知识图谱在生物医学领域的研究综述

学号：S191000854　　姓名：潘小琴

摘要：

随着医疗大数据时代的到来，知识互联受到了广泛的关注。如何从海量的数据中提取有用的医学知识，是医疗大数据分析的关键。知识图谱技术提供了一种从海量文本和图像中抽取结构化知识的手段，知识图谱与大数据技术、深度学习技术相结合，正在成为推动人工智能发展的核心驱动力。知识图谱技术在医疗领域拥有广阔的应用前景，该技术在医疗领域的应用研究将会在解决优质医疗资源供给不足和医疗服务需求持续增加的矛盾中产生重要的作用。目前，针对医学知识图谱的研究还处于探索阶段，现有知识图谱技术在医疗领域普遍存在效率低、限制多、拓展性差等问题。首先针对医疗领域大数据专业性强、结构复杂等特点，对医学知识图谱架构和构建技术进行了剖析；其次，分别针对医学知识图谱中知识表示、知识抽取、知识融合这三个模块的关键技术和研究进展进行综述，并对这些技术进行实验分析与比较。此外，介绍了医学知识图谱在临床决策支持、医疗问答等医疗服务中的应用现状。最后对其发展前景进行了展望。

1.背景知识概述：

知识图谱（knowledge Graph）是以图的形式表现客观世界中的实体（概念、人、事物）及其之间关系的知识库。知识图谱的概念。知识图谱的概念于 2015 年 5 月被 google 正式提出[1]，其原始目的是为了提高搜索引擎的能力，提高搜索结果质量并提升用户的搜索体验。2013 年之后，随着智能信息服务和应用的不断发展，知识图谱已在学术界和工业界普及，并在智能搜索、智慧问答大数据风控、推荐系统等应用中发挥着重要的作用。目前，医学是知识图谱应用最广的垂直领域之一，也是目前国内外人工智能领域研究的热点，在如疾病风险评估、智能辅助诊疗、医疗质量控制及医疗知识问答等智慧医疗领域都有着很好的发展前景[2]。目前很多公司均构建了自己的知识图谱，如 IBM 的 Watson Health、阿里健康的“医知鹿”医学智库、斯坦福大学的 GNBR 数据集[3]等。

随着区域卫生信息化及医疗信息技术的发展，积累了海量的医学数据，如何从这些数据中提炼信息并加以应用，是推进智慧医疗辅助的关键[4]，也是医学知识检索、辅助诊疗、医疗质量控制、电子病历及健康智能化管理应用的基础，对于提高医生诊疗水平、减轻医生负担具有非常重要的意义。

1.1 知识图谱定义

知识图谱是语义网（Sematic web）技术之一，是一种基于图的数据结构，由节点（实体）和标注的边（实体间的关系）组成[5]，它本质上是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述[6]。知识图谱通常用三元组的形式来表示，，由两个具有语义连接关系的医疗实体和实体间的关系组成，是医学知识的直观表示，G =(head,relation,tail)，其中 head 和 tail 分别是头实体和尾实体；relation = {r1,r2,r3・・・}是知识库中的关系集合，三元组中主要包含实体、关系、概念、属性和属性值等。

在生物医学数据中，实体通常指疾病、药物、症状等；关系存在于不同实体之间，例如临床表现、药理作用、发病机制等；概念主要指对象类别，事务的类别等如一级手术、慢性病等；属性主要包含疾病特征，药物规格等；属性值指对象特定的属性值。实体实体之间的内在特征通过属性-属性值来刻画，实体之间的关联通过关系来描述。最终将所有的三元组共同构成一个异构网络，其中节点表示实体，有向边表示关系，不同的关系通过不同类型的边来表示，即标记不同标签。如图 1 所示。

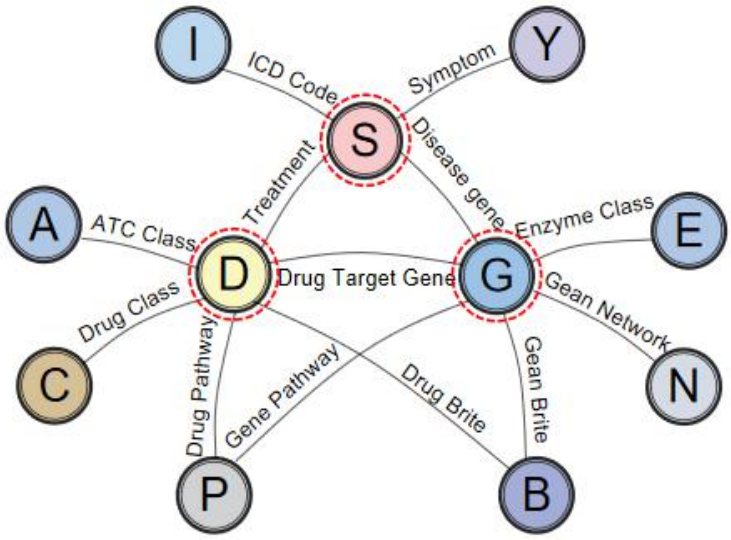


图 1：医学知识图谱

1.2 构建医学知识图谱

医学知识图谱的构建流程可以被归纳为 3 个模块，即医学知识抽取、医学知识融合以及医学知识计算。医学知识抽取通过从大量结构化、半结构化或非结构化的医学数据中提取出实体、关系、属性等知识图谱的组成元素，并选择合理高效的方式将元素存入知识库中；医学知识融合对医学知识库的内容进行整合、消歧、加工，增强知识库内部的逻辑性和表达能力，并为医学知识图谱更新旧知识或补充新知识；医学知识计算借助知识推理，推断出缺失事实，自动完成疾病诊断与治疗。

构建医学知识图谱的主要目的是抽取大量的、让计算机可读的医学知识。生物医学界对化学物质，基因和表型如何相互作用的集体理解分布在 2400 万篇研究文章的全文中。这些

相互作用提供了对更高阶生化现象背后机制的见解，例如药物与药物之间的相互作用以及个体间药物反应的变化。为了帮助他们大规模地进行策划，我们必须了解可能的关系类型，并将非结构化的自然语言描述映射到这些结构化的类上。[7]中提出了一种从两千多篇医学文献摘要中提取医学知识图谱的方法。使用 NCBI 的 PubTator 批注，在 Medline 摘要中标识化学，基因和疾病名称的实例，并应用了 Stanford 依赖性分析器来查找单个句子中实体对之间的连接依赖性路径。他们将已发布的集成二类聚类算法（EBC）与分层聚类相结合，并将依赖关系路径分为与语义相关的类别，使用标签或“主题”（例如，“禁止”和“激活”）进行了注释如图 2 所示。最终根据六个人为管理的数据库（DrugBank，Reactome，SIDER，治疗目标数据库，OMIM 和 PharmGKB）评估了主题任务。

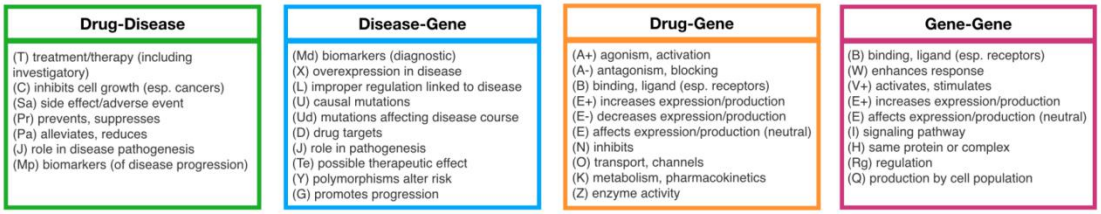


图 2：GNBR 中的主题总结

2.国内外研究现状及创新点

理解、推理和归纳能力是人类智力的核心[8]。然而对于机器而言，想要理解和推理出两个实体之间的关系具有很大的挑战。现实世界中的医疗关系实体具有非常复杂的属性，因此，医学知识图谱的构建与应用需要多种智能信息处理技术的支持[9]。通过知识抽取技术，可以从半结构化、非结构化数据中提取知识要素。借助知识融合技术，可以消除实体、关系、属性与对象之间的歧义，形成高质量医学知识库。医学知识计算是在已有知识的基础上进一步挖掘隐含知识，从而丰富、扩展医学知识库。本节将从医学知识表示、医学知识抽取、医学知识融合所运用的关键技术为重点，详细说明其中的相关研究。

2.1 医学知识表示

通过知识图谱中三元组的的形式来表示知识已经受到广泛的使用及认可，但当应用在医学领域时却会出现计算效率低下的问题。随着近年来人工智能、机器学习、深度学习等技术的发展，知识图谱的表示学习取得了突破性进展。将医学实体中的语义信息映射为稠密低维的实数向量，从而可以在低维空间中计算实体和关系的复杂语义关联，这对于医学知识库的构建过程有重要意义。医学知识表示早期的计算方式主要是基于距离的平移模型，利用基于距离的评分函数对事实的合理性进行评判，代表模型有翻译模型 TransE 以及基于其延伸出的复杂关系模型 TransH，TorusE 等。随着深度学习的发展，基于知识图谱嵌入的表示学习模型也有了快速发展，极具代表的有 convE，RUGE 等。

1）平移模型

TransE[10]是最具代表性的距离平移模型，它将实体和关系表示为同一空间的矢量如图4所示，三元组中的关系矢量$l_{relation}$可以被看作头实体矢量l_{head}到尾实体矢量l_{tail}的翻译，并满足关系：

$$l_{head} + l_{relation} = l_{tail}$$

评价函数为

$$f_{relation}(head, tail) = \left| l_{head} + l_{relation} - l_{tail} \right|_{L_1 / L_2}$$

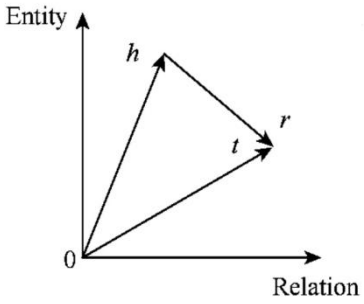


图3：TransE模型

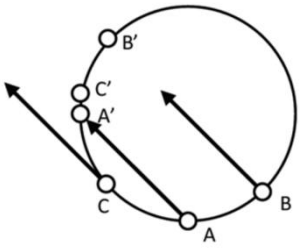
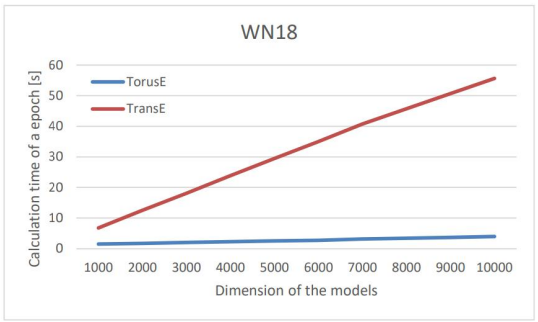


图4：实体间连接预测

TranE翻译模型的参数较少，计算复杂度低，且适用于大规模稀疏医学知识库，性能和扩展性都比较好。然而TransE的正则约束会迫使实体的向量表示在一个球面上，而这与之前的优化条件又是相互矛盾的。这种矛盾还会影响实体间连接预测的准确性。以图3为例，箭头方向表示关系r，A,B,C及表示实体，对于(C, r,C')和(B,r ,B')的正则约束和优化目标是相互矛盾的。因此[11]提出了一种基于李氏群的知识图谱嵌入表示学习 TorusE 拥有类似 TransE 遵循的优化目标和正则项。为了避免上 述的正则项带来的矛盾，TorusE 不再将特征学习到一个开流形（open manifold）的欧式空间，而是在紧空间（compact space）上学习知识图谱的嵌入表示。最终实验证明，TorusE 具有比 TransE 更低的计算复杂度，如图5所示。并在链路预测的任务上，TorusE 比当今最好的模型仍要表现出色。



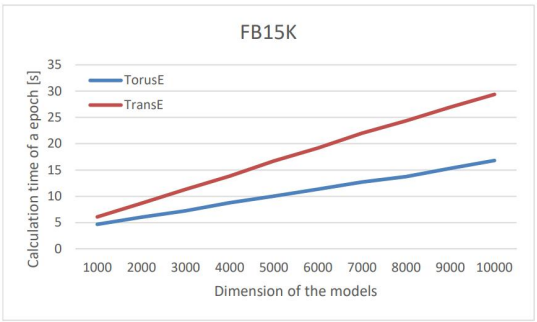


图5：TorusE 和 transE 分别在 WIN18 和 FB15K 的计算时间

2）深度学习模型：

[12]中提出一种多层卷积神经网络模型用于知识图谱的嵌入及链接预测任务。与自然语言处理中常用的一维卷积不同，文章通过把多个向量堆叠成矩阵，就可以像图形一样用二维卷积核来抽取 embedding 之间的关系。

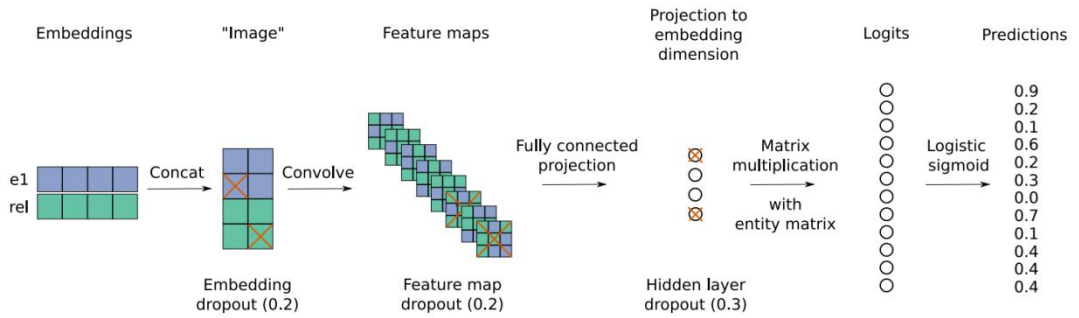


图 6：基于二维卷积的知识图谱嵌入表示学习、

值得一提的是，与传统模型对三元组关系（s, r, o）打分的 1-1 scoring 模式不同，ConvE 以实体关系对（s, r）作为输入，同时对所有实体 o 进行打分，即 1- N scoring。这种方式极大加快了计算速度。实验结果表明，即使实体个数扩大 10 倍，计算时间也只是增加了 25%。虽然该方法计算西能有所提高，但是在进行表示学习的时候，并没有关注到逻辑规则问题，因此[13]又提出来了一种基于规则迭代引导的增强知识图谱表示学习 RUGE 模型。

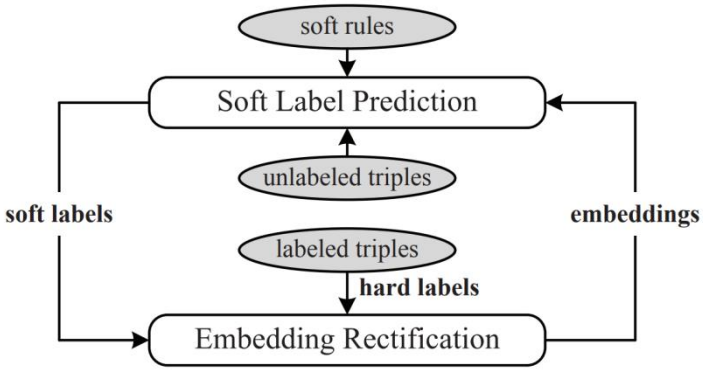


图 7：RUGE 框架图

图 7 是 RUGE 框架图，可以看出 RUGE 同时利用标注三元组（labeled Triples）、未标注三元组（Unlabeled Triples）、自动抽取出的软规则（soft rules）这三种资源以迭代的方式进行知识图谱表示学习。通过这个迭代过程，RUGE 可以成功建模分布式知识表示学习和逻辑推理二者间的交互性，逻辑规则中蕴含的丰富知识也能被更好地传递到所学习的分布式表示中。实验结果表明：RUGE 相比基线方法取得了较好的结果。[13]的创新性在于提出了软规则，并可以成功建模分布式知识表示学习和逻辑推理二者间的交互性，逻辑规则中蕴含的丰富知识也能被更好地传递到所学习的分布式表示中。

2.2 医学知识抽取

医学知识抽取是面向开放的医疗数据，通过人工或自动化技术抽取出可用的知识单元，知识单元包括实体、关系及属性这 3 个知识要素，并以此为基础，形成一系列高质量的事实表达，为上层模式层的构建奠定基础。

近年来深度学习被广泛应用于实体抽取中。目前 BiLSTM-CRF 是医学领域实体抽取中最主流的深度学习模型。文献[14]通过实验对比 BiLSTM-CR 与其他机器学习模型在医学电

子病历的实体抽取的效果，实验结果表明 BiLSTM-CR 对提高结果的准确率是有效的。

知识学习和深度学习的方法大多需要搜集大量语料，或过多依赖于专家的标注，而远程监督（distant supervision）能够减少对标注数据的需求，因此被大量应用于从非结构化医学文本中进行关系抽取文献[15]首先证明由于医学知识库的不完整，大量标记过程产生的否定标签为假否定，并基于此提出一种仅从实体对正标签进行学习的远程监督提取算法，并通过实验证明了此算法的有效性。

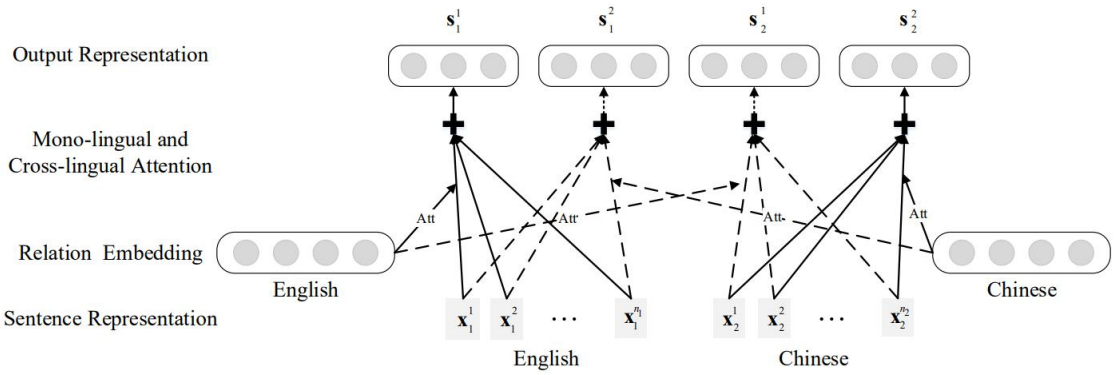


图 8：远程监督监督网络模型

文献[16]提出一种基于远程监督的卷积神经网络模型，利用卷积神经网络抓取实体的描述特征，丰富实体表示，并通过计算实体间关系与句子间的相似度赋予句子不同的权重。

上述的抽取方法主要关注二元实事，即两个实体之间的关系。然而在实际中，我们往往需要考虑三元关系甚至是更加高阶的关系，这种高阶的关系能够更加精确地表达一个事实。比如在医疗领域，我们需要知道一种药品治疗哪一种疾病，这种药品的使用剂量是什么以及适合哪种类型的患者（例如，儿童或者是成人）。

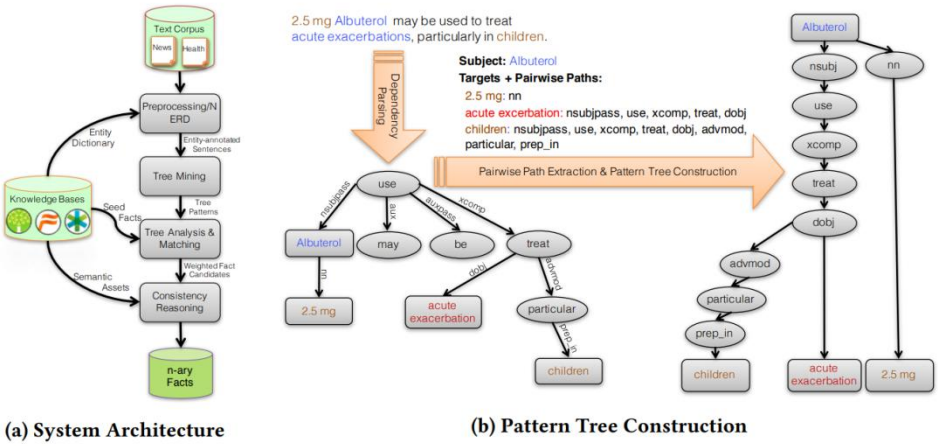


图 9：HighLife 模型

[17]中提出了一种从文本中获取高阶实事的 HighLife 模型，基于远程监督的假设，从一些种子集合出发。一方面为了提高召回率，利用二元的事实模板去发现 尽可能多的事实；另一方面为了提高准确率，设计了一种基于约束的推理方法来 去除错误的候选。主要的创新点是解决了高阶事实在文本中的表达不集中，分布不规律的问题。例如，一句话可以指一

种药物、一种疾病和一组患者，而另一句话则是指药物、其剂量和目标人群，而没有提到疾病。这篇文章的方法在模式学习和约束推理阶段都能很好地处理这些部分观察到的事实。对健康相关文档的实验证明了这种方法的可行性。

2.3 医学知识融合

由于医学数据库中的知识来源复杂，存在知识质量良莠不齐、不同数据源知识重复、知识间关联关系模糊等问题[18]，所以必须将来自不同数据源的多源异构、语义多样、动态演化的医学知识在同一框架规范下进行异构数据的整合、消歧、加工、推理验证、更新等,再将验证正确的知识与通过对齐关联，合并计算后组成知识库。 医学知识融合的关键技术有实体对齐技术、实体链接技术和关系推演技术:

1)实体对齐：

实体对齐用于消除异构数据中的实体冲突、指向不明等不一致问题，从而从顶层创建一个大规模的统一知识库，从而帮助机器理解多源异质数据，形成高质量知识。

2）实体链接

实体链接的主要作用是利用医学知识库中的实体对从医疗大数据的文本中获取的实体指代进行消歧，识别每一个实体指代在医学知识库中与其对应的映射实体。按照实体链接利用的信息不 同，现有工作主要分为基于实体属性（Entity attributes based，EA）的实体链接方法[19]基于实体流行(Entity Population based,EP)的实体链接方法[20]、基于上下文（Context based,CB）的实体链接方法[21]。

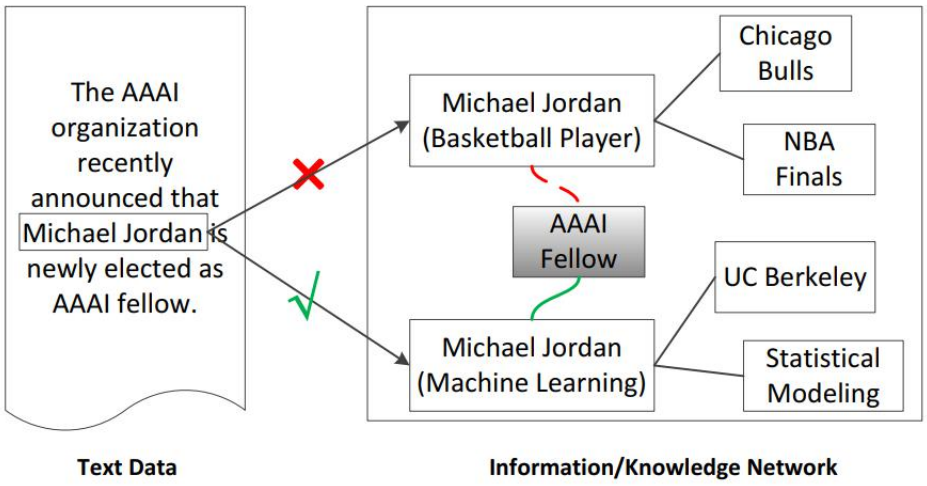


图 10：基于实体流行的实体链接方法

3.医学知识图谱的应用

知识图谱为医疗信息系统中海量、异构、动态的医疗大数据的表达、组织、管理及利用提供了一种更为有效的方式，使系统的智能化水平更高，更加接近于人类的认知思维。目前医学知识图谱技术主要用于临床决策支持系统、医疗智能语义搜索引擎、医疗问答系统、慢病管理系统等

3.1 临床决策支持

利用知识图谱技术可以辅助医疗行业和领域的大数据分析与决策，根据患者症状、检验、检查等数据，自动生成诊断、治疗方案，还可以对医生的诊疗方案进行智能化分析，有效减少误诊情况的发生。文献[22]提出一种面向重症监护室的急性心肌梗死患者的智能监测和决策支持系统，该系统的知识库由 OWL 本体和 1 组表示专家知识的规则组成，能够分析患的情况，并给出了治疗建议；文献[23]通过自然语言处理方法建立 3 层疾病结构知识图谱（疾病－症候－特征），运用正则表达式、隐马尔科夫模型等人工智能技术解决了构建医学知识图谱过程中效率低、耗时长等问题。

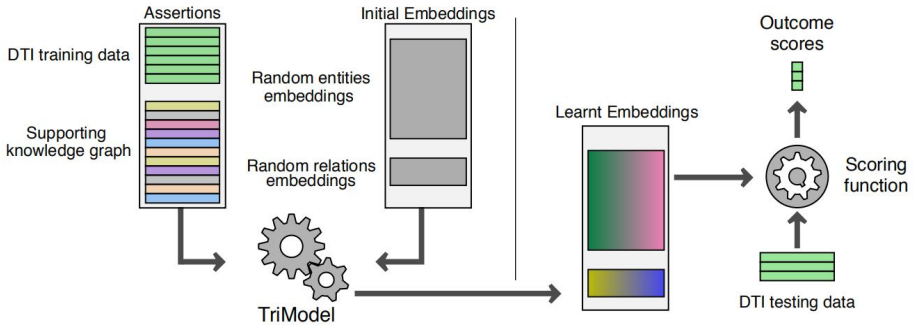


图 11：TriModel 模型

此外预测药靶点-靶点相互作用(DTIS)的计算方法可以为药物作用机制提供有价值的见解，意义重大，并且可以帮助快速识别药物的新的有希望的或意想不到的效果。但是，现有模型面临若干挑战。许多只能处理有限数量的药物和/或蛋白质组覆盖率差。当前的方法还经常遭受高的假阳性预测率。因此[24]提出了一种预测药物靶蛋白的新颖计算方法。该方法基于将问题转化化为知识图中的链接预测，他们使用生物医学知识库来创建与药物及其潜在靶标相关的实体的知识图，并出了一个特定的知识图嵌入模型 TriModel，以学习创建的知识图中所有药物和靶标的向量表示形式，基于这些表示再用于经过训练后的 TriModel 模型，从而计算出的候选药物靶标相互作用。

3.2 医疗问答

医疗问答系统是医疗信息检索系统的一种高级形式，能够以准确简洁的自然语言形式为用户提供问题的解答。多数基于知识图谱的医疗问答系统将给定的问题分解为多个小的问题，然后逐一去知识库抽取匹配的答案，并自动检测答案在时间和空间上的吻合度等，最后

将答案进行合并，以直观的方式展示给用户。

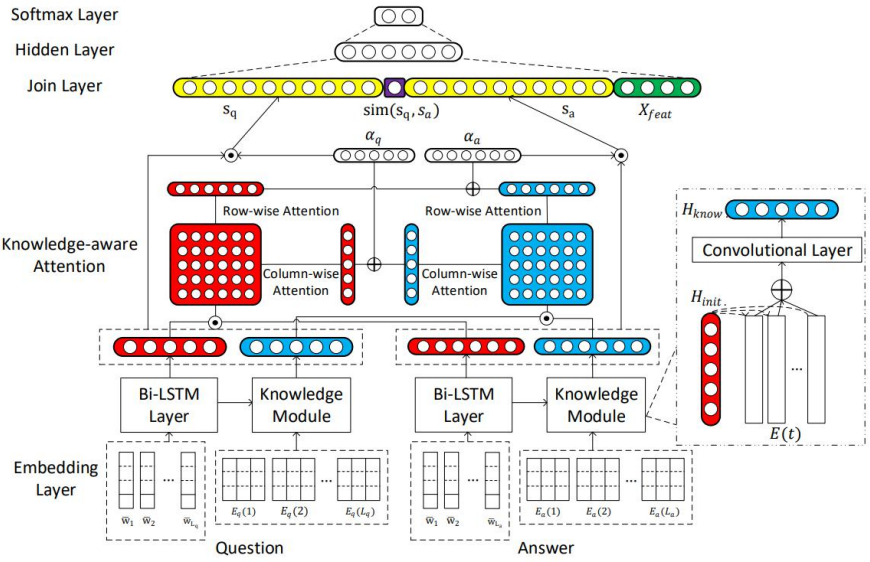


图 12：[25]双向长短记忆模型

[25]提出了一种具有知识感知能力的双向长短记忆模型，它利用医学知识图谱引入的背景知识来丰富问答的表征学习。模型的核心是一个上下文引导的注意力神经网络，通过将知识图谱中的背景知识嵌入整合到句子表示中，并结合知识型注意力机制模块，对问题和答案中的各个部分进行有效的相互关联。通过实验验证了该方法在 WikiQA 和 TREC QA 数据集上的均有不错的效果，但是在具有噪声的场景下，问句中的实体很难直接准确的匹配到知识库上。因此[26]提出了一个端到端的知识库问答模型 VRN 来解决这一问题。VRN 一种变分推理网络，模型分为两部分：通过概率模型来识别问句中的实体（得到图谱中每个实体是问句中实体的概率），这避免了语义解析带来的误差。具体而言，就是将问句（基于语音或者文本）映射成向量，然后对其做 softmax 多分类，以计算问句中的实体的概率。实验结果证明：该方法对于医疗问答准确性的提升具有一定有效性。

4.总结

在医疗领域中，随着医学信息化水平的逐步深入，积累了大量医学数据，医疗数据的有效使用对精准医疗、疾病防控、研发新药、医疗费用控制、攻克顽疾、健康管理等工作都有着重要的意义。构建医疗领域的知识图谱提供了一种从海量医学文本和图像中抽取结构化知识的手段，具有广阔的应用前景。本文从医学知识图谱构建的视角出发，对医学知识图谱的架构、医学知识图谱构建关键技术以及研究应用发展现状进行了全面调研和深入分析，并对医学知识图谱构建工作所面临的重要挑战和关键问题进行了总结。

知识图谱在医疗领域的意义不仅在于它是一个全局医学知识库，也是支撑例如辅助诊疗、智能搜索等医疗智能应用的基础，而且在于它是一把打开人类知识宝库的钥匙，它能够推进医学数据自动化和智能化处理，为医疗行业带来新的发展契机。

参考文献：

[1]Pujara J, Miao H, Getoor L, et al. Knowledge Graph Identification[C]. international semantic web conference, 2013: 542-557.

[2]Paulheim H. Knowledge graph refinment:A survey of approaches and evaluation methods[J]. Semantic Web,2017,8(3):489-508

[3]Percha B, Altman R B. A global network of biomedical relationships derived from text[J]. Bioinformatics, 2018, 34(15): 2614-2624.

[4]Murdoch T B, Detsky A S. The Inevitable Application of Big Data to Health Care[J]. JAMA, 2013, 309(13): 1351-1352.

[5]Wikipedia. Knowledge graph[OL]. [2016-05-09]. http://en.wikipedia.org/wiki/Knowledge_G.

[6]Nickel M, Murphy K, Tresp V, et al. A Review of Relational Machine Learning for Knowledge Graphs[J]. arXiv: Machine Learning, 2017, 104(1): 11-33.

[7]Percha B, Altman R B. A global network of biomedical relationships derived from text[J]. Bioinformatics, 2018, 34(15): 2614-2624.

[8]Maclennan A. The artificial life route to artificial intelligence: Building embodied, situated agents[J]. Journal of the Association for Information Science and Technology, 1996, 47(6): 482-483.

[9]Martinezgil J. Automated knowledge base management: A survey[J]. Computer Science Review, 2015: 1-9.

[10]Bordes A, Usunier N, Garciaduran A, et al. Translating Embeddings for Modeling Multi-relational Data[C]. neural information processing systems, 2013: 2787-2795.

[11]Ebisu T, Ichise R. TorusE: Knowledge Graph Embedding on a Lie Group[C]. national conference on artificial intelligence, 2018: 1819-1826.

[12]Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D Knowledge Graph Embeddings.[J]. arXiv: Learning, 2017.

[13]Guo S, Wang Q, Wang L, et al. Knowledge Graph Embedding with Iterative Guidance from Soft Rules[J]. arXiv: Artificial Intelligence, 2017.

[14]Jagannatha A, Yu H. Structured prediction models for RNN based sequence labeling in clinical text[J]. arXiv: Computation and Language, 2017.

[15]Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance Multi-label Learning for Relation Extraction[C]. empirical methods in natural language processing, 2012: 455-465.

[16]Lin Y, Liu Z, Sun M, et al. Neural Relation Extraction with Multi-lingual Attention[C]. meeting of the association for computational linguistics, 2017: 34-43.

[17]Ernst P, Siu A, Weikum G, et al. HighLife: Higher-arity Fact Harvesting[J]. the web conference, 2018: 1013-1022.

[18]Lin Hailun, Wang Yuanzhou,Jia Yantao,et al.Network big data oriented knowledge fusion methords:A survey [J].Chinese Journal of Computers, 2017,23(1):1-27(in Chinese)

[19]Chen R, Bau C, Yeh C, et al. Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques[J]. Applied Soft Computing, 2011, 11(2): 1908-1923.

[20]Li Y, Wang C, Han F, et al. Mining evidences for named entity disambiguation[C]. knowledge discovery and data mining, 2013: 1070-1078.

[21]Bhattacharya I, Getoor L. Collective entity resolution in relational data[J]. ACM Transactions on Knowledge Discovery From Data, 2007, 1(1).

[22]Martinezromero M.Vazqueznaya J M,Pereira J,et al. The iOSC3 system:Using ontologies and SWRL rules for disorders[J]. Computational and Mathematical Methords in Medicine,2013,2013(5904):650-671

[23]Nie Lili, Li Chuanfu, Xu Xiaoqian, et al. Study onapplication of artificial intellifence in the building of medicaldiagnosis knowledge-graph[J]. Journal of Medical Informatics,2018,5(6):7-12(in Chinese)

[24]Mohamed S K, Novacek V, Nounu A, et al. Discovering Protein Drug Targets Using Knowledge Graph Embeddings[J]. Bioinformatics, 2019.

[25]Shen Y, Deng Y, Yang M, et al. Knowledge-aware Attentive Neural Network for Ranking Question Answer Pairs[C]. international acm sigir conference on research and development in information retrieval, 2018: 901-904.

[26]Zhang Y, Dai H, Kozareva Z, et al. Variational Reasoning for Question Answering with Knowledge Graph[C]. national conference on artificial intelligence, 2018: 6069-6076.