 41(Database Issue): 1089-1095.

[80] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5. 0: a major update to the drugbank database for 2018 [J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.

[81] OUYANG D，HE B，GHORBANI A，et al. Video-based ai for beat-to-beat assessment of cardiac function [J]. Nature，2020，580(7802)：252-256.

[82] RUDZICZ F，HIRST G，LIESHOUT P V. Vocal tract representation in the recognition of cerebral palsied speech [J]. Journal of Speech Language and Hearing Research，2012，55(4)：1190-1207.

[83] NAGRANI A，CHUNG J S，XIE W，et al. Voxceleb：large-scale speaker verification in the wild [J/OL]. Computer Speech & Language，2020，60. https：//doi. org/10. 1016/j. csl. 2019. 101027.

[84] CHUNG J S，NAGRANI A，ZISSERMAN A. VoxCeleb2：Deep Speaker Recognition[C]// International Speech Communication Association. 2018：1086-1090.

[85] NAGRANI A，CHUNG J S，ZISSERMAN A. VoxCeleb：a Large-Scale Speaker Identification Dataset[C]// International Speech Communication Association. 2017：2616-2620.

[86] CHEN S，JU Z Q，DONG X Y，et al. MedDialog：a large-scale medical dialogue dataset [J]. arXiv：2004. 03329，2020.

[87] WANG P，WU H. Discussion on the current situation and future development trend of mobile Internet medical applications at home and abroad [J]. China Digital Medicine，2014，9(1)：8-10.

[88] YANG S，WANG X H，ZHAO Z G，et al. Research on the construction and application of covid-19 knowledgegraph [J]. Journal of Qingdao University(Engineering & Technology Edition)，2021，36(4)：22-29.

[89] YU T，LI J H，ZHU L，et al. Construction and application of clinical knowledge map of traditional Chinese medicine [J]. New Era of Science and Technology，2017(4)：51-54.

[90] YIN Y T，ZHANG L，WANG Y G，et al. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis b [J/OL]. BioMed Research International，2022. https：//doi. org/10. 1155/2022/7139904.

[91] GELETA D，NIKOLOV A，EDWARDS G，et al. Biological insights knowledge graph：an integrated knowledge graph to support drug development [J/OL]. bioRxiv 2021. 10. 28. 466262. https：//doi. org/10. 1101/2021. 10. 28. 466262.

[92] HIMMELSTEIN D S，LIZEE A，HESSLER C，et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing [J/OL]. eLife，2017，6. https：//elifesciences. org/articles/26726.

[93] KOSCIELNY G，AN P，CARVALHO-SILVA D，et al. Open Targets：a platform for therapeutic target identification and validation [J]. Nucleic Acids Research，2017，45(D1)：985-994.

[94] YU S，YUAN Z，XIA J，et al. BIOS：an algorithmically generated biomedical knowledge graph [J]. arXiv：2203. 09975，2022.

JIANG Chuanyu，born in 1999，postgraduate，is a member of China Computer Federation. His main research interest is knowledge graph.

GU Yang，born in 1975，master. His main research interests include image recognition and data analysis.

（责任编辑：何杨）