**1 介绍**

深度学习模型在机器阅读理解方面取得了重 大进展，甚至在包括SQUAD在内的单段问答 （QA）基准测试中表现优于人类然而，为了跨越机器与人之间阅读理 解能力的鸿沟，未来面临三大挑战：

**1）推理能力。**

段QA的模型倾向于在与问题匹配的句子中寻找答案，这不涉及复杂的推理。

因此，多跳QA成为下一个征服的前沿.

**2）可解释性。**

显式推理路径可以验证逻辑严谨性，对QA系统的可靠性至关重要。HotpotQA

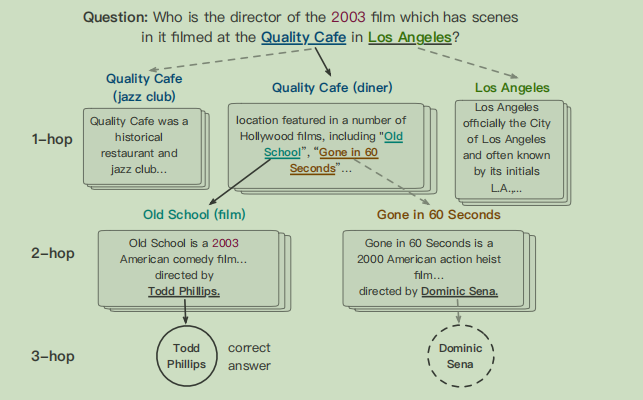


图1：多跳QA的认知图示例。

每个节点对应于实 体（例如，“洛杉矶”），其后是其介绍段落。 圆圈表示ans节点，回答问题的候选者。认知图模仿人类推理过程。在将实体称为“思维”时构建边缘。纯黑色边缘是正确的推理路径。 提供支持句子的模型，这意味着无序和句子级别的可解释性，然而人类可以通过逐步解决方案来解释答案，表明有序和实体级别的可解释性。

**3）可扩展性。**

对于任何实用的QA系统，可扩展性是必不可少的。基于机器理解的现有QA系统通常遵循DrQA中的检索 - 提取框架（陈等人。, 2017），通过预检索将源的范围缩小到几个段落。这个框架是单段QA和可扩展信息检索之间的简单折衷，与人类利用大容量存储器中的知识进行推理的能力相比（王等人。, 2003).

因此，可以从人类的认知过程中获得对这些挑战的解决方案的见解。双重过程理论（埃文斯, 1984, 2003, 2008; 斯洛曼, 1996）建议我们的大脑首先检索相关信息如下通过称为系统1的隐含的，无意识的和直观的过程引起注意，基于该过程，然后进行另一个明确的，有意识的和可控制的推理过程，系统2。系统1可以根据请求提供资源，而系统2可以通过在工作记忆中执行顺序思考来深入探讨关系信息，这种思维较慢但具有人类独特的合理性（巴德利, 1992).对于复杂的推理，两个系统协调进行快速和慢速思考（卡恩- 男人和伊根, 2011）迭代地。

在本文中，我们提出了一个框架，即认知图QA（CogQA），有助于解决上述所有挑战。受双重过程理论的启发，该框架包含功能不同的系统1和2模块。系统1从段落中提取与问题相关的实体并回答候选，并对其语义信息进行编码。提取的实体被组织为认知图（图1），类似于工作记忆。系统2然后在图上进行推理过程，并收集clues以指导系统1更好地提取下一跳实体。迭代上述过程直到找到所有可能的答案，然后根据系统2的推理结果选择最终答案。基于BERT的有效实现（Devlin等 人 。 , 2018 ） 和 图 神 经 网 络 （ GNN ）（Battagliaet al., 2018）介绍。

我们的贡献如下：

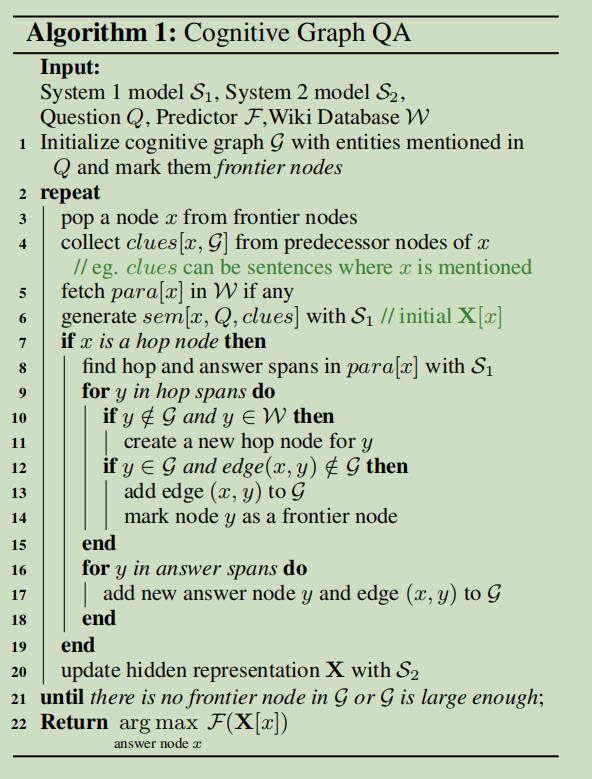
我们根据人类认知提出了大规模CogQA框架，用于大规模多跳阅读理解QA。

我们证明们框架中的认知图结构提供了有序和实体级可解释性，适合于关系推理。

我们基于BERT和GNN的实现在所有指标上都超过了之前的作品和其他竞争对手。

**2 认知图QA框架**

人类的推理能力主要取决于信息的关系结构。直观地，我们采用有向图结构进行多跳QA认 知过程的逐步演绎和探索。在我们的阅读理解设置中，该认知图中的每个节点对应于实体或可能的答案x，也可互换地表示为节点x。



提取模块系统1，读取实体x的引言段落[x]，并从段落中提取答案候选者和有用的下一跳

实体。然后使用这些新节点进行扩展，为推理模块系2提供显式结构。在本文中，我们

假设系统2通过计算节点的隐藏表示X来进行基于深度学习而不是基于规则的推理。因此，

系统1还需要在提取跨度时将para [x]概括为语义向量作为初始隐藏表示。然后，系统2基于图结构更新X作为下游预测的推理结果。

由于认知图中的显式推理路径，可以理解可解释性。除了简单的路径之外，认知图还可以清晰地显示联合或循环推理过程，其中新的前辈可能会为答案带来新的clues。我们框架中的clues是一种形式灵活的概念，指的是前人提供的信息，用于指导系统1更好地提取跨度。除了新添加的节点之外，由于新clues，具有新传入边缘的节点也需要重新访问。我们将它们都称为边界节点。可伸缩性意味着QA的时间消耗不会随着段落的数量而显着增长。我们的框架可以扩展

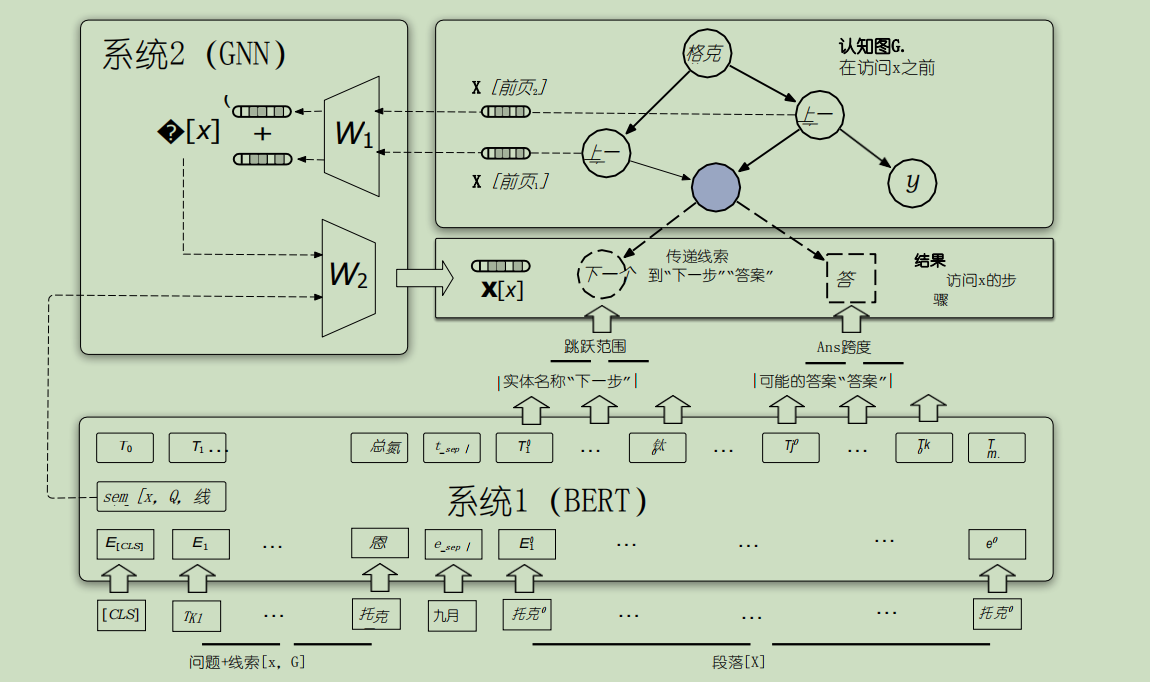
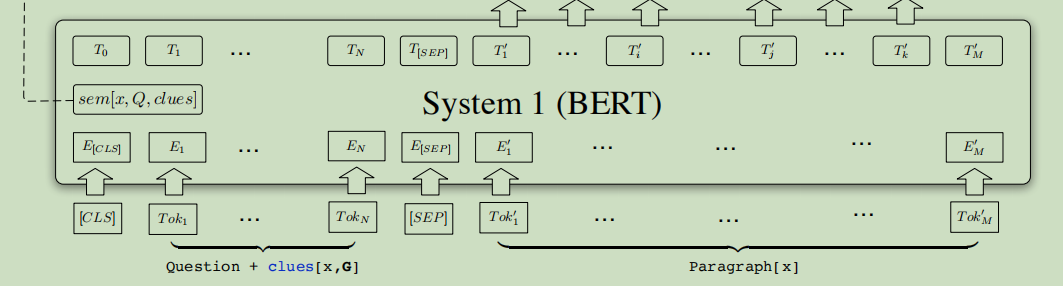


图2：CogQA实现概述。当访问节点x时，系统1基于系统2发现的clues[x，]生成新的跳跃和应答节点。它还创建初始表示sem [x，Q，clues]，基于其在系统2中的GNN更新隐藏的表示X [x]。

在本质上，因为所有段落的唯一操作是通过其标题索引访问某些特定段落。对于多跳问

题，传统的检索 - 提取框架可能会牺牲后续模型的潜力，因为距离问题多跳的段落可能与问题共享很少的常用词和很少的语义关系，导致检索失败。但是，这些段落可以通过我们的框架中的clues迭代扩展来发现。算法1 描述了我们的框架CogQA的过程。初始化之后，开始用于图形扩展和推理的迭代过程。在每个步骤中，我们访问前沿节点x，系统1在clues和问题Q的指导下读取para [x]，提取跨度并生成语义向量sem [x，Q，clues]。同时，系统2更新隐藏的表示X并为任何后继节点y准备clues[y，g]。最终预测基于X.

**3 实现**

实现CogQA框架的主要部分是确定系统1和2的具体模型，以及clues的形式。

我们的实现使用BERT作为系统1，GNN作为系统2.同时，clues[x，G]是x的上一节点的段落中的句子，从中提取x。我们直接将原始句子作为clues传递，而不是任何形式的计算隐藏状态，以便于系统1的训练。因为原始句子是独立的并且独立于先前迭代步骤的计算，所以在不同的迭代步骤的训练被解耦，导致在训练期间提高效率。细节介绍于3.4.每次通过GNN的传播步骤更新图节点的隐藏表示X.我们的整体模型如图所示2.

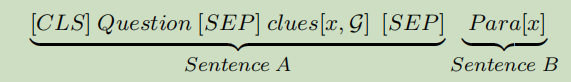
**3.1 系统1**

System 1模型的提取能力是构建认知图的基础，因此需要一个强大的模型。最近，BERT

（去 vlinet al., 2018）已成为各种NLP任务中最成功的语言表示模型之一，包括 SQuAD

（Rajpurkar等人。, 2016).BERT由多层变压器组成（Vaswaniet al., 2017），一个基于自我关注的架构，并在大型语料库上精心预先训练。输入句子由两个不同的功能部分A和B组成。我们使用BERT作为系统1，并在其输入时使用

访问节点x如下：



其中clues[x，g]是从上一节点传递的句子。

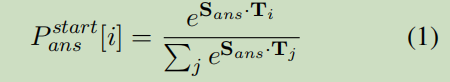
BERT的输出向量定义为，

其中L是输入序列的长度，H是隐藏表示的维度大小。值得注意的是，对于答案节点x，可能缺少P ara [x]。因此，我们不提取跨度，但仍然可以根据“句子A”部分计算sem [x，Q，clues]。当从问题中提取单跳节点进行初始化时，我们不计算语义向量，并且输入中只存在问题部分。跨度提取答案和下一跃点实体具有不同的属性。答案提取在很大程度上依赖于问题所指示的字符。例如，“纽约市”更可能是问题的答案而不是“2019”，而下一跳实体通常是其描述与问题中的陈述匹配的实体。

因此，我们分别预测答案跨度和下一跃点跨度。我们引入“指针向量”Shop, Ehop, Sans, Eans

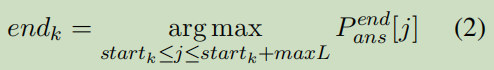
作为预测目标跨度的额外可学习参数。第i个输入关键字作为答案跨度

概率计算如下：



我们将剩余的预测答案跨度扩展为认知图，作为新的“答案节点”。遵循相同的过程以通过用**S***hop,* **E***hop* 来替换**S***ans,* **E***ans*  来扩展“下一跳节点”。

设 是第i个输入关键字成为答案跨度结束的概率，可以按照相同的公式计算。 我们只关注顶部K开始概率{}的位置。对于每个k，结束位置由以下给出：



其中，maxL 是最大可能的跨度长度。为了识别不相关的段落，我们利用§3.4.1中引入的负抽样来训练系统1以产生负阈值。在顶部K跨中，那些启动概率小于负阈值的将被丢弃。 因为第0个关键字[CLS]是预先训练的，用于合成下一个句子预测任务的所有输入关键字。作为我们实现中的阈值。

我们扩展了认知图，剩余的预测答案跨度为新的“答案节点”。 通过用Shop、Ehop替换Sans、Eans来扩展“下一跳节点”，遵循同样的过程。

语义生成如上所述，位置0的BERT输出具有汇总序列的能力。因此，最直接的方法是使用T0 作为sem [x，Q，clues]。然而，BERT中的最后几层主要负责转换跨度预测的隐藏表示。在我们的实验中，位置0处第三到最后一层输出的使用为sem [x，Q，clues]表现最佳。

**3.2 系统2**

系统2的第一个功能是为边界节点准备clues[x，g]，我们将其实现为收集提到

x的x的上一节点的原始句子。

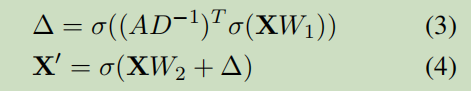
第二个函数，用于更新隐藏的表示X，是系统2的核心功能。隐藏的表示代表对所有n个实体的理解。为了完全理解实体x和问题Q之间的关系，仅仅分析语义

sem [x，Q，clues]是不够的。GNN已被提议在图形上进行

深度学习(Kipf and Welling, 2017)，特别是由于图结构的归纳偏差而产生的关系推理（Battaglia et al., 2018).

在我们的实现中，GNN的一个变体被设计成系统2。对于每个节点x，

初始隐藏表示是 系统1中的语义向量sem [x，Q，clues]。 让X’ 成为GNN传播步骤后的新隐藏表示，并且 ∆是从传播中的邻居传递的聚合向量。 X的更新公式如下：



其中σ是激活函数，W1，W2 是权重矩阵。A是G的相邻矩阵，其被列标准化为，此时。变换隐藏向量左乘以，这可以用Defferrard et al. (2016) 局域光谱滤波器 来解释。

在访问前沿节点x的迭代步骤中，其隐藏表示**X**[x]按照方程（3)(4)更新。 在实验中，我们观察到这种“异步更新”在G定型后通过多个步骤一起更新所有节点的X在性能上没有明显的差异，这是更有效的，并在实践中采用。

**3.3 预报器**

**HotpotQA**数据集中的问题一般分为三类：特殊问题、替代问题和一般问题，它们被视为三种不同的下游预测任务将**X**作为输入。 在测试集中，它们也可以很容易地根据疑问词进行分类。

特殊问题是最常见的情况，要求在段落中找到跨度，如地点、日期或实体名称。 我们使用两层全连接网络(**FCN**)作为预测器：



备选问题和一般问题都旨在比较HotpotQA中实体x和y的某一性质，分别用实体名称和“是或否”回答”。 这些问题被认为是二元的 用输入**X**[x]-**X**[y]分类，并由另两个相同的**FCNs**求解。

**3.4 训练**

我们的模型是在负抽样的监督范式下训练的。 在训练集中，下一跳和回答跨度是预先提取的段落。 更准确地说，对于与问题Q相关的每一个para[x]，我们有跨越数据



其中，para[x]中从到的跨度与实体或答案的名称模糊匹配。 详见第4.1节。

**3.4.1任务＃1：跨度提取**

的正确标注是基于D[x，Q]构造的。 每个段落最多有一个答案跨度(y，start，end), 因此，是一个one-hot向量，其中[start]=1。 但是多个不同的下一跳跨度可能出现在一段中，因此[]=1/k，其中k是下一跳跨度的数目.

为了能够区分不相关的段落，将不相关的负跳节点提前添加到*G*中。 如第3.1节所述， [CLS]，T0的输出负责产生负阈值。 因此，每个负跳节点negative hop node的是[0]=1的one-hot向量。

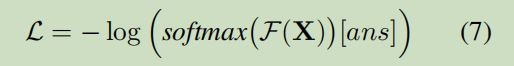
利用交叉熵损失来训练系统1中的跨度提取任务。 结束位置和下一跳跨度的损失定义如下。



**3.4.2 任务＃2：应答节点预测**

要掌握推理能力，我们的模型必须学会从认知图中识别正确的答案节点。 对于训练集中的每个问题，我们为这个任务构造一个训练样本。 每个训练样本都是黄金图的组成，黄金图是所有正确推理路径和负节点的结合。 负节点包括任务#1和两个n中使用的负跳点节点主动回答节点。 从随机选择的跳节点中随机提取的跨度构造负答案节点。

对于特殊问题，我们首先通过对*F*的输出执行Softmax来计算每个节点的最终答案概率。损失*L* 定义为概率与单热之间的交叉熵 答案节点ans的向量。



备选问题和一般问题用二元交叉熵以类似的方式进行优化。 该任务的损失不仅被反向传播，以优化预测器和系统2，而且通过系统1语义向量*sem*[*x, Q, clues*]进行微调.

**4 实验**

**4.1 数据集**

我们使用HotpotQA的全维基设置来进行我们的实验。 112，779个问题由众包根据维基百科文件中的第一段收集，其中84%需要多个段落- 跳跃推理。 数据被分成训练集（90,564个问题)、开发集(7,405个问题)和测试集(7,405个问题）。 开发和测试中的所有问题都是困难的 我跳的案子。但是，一个段落中可能出现多个不同的下一跳跨度，gt开始 [starti] = 1 / k其中k是下一跳跨度的数量。为了区分不相关段落的能力，将不相关的负跳节点预先添加到G.如§中所述3.1,维基百科文档中的图表，其中84％需要多跳推理。数据分为训练集（90,564个问题），开发集（7,405个问题）和测试集（7,405个问题）。开发和测试集中的所有问题都是hard multi-hop案例。

在训练集中，对于每个问题，提供了2个金的（有用）实体的答案和段落，其中包含多个支持事实，包含推理关键信息的句子，标记出。还有8个无用的负面段落用于训练。在评

估期间，只提供问题，同时除了答案之外还需要支持事实。

为了构建训练的认知图，基于Levenshtein距离的模糊匹配从支持事实推断出纯金认知图中的边缘(Navarro, 2001).对于para [x]中的每个支持事实，如果任何黄金实体或答案（表示为y）与支持事实中的跨度模糊匹配，则添加边（x，y）。

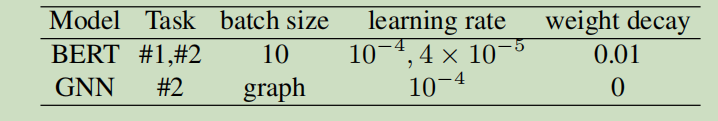
**4.2 实验细节**

我们使用预先训练的BERT基础模型(Devlin et al., 2018)

在系统1中，隐藏大小H为768，在 GNN和预测变量的节点向量中保持不变。我们

模型中的所有激活函数都是gelu(Hendrycks and Gimpel, 2016)

我们在任务＃1上训练模型1个epoch，然后在任务＃1和＃2上共同训练1个epoch。训练中的超参数如下：

 BERT和GNN由两个不同的Adam优化器优化，其中β1 = 0.9，β2 = 0.999。预测变量与GNN共享相同的优化器。在前10％步骤中BERT预热参数的学习率，然后线性衰减到零。

为了选择支持事实，我们只是将图中任何节点clues中的句子视为支持事实。在初始化中，这些单跳跨度存在于问题中，并且还可以通过与训练集中的支持事实的模糊匹配来检测。我们的框架提取的1跳实体可以改善其他模型的检索阶段（参见4.3），这促使我们将单跳实体的提取分离到另一个基BERT的模型，以便在实现中重用。

**4.3 基础**

**第一类是以前的工作或竞争对手：**

杨等人。 (2018）HotpotQA原始论文中提出的强基线模型（杨等人。, 2018).它遵循DrQA的检索提取框架（2017并且包含QA中的高级技术，例如自我关注，角色级别模型，双重关注。

GRN，QFE，DecompRC，MultiQA排行榜上 的其他模型。

BERT最先进的单跳QA模型。原始论文中的 BERT需要单段输入，预训练的BERT几乎不 能处理最多512个关键字的段落，远小于连 接段落的平均长度。我们将认知图中的上一节点的相关句子添加到每个段落，并在所有段落中以最大起始概率报告答案跨度。

杨等人。 (2018）-IR：杨等人。 (2018）改进的检索。杨等人。(2018）使用传统的倒排索引过滤策略来检索相关段落。有效性可能会受到挑战，因为它有时无法找到有关的实体。主要原因是检索中的词级匹配通常忽略语言模型，这表明词语的重要性和POS。我们通过在模型中添加问题中的1-hop实体来改进检索，将支持事实的覆盖范围从56％增加到72％。

**另一类是消融研究：**

CogQA-onlyR模型使用检索到的相同实体 进行初始化杨等人。(2018）作为单跳实体，主要用于公平比较。

CogQA-onlyQ仅使用1跳进行初始化从问题中提取的实体，没有检索到的段落。完整的CogQA实现使用两者。

CogQA-sys1仅保留系统1，并且在系统2中缺少级联推理。

**4.4 结果**

以下杨等人。 (2018），答案和支持事实的评估包括两个指标：完全匹配（EM）和F1 得分。

如果答案字符串和支持事实都严格正确，则联合EM仅为1。联合精度和召回是Ans和Sup的产品，然后计算联合F1 。这些指标的所有结果都在测试集上取平均值。4 实验结果表明我们的方法在多个方面具有优势：

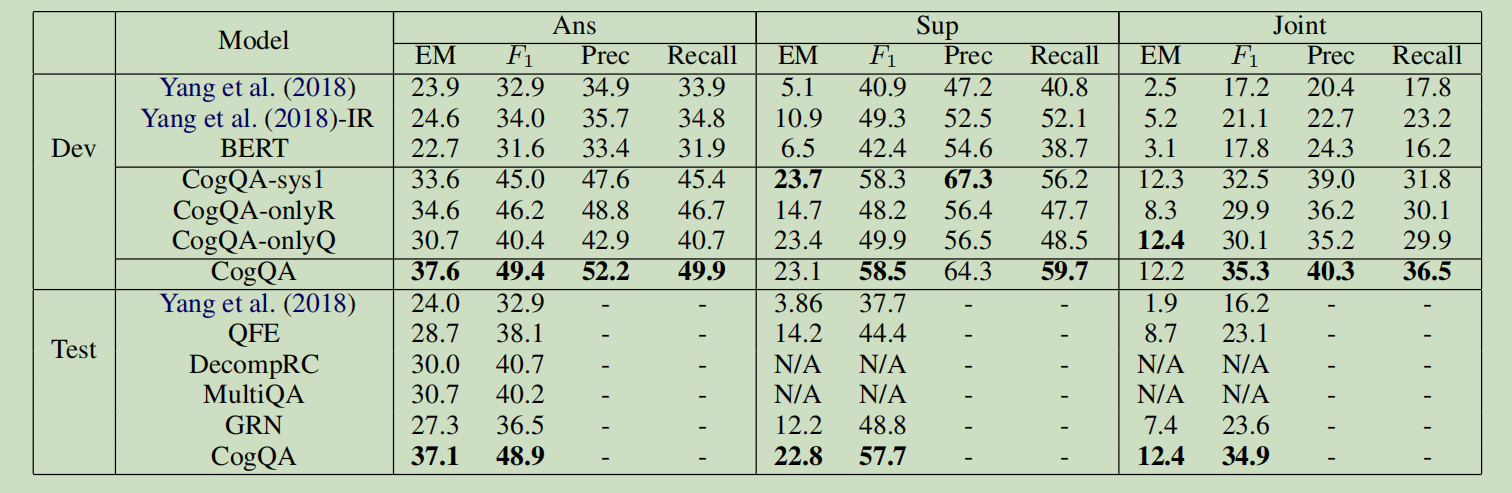


表1：HotpotQA的结果

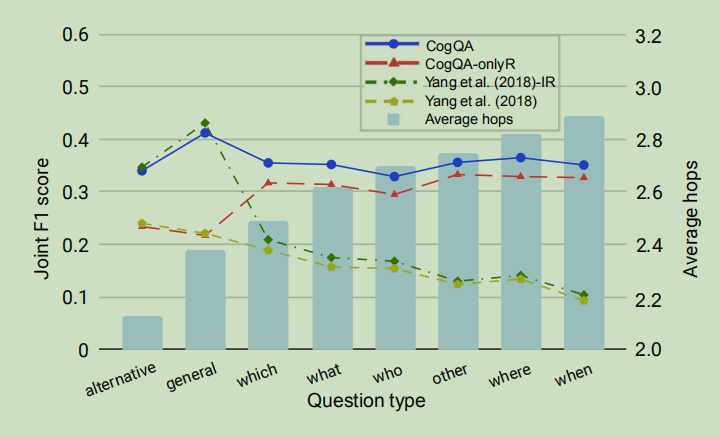
HotpotQA的维护者只为每次提交提供EM和F1 。

N / A表示模型找不到支持事实。

整体表现我们的CogQA在所有指标上的表现均 优于所有指标（见表格）1).性能的飞跃主要源于CogQA框架优于传统检索提取方法的优越性。由于多跳的段落可能与字面上几乎没有共同的单词或甚至与问题的语义关系很少，因此检索 - 提取框架只有在找到与其相关的推理clues后才能找到相关的段落。然而，我们的框架逐渐发现相关实体的clues。

Logical Rigor QA系统经常被批评用浅模式匹配来回答问题，而不是基于推理。为了评估质量保证的逻辑严谨性，我们使用 J整数 正确答案中“联合正确答案”的比例。联合正确答案是从所有必要和正确的支持事实中推导出的答案。因此，这个比例代表了推理的逻辑严谨性。我们的方法比例高达33.4％，远远超过7.9％杨等人。(2018）和30.3％的QFE。

图3：8种问题的模型表现



**多跳推理图**3 说明联合F1 得分和8种问题的平均跳数，包括具有不同疑问词的一般问题，替代问题和特殊问题。 随着跳数的增加，性能也随之提高。我们的方法非常强大。但是，替代性和一般 性问题没有改善，因为判断的证据不能从支持事实中推断出来，导致缺乏监督。需要进一步的人工标记来回答这些问题。

消融研究为了研究初始实体在认知图中的影响，CogQA-onlyR从相同的初始段开始.我们发现CogQA-onlyR仍然表现得更好。与CogQA相比，性能略有下降，表明贡献主要来自框架。

为了与检索 - 提取框架进行比较，CogQAonlyQ被设计为仅从出现在问题中的实体开始。没有精心设计的检索方法，这种设置可以被视为人类的自然思维模式，其中只有明确可靠推理需要关系。CogQA-onlyQ仍然优于所有基线，这可能揭示了CogQA框架优于检索 - 提取框架的优势。

BERT不是改善的关键因素，尽管它起着必要的作用。香草BERT表现相似甚至稍差在这个多跳QA任务中，可能是因为有针对性的架构，更好地利用监督 支持事实。

为了研究缺少系统2的影响，我们设计了一个仅限系统1的方法，

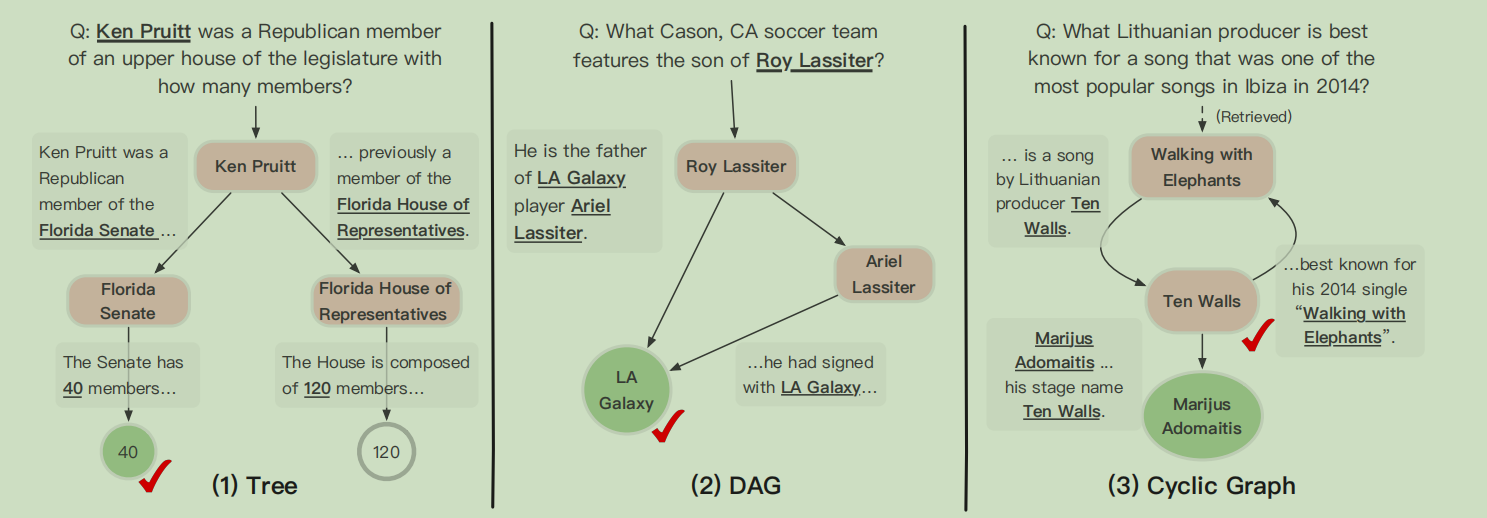


图4：案例研究。我们的结果中有不同形式的认知图，即树，有向无环图（DAG），循环图。圆是候选答案节点，而圆角矩形是跳节点。绿色圆圈是CogQA给出的最终答案，复选标记代表注释的基本事实。

CogQA-sys1，它继承迭代框架但输出具有最大预测概率的答案跨度。在Ans指标上，对最佳竞争对手的改进减少了约50％，突出了GNN在认知图上的推理能力。

**案例研究** 我们展示了认知图如何在图中的实验中清楚地解释复杂的推理过程4.认知图突出了案例（1）中问题的核心 - 即，在两个房屋中的成员数量之间进行选择。CogQA根据“参议院”和“上议院”之间的语义相似性做出了正确的选择。情况（2）说明了通过探索并行推理路径可以提高答案的稳健性。案例（3）是一个语义检索问题，没有提到任何实体，这对于CogQA-onlyQ甚至是人类都是难以处理的。一旦与信息检索相结合，我们的模型最终得到了答案“Marijus Ado-maitis”，而注释的基本事实是“十墙”。然而，当在认知图中回溯推理过程时，我们发现该模型已经达到“十墙”并以他的真实姓名回答， 这是可接受的，甚至更准确。黑盒模型不具 备这种可解释的优点。

**6 讨论和结论**

我们提出了一个新的框架CogQA来大规模解决多 跳机读取问题。推理过程被组织为认知图，达 到了前所未有的实体级可解释性。我们基于 BERT和GNN的实现在HotpotQA数据集上获得了最先进的结果，