北京交通大学
BEIJING JIAOTONG UNIVERSITY

文献报告

Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction[1]

将医疗知识图谱融入大语言模型以进行诊断预测

张鳊沣 22281052

2024年10月12日

2024–2025学年 《知识表示与处理》 实验2

[1] Gao Y, Li R, Caskey J, Dligach D, Miller T, Churpek MM, Afshar M. Leveraging a medical knowledge graph into large language models for diagnosis prediction[J]//arXiv preprint arXiv:2308.14321. 2023.

北京交通大学
BEIJING JIAOTONG UNIVERSITY

Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction

将医疗知识图谱融入大语言模型以进行诊断预测

CONTENTS

01 问题定义

02 方法原理

03 研究结果

04 创新特色

05 个人体会

核心问题

探讨基于医疗知识图谱的模型如何为生成诊断的基础模型提供知识，提高自动诊断生成的准确性和可解释性

文献综述

随着**电子健康记录（EHR）**（如图1）的广泛使用，临床笔记中包含的大量复杂信息给医疗从业人员带来了认知负担，可能导致诊断错误[2]。为了解决这一问题，本篇文献提出了一种将**医学知识图谱（KG）**融入**大语言模型（LLMs）**的创新方法，用于自动生成诊断。研究引入了一种名为**DR.KNOWS的图模型**，借助临床诊断推理过程，利用来自美国国家医学图书馆的统一医学语言系统（UMLS）构建知识图谱。研究表明，通过结合LLMs与KG，这种方法能够提高自动诊断的准确性，并提供可解释的诊断路径，为未来的人工智能诊断决策支持系统奠定了基础。

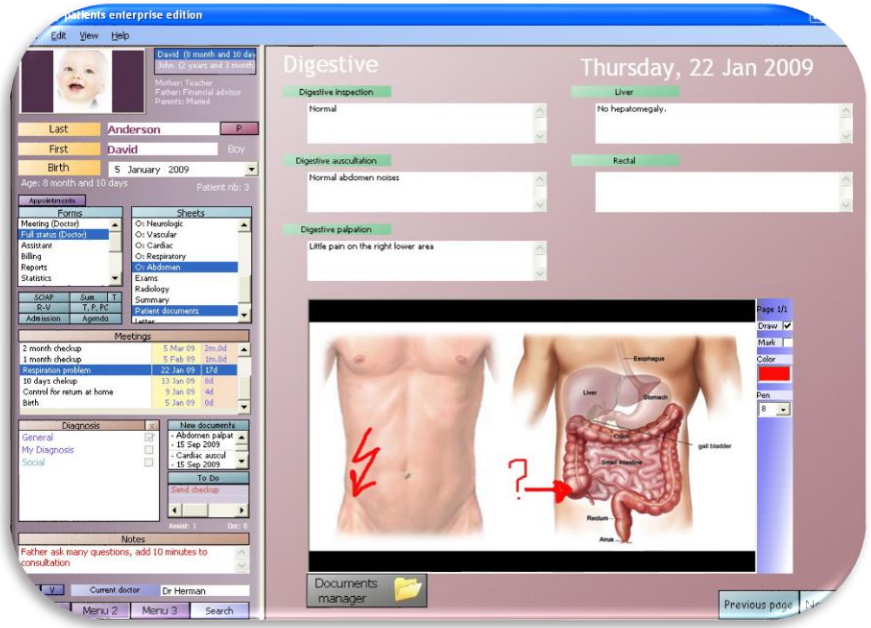

图1 Sample view of an electronic health record（源自Wikipedia）

TIPS: 电子健康记录（EHR）

电子健康记录是存储于计算机系统之中、面向个人提供服务、具有安全保密性能的终身个人健康档案。[3] 电子健康记录的一些组成部分包括叙述性文本、结构化数据输入和/或直接导入的患者数据，例如生命体征、实验室结果或药物清单。[4]

[2] Croskerry P. Diagnostic failure: A cognitive and affective approach[C]//Henriksen K, Battles JB, Marks ES, et al., editors. Advances in patient safety: From research to implementation (Volume 2: Concepts and methodology). Rockville (MD): Agency for Healthcare Research and Quality (US), 2005. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20487/
[3] 董建成. 医学信息学概论[M]. 北京: 人民卫生出版社, 2010: 133.
[4] Brown PJ, Marquard JL, Amster B, Romoser M, Friderici J, Goff S, Fisher D. What do physicians read (and ignore) in electronic progress notes?[J]//Applied clinical informatics. 2014, 5(02): 430-444.

国内外研究现况——LLM & 医学领域

近年来，LLM在多个领域展现了广泛的应用潜力。LLM基于深度学习和自然语言处理技术，已被广泛应用于生成式人工智能任务，如文本生成、对话系统等[5]。在医学领域，LLM正在改变诊断、病例撰写等多方面的工作方式。**LLM能够辅助医生生成医疗报告、改进患者沟通，并简化文档处理流程[6,7]**。例如ChatGPT和BioGPT在医学诊断和文本生成中取得了与人类专家相当的表现[7]，周雪忠等人为 LLM 中的生物医学知识设计了五个具体的识别和评估任务对LLM的KB进行评估研究[8]。

然而，LLM在医学应用中仍面临许多挑战，包括对复杂医学上下文的理解有限、可能产生偏见或错误信息、以及在任务执行中的不一致性[7,9]。例如LLM可能无法准确评估医学知识的准确性和全面性，这在医疗决策和诊断中至关重要。为此许多学长也在优化这方面的缺陷，例如 Harrer S 等人提出了基于医学知识图谱的评估框架，以量化LLM在医学知识处理中的表现[9]。

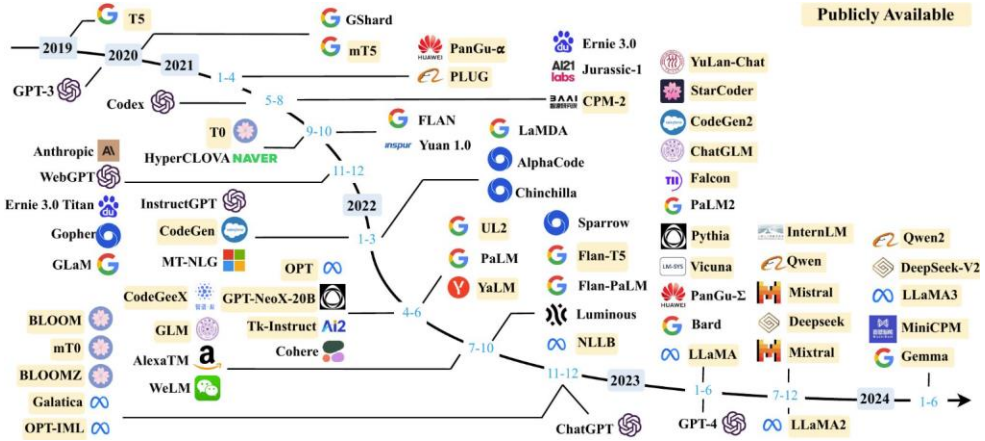

图2 近年来现有大语言模型（大小超过 10B）的时间线[2]

该时间线主要根据模型技术论文的发布日期（如提交到 arXiv 的日期）建立。如果没有相应的论文，图作者将模型的日期设置为其公开发布或宣布的最早时间。作者将具有公开可用模型检查点的 LLM 标记为黄色。由于图表的空间限制，作者只包含具有公开报告的评估结果的 LLM。

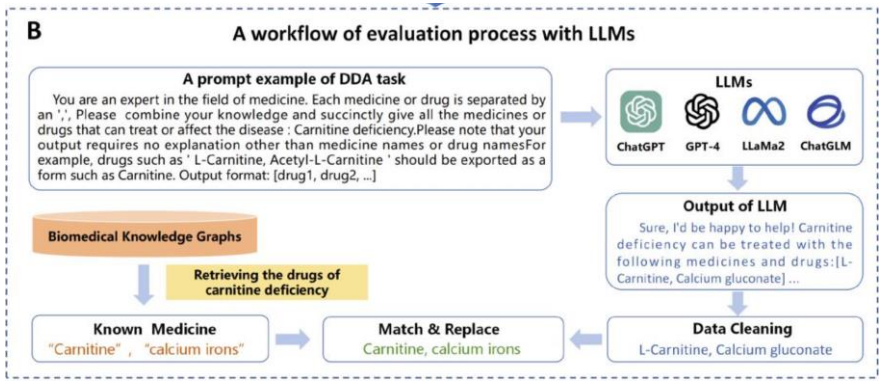

图3 用于大语言模型的生物医学评估框架概述——说明大语言模型评估过程的工作流程图[3]

[5] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data[C]//Advances in Neural Information Processing Systems. 2013: 2787-2795.
[6] Zhao W X, Zhou K, Li J, et al. A Survey of Large Language Models[C]//arXiv Preprint arXiv:2303.18223, 2024.
[7] Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: A scoping review[J]//iScience. 2024, 27: 109713.
[8] Zhang F, Yang K, Zhao C, et al. Benchmarking biomedical relation knowledge in large language models[C]//Bioinformatics Research and Applications. 2024, 14955: 1-14.
[9] Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine[J]//eBioMedicine. 2023, 90: 104512.

国内外研究现况——EHR运用现况

电子健康记录（EHR）作为日常护理记录的电子标准化文档，已经在医疗行业里面广泛使用，对于确保患者护理的连续性至关重要。这些记录详细记录了患者的健康状况、诊断及治疗方案。[3,10]

然而，随着EHR的不同规范的指定，以及临床中叙述的复杂性和冗长性语句不断增加，大量的EHR包含了许多重复信息，这给医疗人员带来了认知负担，并可能导致诊断错误。医生常常跳过冗长和重复的内容，依赖于决策捷径（决策启发式）[2]，这进一步增加了误诊的风险。

目前自动生成ERH主要依赖于大语言模型，许多学者运用T5和GPT进行开发与评估[11,12]，这表明大语言模型与ERH的结合研究受到了广泛关注。然而，自动诊断生成需要极高的准确性和可靠性。但模型可能生成误导性或虚假信息，导致诊断准确性受到影响[13]，因此大语言模型在诊断预测的应用引发了一定的担忧。

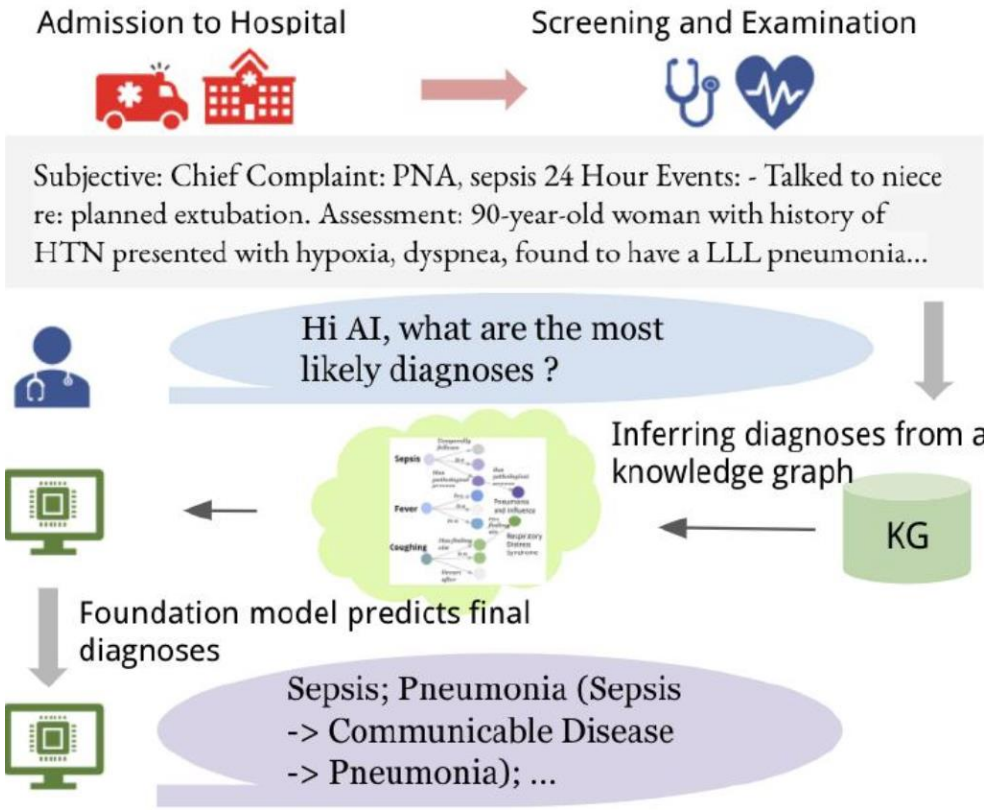

图4 人工智能辅助诊断流程[1]

首先病人入院并经过检查，医生输入患者的症状和病史信息后询问AI可能的诊断。AI通过一个KG推理出相关的疾病和症状关系，然后模型根据推理结果进行最终诊断预测，最终为医生提供一个可能诊断（例如图中根据输入内容诊断出了脓毒症和肺炎），供医生做出临床决策。

[10] Brown PJ, Marquard JL, Amster B, Romoser M, Friderici J, Goff S, Fisher D. What do physicians read (and ignore) in electronic progress notes?[J]//Applied clinical informatics. 2014, 5(02): 430-444.
[11] Gao Y, Dligach D, Miller T, Xu D, Churpek MM, Afshar M. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2979-2991.
[12] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences[J]//Minds and Machines. 2020, 30: 681-694.
[13] Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine[J]//Clinical and translational medicine. 2023, 13(3).

相关工作

作者通过研究整理集中在临床笔记的摘要，包括出院总结、患者就诊的实时总结以及问题和诊断列表，为后续进行的模型训练进行整理。作者的工作遵循先前研究者对于上述问题的研究进行整理，以相关患者病历问题和诊断摘要作为研究方向。

作者进行整理发现，将KG整合到语言模型（LM）中已成为一种新兴趋势，因为这在一定程度上可以增强专家系统与大语言模型的事实知识可信度，特别是在特定领域的问答任务中。作者通过KG的整理，整合到LLMs中来预测诊断，并使用一种新的图模型来生成基于路径的提示。

研究将要解决的问题

SOAP格式的每日进展记录将患者的主观症状、客观数据、评估和治疗计划进行结构化记录。**本研究第一个任务是预测记录中计划部分中包含的问题和诊断列表。**

UMLS知识图谱通过语义关系和概念分类，帮助从患者病历中识别潜在诊断。**本研究第二个任务是利用UMLS中的相关属性对概念进行分类，实现高效探索疾病关联，以及探索跨各种医学词汇的语义理解和知识发现。**

数据集和环境

本研究使用了来自不同临床环境的两组进展记录：MIMIC-III 和 IN-HOUSE HER 数据集。

MIMIC-III是最大的**公开**数据库之一，包含来自 Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center (BIDMC)，包含了2001年至2012年超过38,000名ICU病人数据。遵循SOAP格式记录病程记录。

IN-HOUSE EHR数据集是EHR的一个私有子集（**强隐私数据**），包含2008年至2021年间在美国一家医院住院的成年患者。与MIMIC集相比，IN-HOUSE集涵盖了所有医院环境的病程记录，包括急诊科、普通内科病房、亚专科病房等。遵循SOAP格式记录病程记录。

项目使用了3种不同的计算环境进行训练：

公共EHR数据集（MIMIC-III）是在Google云计算（GCP）上完成的，利用1-2个NVIDIA A100 40GB GPU和一台配备1个RTX 3090 Ti 24GB GPU的传统服务器。

受保护健康信息的私有EHR数据集（IN-HOUSE EHR）在存储在医院研究实验室内的工作站上进行训练，配备一个NVIDIA V100 32GB GPU，确保受电子保护健康信息 (ePHI) 的机密性、完整性和可用性。

Dataset	Dept.	#Input CUIs	#Output CUIs	% AbstCon
MIMIC-III	ICU	15.95	3.51	48.92%
IN-HOUSE	All	41.43	5.81	<1%

图5 两个 EHR 数据集（MIMIC-III 和 IN-HOUSE HER 数据集）的输入和输出CUI（UMLS中概念唯一标识符）的平均数量。[1]

IN-HOUSE数据集包含所有医院环境，MIMIC-III数据集则专注于ICU。“% AbstCon”代表模型在预测诊断时使用抽象概念的比例，反映了CUI与输出的诊断概念中，通过抽象推理生成的比例，反应了抽象推理的贡献大小。

UMLS知识图谱推断原理

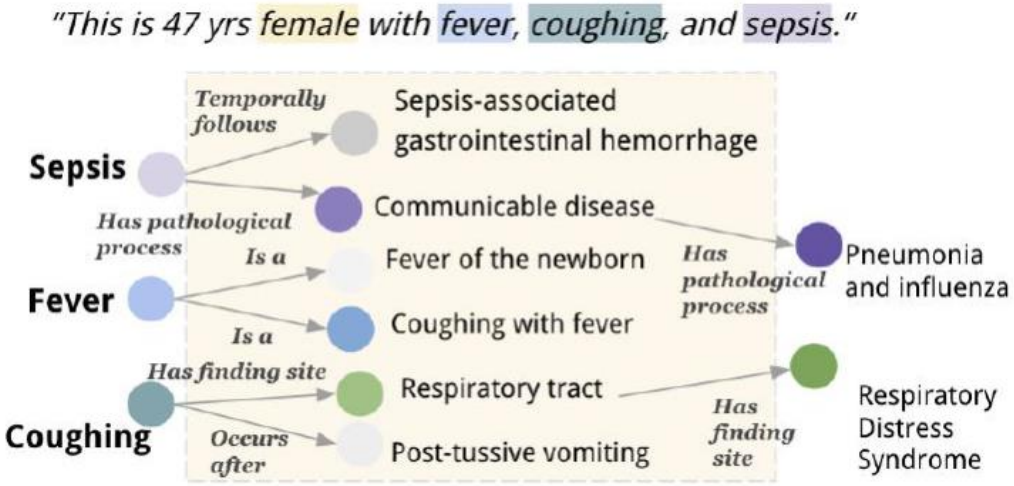

图5 UMLS知识图谱推断示意图，其中省略了“女性”的子图

根据患者的病历描述， UMLS 知识图谱可以推断出一定范围内的可能诊断。在图5中，作者用彩色方框突出显示 UMLS 医疗概念（“female”、“sepsis 败血症”等），其中概念是顶点，语义关系是边，每个概念都有自己的子图。在第一跳（hop）中可以识别出与输入描述最相关的邻近概念。顶点的颜色越深，它们与输入描述的相关性就越高。可以根据最相关的节点进一步执行第二跳，并最终得出诊断“Pneumonia 肺炎 和 Influenza 流感”和“Respiratory Distress Syndrome 呼吸窘迫综合征”。

DR.KNOWS 的架构设计

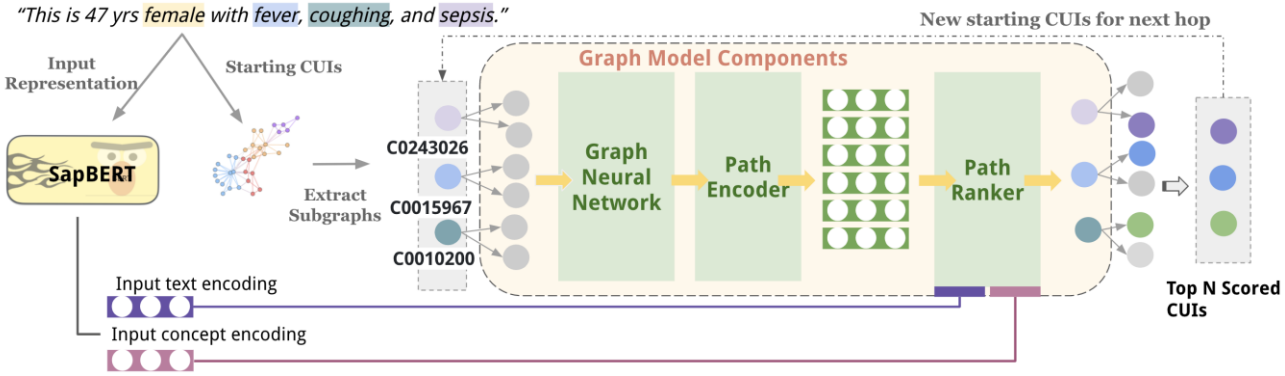

图6 DR.KNOWS 模型架构

在DR.KNOWS模型中输入患者文本，识别出CUI （SNOMED 概念标识符）的 UMLS 概念，在**UMLS KG**中检索出第一跳（ 1-hop ）子图。这些子图由堆叠图同构网络 (SGIN) 编码表示，然后馈送到**Path Encoder 路径编码器**，生成路径表示。**Path Ranker 模块**通过考虑路径与输入文本和概念的语义和逻辑关联，评估1-hop 路径，并根据路径表示、输入文本和概念表示生成评分。在所有指向这些节点的路径中，选取得分排名前N的节点，指导后续跳跃的路径探索。如果未能找到合适的诊断节点，算法将终止于指向当前节点。

上下文化节点表示

作者定义了基于SNOMED CUI和语义关系的确定性UMLS知识图谱 $G = (V, E)$，其中 V 是CUI的集合，E 是语义关系的集合。给定包含源CUI集合 $V_{src} \subseteq V$ 和其1-hop关系 $E_{src} \subseteq E$ 的输入文本x，作者可以为每个 $\langle v_i \rangle_{i=1}^{I} \subseteq V_{src}$ 构建关系路径 $P = p_1, p_2, \ldots, p_J$，其中每个路径 $p_j = v_1, e_1, v_2, \ldots, e_{t-1}, v_t$，且 t 是预定义的标量，J 是非确定性的。

节点通过SapBERT编码 l_i 获得语境化表示 h_i，并通过Stack Graph Isomorphism Network (SGIN) 更新，公式如下：

$$h_i^{(k)} = MLP^{(k)}\left(\left(1 + \epsilon^{(k)}\right)h_i^{(k)} + \sum_{s \in N(v_i)} ReLU\left(h_s, e_{s,i}\right)\right)$$

最终节点表示通过叠加多层GIN（图同构网络）计算：

$$h_i = \left[h_i^{(1)}; h_i^{(2)}; \ldots; h_i^{(K)}\right]$$

其中，$N(v_i)$ 表示节点 v_i 的邻居节点集合，$h_i^{(k)}$ 是第 k 层的节点表示，$\epsilon^{(k)}$ 是可学习参数，MLP 是多层感知机。

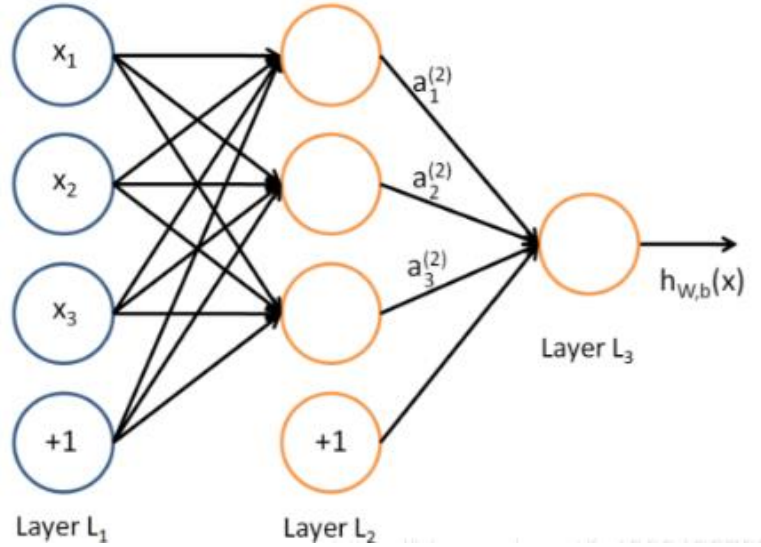

图7 多层感知机MLP

MLP最底层是输入层，中间是隐藏层，最后是输出层。 MLP层与层之间是全连接的。

PS：多层感知机（Multilayer Perceptron，MLP）

MLP是神经网络（Neural Network）的一种特殊类型。神经网络是一种模拟人脑神经元网络的计算模型，由大量简单处理单元相互连接组成，能够进行分布式信息处理。而MLP特指包含至少一个隐藏层的神经网络，其中隐藏层负责从输入层接收信息并传递到输出层。

两者之间存在显著差异。**神经网络可以拥有多个隐藏层，而多层感知机通常指只有一层隐藏层的神经网络。**此外，神经网络的连接权重可以随着训练过程中迭代次数的增加而不断调整，以优化网络的性能。而多层感知机的权重通常是固定的，不会随着训练次数的增加而改变。[14]

[14] c4t. 多层感知机与神经网络的联系与区别[EB/OL]. Baidu Developer. Available from: https://developer.baidu.com/article/detail.html?id=3157781.

路径推理与排序

每个节点表示 h_i 的邻居节点路径嵌入生成如下：

$$\mathrm{p_i} = \begin{cases} h_i, n = 1 \\ \mathrm{p_{t,i}^{(n-1)}}, \text{others} \end{cases}$$

并通过以下公式计算路径嵌入：

$$p_{t,i}^{(n)} = \mathrm{FFN}\left(W_i h_i^{(n)} + W_t\left[e_{t,i}^{(n)}, h_{t,i}^{(n)}\right]\right)$$

通过多头注意力机制（MultiHead attention, MultiAttn）或三线性注意力机制（Trilinear attention, TriAttn）计算逻辑关系：

$$H_i = [h_x; p_i; h_x - p_i; h_x \odot p_i], H_i \in R^{4D}$$
$$Z_i = [h_v; p_i; h_v - p_i; h_v \odot p_i], Z_i \in R^{4D}$$

注意力分数：

$$\alpha_i = \mathrm{MultiAttn}(H_i \odot Z_i)$$

并通过softmax函数聚合分数，选择下一个跳跃的Top N节点：

$$\beta = \mathrm{Softmax}\left(\sum_{i=1}^{V_{src}}\sum_{t=1}^{T} S_{i,t}^{\mathrm{Tri}}\right)$$

损失函数

作者定义了预测损失 L_{Pred} 和对比学习损失 L_{CL}，总损失为：

$$L = L_{Pred} + L_{CL}$$

其中预测损失使用二元交叉熵计算：

$$L_{Pred} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(y_{m,n}\log(v_{m,n}) + (1 - y_{m,n})\log(1 - v_{m,n})\right)$$

对比学习损失通过余弦相似度计算：

$$L_{CL} = \sum_{i}\max(\cos(A_i, f_i^+) - \cos(A_i, f_i^-) + \mathrm{margin}, 0)$$

PS：前馈神经网络（Feed-Forward Neural Network，FFN）

FFN是深度学习中最基础的网络结构之一。它由输入层、隐藏层和输出层组成，数据从输入层到输出层单向传播，不发生回流。FFN是一个顺序结构：包括一个全连接层(FC) + relu激活层 + 第二个全连接层，公式表达如下：

$$\mathrm{FFN}(x) = f\left(x \cdot W_1^{\top}\right) \cdot W_2$$

DR.KNOWS 在 CUI 预测任务上的评估

作者在 IN-HOUSE 和 MIMIC 数据集上训练 DR.KNOWS（TriAttn 和 MultiAttn）及其加权变体（TriAttn_W和 MultiAttn_W）。

作者在 MIMIC 数据集上获得了 600、81 和 87 的数据分割，在 IN-HOUSE 数据集上获得了 3885、520 和 447 的数据分割。

在评估任务中，作者要评估 DR.KNOWS 使用 CUIs 预测诊断的准确性，为此作者使用概念提取器分析计划部分的文本，并提取属于语义类型 T047 疾病和综合征的 CUIs。具体来说，作者包含了那些保证至少有一条路径，其最大长度为目标 CUI 和输入 CUI 之间 2-hop的 CUI。这些选定的 CUI 构成了最优 CUI 集，用于训练和评估模型的性能。

作者将 DR.KNOWS 与两个不同的基线系统进行比较：QuickUMLS是一个概念提取基线系统，用于识别医学概念。Vanilla-T5 是一种生成式语言模型，针对此任务作者进行了微调，并预测有关诊断的概念。作者使用 QuickUMLS 将 T5 生成的文本解析为医学概念列表。图7、图8展示了 DR.KNOWS 内在评估结果。不同数据集长度的差异决定了选择不同的前 N 个值。**在两个数据集上，DR.KNOWS 的表现都显示出更高的精确度和 F 分数，表明其在预测正确诊断 CUI 方面的优势。**T5 的表现不佳可以归因于它倾向于生成与诊断和疾病无关的概念，而不是专注于所需的概念。

Model	Top N	Recall	Precision	F-Score
Concept Ex.	-	**56.91**	13.59	21.13
T5-L	-	12.95	9.77	9.61
MultiAttn	4	28.80	24.15	**24.77**
	6	30.76	17.41	20.87
MultiAttn_W	4	26.91	22.79	23.10
	6	29.14	16.73	19.94
TriAttn	4	27.77	**24.44**	24.63
	6	32.96	16.77	19.94
TriAttn_W	4	29.85	17.61	20.93
	6	29.85	17.61	20.93

图8 MIMIC-III 数据集上目标 CUI 预测的概念提取基线 (Concept Ex.)、T5-Large 基线 (T5-L) 和 DR.KNOWS的Recall、Precision、F-Score值

Model	Top N	Recall	Precision	F-Score
Concept Ex.	-	**90.11**	12.38	20.09
MultiAttn_W	6	24.68	15.82	17.69
	8	28.69	15.82	17.33
TriAttn_W	6	34.00	**22.88**	23.39
	8	44.58	22.43	**25.70**

图9 使用内部数据集进行目标 CUI 预测时，概念提取 (Concept Ex.) 与两种 DR.KNOWS 变体之间的的Recall、Precision、F-Score值及对比。

基础模型的路径提示&大语言模型评估

作者**将 DR.KNOWS 作为基于知识路径的提示提供给大语言模型**，观察大语言模型在诊断摘要方面受到的影响。作者探究有7.7亿参数的T5-Large模型和有1540亿参数的 GPT-3.5-turbo模型。类似于T5 这样的较小模型提供更易于控制，而类似于 GPT 这样的较大模型则以更大的规模和能力生成文本。**作者的研究重点是在 PROBSUM 测试集上评估微调后的T5模型与和零样本设置的GPT模型的性能。**

在BETTERPROMPT框架指导下，作者手动设计了五组提示以整合路径输入（如图10），前三组提示由非医学领域专家（计算机科学家）设计，而最后两组提示则由医学领域专家（重症监护医师和医学信息学家）开发。作者将最后两个提示指定为“主题提示”，将前三个提示指定为“非主题提示”。

生成的诊断摘要的质量使用 ROUGE 分数 和 CUI F-score 进行评估，CUI F-score 是一种临床指标，它结合了 CUI Recall 和precision。ROUGE 分数定量地衡量了系统生成的摘要和参考摘要之间的重叠，但难以处理缩写和首字母缩略词。CUI F-score 利用 UMLS 概念，考虑了医学术语。这些指标共同提供了对模型性能的全面评估，确保了对生成输出的整体评估。

Group	Output Customization Prompts and No.	Perplexity
Non-Subj.	A. <Explain> B.You may utilize these facts: C. You may find these facts helpful:	3.86e-13
Subj.	D. /E. *Act as a medical doctor, and list the top three direct and indirect diagnoses from the Assessment.*(You will be provided with some hints from a knowledge graph.) Explain the reasoning and assumptions behind your answer.	1.03e-03
No.	**Context Control with Path Presentation**	
1	Structural.e.g. *“Infectious Diseases –> has pathological process –> Pneumonia "*“–>" is added to tokenizers as a new token.	
2	Clause. e.g. *“Infectious Diseases has pathological process Pneumonia"*	

图10 为路径提示实验创建的五个手动设计的提示（输出定制）和两种路径表示风格（上下文控制）

总共有 10 种提示模式（5 种输出定制 x 2 种上下文控制）。对于每个输出定制提示，我们使用 ChatGPT 生成 50 个释义，并运行 BETTERPROMPT 获取困惑度。此表还包括每个提示的平均困惑度。带有 * 的提示也被用于无路径 T5 微调（基线）。

Model	R2	RL	CUI-R	CUI-P	CUI-F
Prompt-based Fine-tuning					
VanillaT5	12.66	29.08	39.17	22.89	26.19
+Path	13.13	**30.72**	**40.73**	**24.28**	**27.78**
FlanT5	11.83	27.02	38.28	22.32	25.32
+Path	**13.30**	30.00	38.96	24.74	27.38
ClinicalT5	11.68	25.84	30.37	17.91	19.61
+Path	12.06	25.97	29.45	22.78	23.17
Prompt-based Zero-shot					
ChatGPT	7.05	19.77	23.68	15.52	16.04
+Path	5.70	15.49	25.33	17.05	18.21

图11 不同基础模型在 PROBSUM 测试集上的最佳性能

此处包括所有提示风格，包括使用和不使用DR.KNOWS路径提示。作者报告了 ROUGE-2 (R2)、ROUGEL (RL)、CUI Recall (CUI-R)、Precision (CUI-P) 和 F-score (CUI-F)，以更好地说明性能差异。

基础模型的路径提示&大语言模型评估

尽管接受相同的输入，ChatGPT 和 T5 表现出不同的行为。图12展示了一个复杂案例，以及 DR.KNOWS 和基础模型的预测结果。注释中的大多数诊断可以直接从输入笔记中提取。对于这些诊断，**DR.KNOWS 预测了带有自循环的路径。它还成功地预测了败血症诊断，这需要对血液和其他症状进行抽象和关联。**

在检查模型输出后，**观察到只有 ChatGPT+Path 使用了败血症诊断路径，准确地预测了诊断为“败血症”。**作者进一步发现，在 237 个测试样本中，38%的 ChatGPT+Path 输出明确提到，诊断是从其推理中的知识路径/图推断或支持的，在其他模型中没有发现这种模式。

另一个重要观察结果是，**ChatGPT+Path 利用路径中出现的概念名称，而其他模型倾向于从输入中复制名称。**例如，系统性动脉高血压与输入中的 HTN 具有相同的 CUI（C0020538），但名称不同。ChatGPT+Path 从路径而不是 HTN 中获取此名称，这可能会导致 ROUGE 分数降低。

此外，作者注意到**图模型可能会产生不太相关的路径，这可能会影响到基础模型。**例如，给定一个患有肺炎、高血压和呼吸窘迫为主要诊断的患者，预测连接急性概念到溃疡应激的路径会导致基础模型生成溃疡应激。作者认为这个问题是未来改进 DR.KNOWS性能的方向。

Model	R2	RL	CUI-R	CUI-P	CUI-F
Prompt-based Fine-tuning					
VanillaT5	12.66	29.08	39.17	22.89	26.19
+Path	13.13	**30.72**	**40.73**	**24.28**	**27.78**
FlanT5	11.83	27.02	38.28	22.32	25.32
+Path	**13.30**	30.00	38.96	24.74	27.38
ClinicalT5	11.68	25.84	30.37	17.91	19.61
+Path	12.06	25.97	29.45	22.78	23.17
Prompt-based Zero-shot					
ChatGPT	7.05	19.77	23.68	15.52	16.04
+Path	5.70	15.49	25.33	17.05	18.21

图11 不同基础模型在 PROBSUM 测试集上的最佳性能

此处包括所有提示风格，包括使用和不使用DR.KNOWS路径提示。作者报告了 ROUGE-2 (R2)、ROUGEL (RL)、CUI Recall (CUI-R)、Precision (CUI-P) 和 F-score (CUI-F)，以更好地说明性能差异。

Input Assessment and Subjective: Mr [****] is a 68 yo M s/p fall while EtOH intox > skull, SAH, SDH. Active issues: PNA resolving, HTN, withdrawal, ileus. Impression: Although presenting with head injury c/b SAH and SDH, his admission has been complicated by an ileus, worsening renal failure and stupor. <Subjective> Chief Complaint: Mr is a 68 yo M s/p fall while EtOH intox > skull, SAH, SDH. Active issues: PNA resolving, HTN, withdrawal, ileus. 24 Hour Events: BLOOD CULTURED . . .
Annotated Diagnoses Summary: Ileus; PNA; sepsis - ATN; uremic; SAH/SDH.

Graph Model Top 6 Predicted Paths: (↻ denotes a self-loop at the second hop, ‖ denotes the path separator.) ileus $\xrightarrow{\text{self}}$ ileus ↻ ‖ Operation $\xrightarrow{\text{occurs after}}$ Chronic ↻ ‖ Renal failure $\xrightarrow{\text{is a}}$ Unspecified chronic renal failure ↻ ‖ Blood $\xrightarrow{\text{has finding site}}$ Sepsis ↻ ‖ Renal failure $\xrightarrow{\text{self}}$ Renal failure ↻ ‖ systemic arterial hypertension $\xrightarrow{\text{self}}$ systemic arterial hypertension ↻

VanillaT5: SAH SDH ileus **VanillaT5+Path**: EtOH intox; fall; SAH SDH **FlanT5**: Although . . . head injury c b SAH and SDH, admission . . . by an ileus, worsening renal failure and stupor **FlanT5+Path**: SDH; Leukocytosis; GI bleed; EtOH intox; SAH **ClinicalT5:**s p fall; acute on chronic renal failure **ClinicalT5+Path:** SDH; Leukocytosis; GI bleed; EtOH intox; SAH

ChatGPT: 1. skull fracture SAH SDH 2. ileus 3. worsening renal failure <Reasoning> The first diagnosis is based on the patient's chief complaint, and the CT scan results, which show skull fracture, SAH, and SDH. The second diagnosis is based on the 24-hour events, which show that the patient has an ileus. The third diagnosis is based on the creatinine levels, which are rising, urine lytes ordered, which suggests worsening renal failure.
ChatGPT+Path: systemic arterial hypertension; ileus; Renal failure; Unspecified chronic renal failure; Sepsis <Reasoning> The input note mentions systemic arterial hypertension, ileus, renal failure, and unspecified chronic renal failure as active issues, so these diagnoses are directly extracted from the input note. Sepsis is inferred from the fact that blood was cultured and sepsis is a possible complication of the input diagnoses. Although there is a history of fall and head injury, the note does not provide any information about the current status of the head injury, so I did not include head injury as a differential diagnosis.

图12 不同基础模型推断结果示例图

顶部：一个单独呈现复杂病例的示例输入笔记，以及带注释的诊断摘要。中部：DR.KNOWS (TriAttnw) 预测了六条路径。底部：T5 模型和 ChatGPT 模型输出带有和不带有路径的结果。

创新点

结合大语言模型(LLMs)与医疗知识图谱(KG)用于诊断预测： 本文提出的DR.KNOWS模型是将基于UMLS的医疗知识图谱与大语言模型相结合的创新方法。与传统依赖于预训练的LLMs不同，本文方法通过直接利用知识图谱中的语义关系进行推理，有效增强了诊断生成任务中的准确性和可解释性。

DR.KNOWS路径选择算法的提出： DR.KNOWS模型通过选择与输入文本和概念关联最紧密的hop路径，并基于这些路径为后续探索提供指导。通过评估多跳路径的得分，模型能够发现最相关的诊断路径，从而提高诊断准确率。

引入多种注意力机制用于路径进行推理：文章中提出了多头注意力（MultiAttn）和三线性注意力（TriAttn）机制，用于计算输入文本与路径之间的逻辑关系。并且引入 DR.KNOWS 作为基于知识路径的提示提供给大语言模型。这些机制能够增强模型在处理复杂医疗文本时的推理能力，在输入概念和上下文时能够生成更加准确的诊断预测。

在实际医疗数据集上的应用与验证： 该模型在MIMIC-III和IN-HOUSE两个真实世界的医院数据集上进行了评估，尤其是IN-HOUSE的私有数据集，显示出比基线模型更高的诊断预测性能。这表明该方法在实际医疗场景中具有广泛的应用潜力，尤其是在帮助医生提高诊断准确性和解释性方面。

通过阅读这篇文章，我认识到将医疗领域中的KG与LLMs结合在一起来解决医疗诊断问题的巨大潜力。这篇文章不仅展示当前技术在医学领域的实际应用，还创新模型设计，提供了一种新的思路：如何运用LLMs在繁杂、模糊的医疗数据中提取出有用信息，从而帮助医生做出更准确、更高效的诊断。

KG与LLMs的结合是本文最具创新性的部分之一。在阅读过程中，我逐渐了解到，LLMs虽然在NLP领域表现出色，但在处理医学诊断这类高复杂度专业性任务时往往存在一定的局限性。医疗记录中充满了大量的专有名词和复杂的病理关系，单靠LLMs难以准确地捕捉和推理出这些关系。通过引入UMLS等医疗KG将这些复杂的语义和病理关系融入诊断推理过程，从而极大提升了诊断的精准度和可靠性。

DR.KNOWS模型中路径选择与多跳推理机制的设计也让我印象深刻。传统的诊断过程通常依赖医生的临床经验和知识积累，但也会出现“经验主义失误”等情况。本文提出的**UMLS知识图谱多跳推理模型**，通过KG中不同节点的语义关联，自动生成排序潜在的诊断路径，模拟医生在临床推理中逐步排除和验证的过程。这种设计增强了模型的可解释性，为自动诊断生成提供了全面的推理过程，提高了模型的诊断效率和准确度。

通过这篇文章，我对如何将人工智能技术应用于医疗领域有了更深刻的理解，也激发了我对相关领域进一步学习和探索的兴趣。借此机会，我也阅读了很多类似的文章，从中找到了一部分思维范式，这对以后我走向计算机+生物医学方向有了一定的知识铺垫，也为后面的实验3、实验4打下了一定的理论基础，提升了知识图谱与医疗层面领域的认知水平。

北京交通大学
BEIJING JIAOTONG UNIVERSITY

文献报告

Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction[1]

将医疗知识图谱融入大语言模型以进行诊断预测

张鳊沣 22281052

2024年10月12日

2024–2025学年 《知识表示与处理》 实验2

[1] Gao Y, Li R, Caskey J, Dligach D, Miller T, Churpek MM, Afshar M. Leveraging a medical knowledge graph into large language models for diagnosis prediction[J]//arXiv preprint arXiv:2308.14321. 2023.