# 利用知识库和文本的早期融合来开放域问题回答

海天太阳* Bhuwan Dhingra* Zaheer满足乐 凯瑟琳mazaitis
鲁斯兰salakhutdinov 威廉 科恩
卡内基梅隆大学计算机科
学学院
{海地人，bdhingra，manzilz，krivard，rsalakhu，wcohen} @ cs.cmu.edu



## 摘要

开放域问答（QA）正在从复杂的流水线系统发展到端到端的深度神经网络。 已经开发了专门的神经模型，用于单独从文本或知识库（KB）中提取答案。 在本文中，我们将看一个更实用的设置，即关于KB和实体链接文本组合的QA，当具有大文本语料库的不完整KB时，这是适当的。 基于图表示学习的最新进展，我们提出了一种新的模型GRAFT-Net，用于从包含文本和KB实体和关系的特定于问题的子图中提取答案。我们为此问题构建了一套基准测试任务，改变了问题的难度，培训数据量和KB完整性。 我们表明，当使用KB或单独的文本进行测试时，GRAFT-Net与最先进的设备相比具有竞争力，并且在组合设置中大大优于现有方法。

## 1 介绍

开放域问答（QA）是查找自然语言提出的问题答案的任务。 从历史上看，这需要一个由多个机器学习和手工制作模块组成的专用管道（Ferrucci等人。，2010）。 最近，范式已转向培训端到端深度神经网络模型的任务（陈等人。，2017；梁等人。，2017；Raison等。，2018； Talmor 和 Berant， 2018； Iyyer 等 。 ，2017）。 然而，大多数现有模型使用单个信息源来回答问题，通常是来自百科全书的文本或单个知识库（KB）。

直观地说，QA信息源的适用性取决于其覆盖范围和从中提取答案的难度。 大文本语料库具有高覆盖率，但是使用许多不同的文本模式来表达信息。 因此，运行这些模式的模型（例如BiDAF（Seo等人。，2017））不要超出他们的训练领域（Wiese等。，2017；Dhingra等。，2018）或新颖的推理类型（Welbl等人。，2018；Talmor和Berant，2018）。 另一方面，KB由于其不可避免的不完整性和受限制的架构而遭受低覆盖率（敏等人。，2013），但更容易从中提取答案，因为它们是为了被查询而精确构造的。

在实践中，一些问题最好使用文本来回答，而其他问题最好使用KB来回答。 因此，一个自然的问题是如何有效地结合两种类型的信息。 令人惊讶的是，之前很少有人研究过这个问题。 在本文中，我们关注一个大规模KB的场景（Bollacker等。，2008；奥尔等人。，2007）和文本语料库可用，但既不足以回答所有问题。

在这种情况下，一个简单的选择是采用为每个来源开发的最先进的QA系统，并使用一些启发式聚合他们的预测（Ferrucci等人。，2010；博迪，2015）。 我们将这种方法称为晚期融合，并表明它可能是次优的，因为模型在不同来源之间汇总证据的能力有限（ 5.4）。相反，我们专注于早期融合策略，其中训练单个模型以从问题子图中提取答案（参见图2）1，左）包含相关的KB事实和文本句子。 早期融合允许更灵活地组合来自多个来源的信息。

为了实现早期融合，在本文中，我们提出了一种新的基于图卷积的神经网络，称为GRAFT-Net（事实和文本网络之间的关系图），专门设计用于在KB事实的异构图上操作

*Haitian Sun和Bhuwan Dhingra对这项工作做出了同样的贡献。

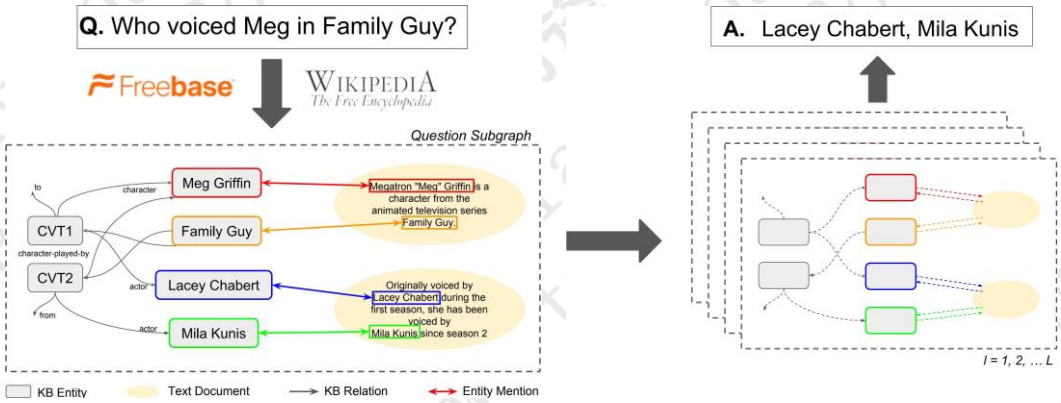



图1：左：为了回答用自然语言提出的问题，GRAFT-Net考虑了从文本和KB事实构建的异构图，因此可以利用两个信息源之间丰富的关系结构。 右：嵌入在图表中传播固定数量的图层（L）并且最终节点表示用于对答案进行分类。

和文字句子。 我们以图表表示学习的最新工作为基础（基普夫和威灵，2016; Schlichtkrull等。，2017），但提出了两个关键的修改，以采用它们来完成QA的任务。首先，我们提出异构更新规则来处理KB节点与文本节点的不同之处：例如，基于LSTM的更新用于将信息传入和传出文本节点（ 3.2）。其次，我们引入了一种定向传播方法，其灵感来自IR中的个性化Pagerank（Haveliwala, 2002），它限制嵌入在图中的传播，以遵循从与问题相关联的种子节点开始的路径（ 3.3）。 根据经验，我们证明这些扩展对QA的任务至关重要。 该模型的概述如图所示1.

我们在一组新的基准测试任务中评估这些方法，以便在存在KB和文本时测试QA模型。 使用WikiMovies （米勒等人。，2016）和WebQuestionsSP（弘毅等。，2016），我们构建具有不同数量的训练监督和KB完整性的数据集，并且具有不同程度的问题复杂性。 我们报告基线以供将来比较，包括关键值存储器网络（磨坊主等。，2016; 达斯等人。，2017c），并表明我们提出的GRAFT-Nets在各种条件下都具有卓越的性能（ 5）。 我们还表明，GRAFT-Nets与专门针对纯文本QA开发的最先进方法相比具有竞争力，并且最先进的方法发展

为仅KB的QA开启（§ 5.4）[1].

[1]源代码和数据可从以下位置获得 https://开头oceanskysun graftnet github.com / /

## 2 任务设置

### 2.1 描述

知识库表示为K = (V, E, R)，其中V是KB中的实体集，并且边E是三元组（s，r，o），其表示关系r∈R保持在主题s∈V和对象o∈V。文本语料库D是一组文档{d1, . . . , d|D|}，其中每个文档是单词序列di = (w1, ..., w|di|)。 我们进一步假设一个（不完美的）实体链接系统

tem已经运行在文档集合上，其输出是一组链接（v，dp）将实体v与文档d中位置p处的单词连接起来，我们用Ed表示所有实体的集合文件中的链接d。 对于在d中跨越多个单词的实体提及，我们包括指向提及中的所有单词的链接。

鉴于自然语言问题，任务是q = (w1, . . . , w|q|)，从中提取答案a ⊆ V。 可能有多个正确的答案，因此我们将其转向一个问题。 在本文中，我们假设答案是来自文档或KB的实体。 我们对各种设置感兴趣，其中KB从高度不完整到完整用于回答问题，我们将介绍用于在这些设置下测试我们的模型的数据集。

为了解决这个任务，我们分两步进行。 首先，我们提取一个子图Gq = (Vq, Eq)，其中包含高概率问题的答案。 此步骤的目标是确保对答案的高度回忆，同时生成足够小的图形以适合基于梯度的学习的GPU内存。 接下来，我们使用我们提出的模型GRAFT-Net来

学习$_q$中的节点表示，以q为条件，用于将每个节点分类为答案。 使用远程监督生成第二步的训练数据。 整个过程模仿基于文本的QA的搜索和阅读范例（Dhingra等。，2017）.

### 2.2 问题子图检索

我们使用两个并行管道检索子图$_q$ - 一个在KB上返回一组实体，另一个在语料库上返回一组文档。 然后将检索到的实体和文档与实体链接组合以产生完全连接的图。

KB检索。 为了从KB检索相关实体，我们首先在问题q上执行实体链接，产生一组种子实体，表示为S$_q$。 接下来我们运行个性化PageRank（PPR）方法（Haveliwala, 2002围绕这些种子来识别可能是问题答案的其他实体。 S$_q$周围的边缘权重在相同类型的所有边缘之间均匀分布，并且它们被加权使得与问题相关的边缘比不具有的边缘具有更高的权重。 具体地，我们平均单词向量以从关系的表面形式计算关系向量v（r），并且从问题中的单词计算问题向量v（q），并使用它们之间的余弦相似度作为边缘权重。 运行PPR后，我们保留了前E个实体v$_1$,. . . ，v$_E$按PPR得分，以及任何边缘 -

补间它们，并将它们添加到G$_q$。

文本检索。 我们使用维基百科作为语料库并在句子级别检索文本，即D中的文档是沿着句子界定的 -
白羊座[2]。 我们分两步执行文本检索：首先我们检索前5个最相关的维基百科

文章，使用DrQA的加权词袋模型（陈等人。，2017）； 然后我们填充一个Lucene[3] 索引与这些文章中的句子，并检索排名最高的d$_1$,. . . ，d$_D$，基于问题中的单词。 对于句子检索步骤，我们发现将文章标题作为Lucene索引中的附加字段包含在内是有益的。由于文章中的大多数句子都涉及标题实体，这有助于检索未明确提及问题中的实体的相关句子。 我们添加检索到的

[2]术语文档将始终引用本文其余部分的句子。

[3]lucene.apache.org https: / / /

文件，连同与其相关的任何实体，到子图G$_q$。

最后一个问题子图是G$_q$= （V$_q$, E$_q$, R$^+$），其中顶点V$_q$由所有检索到的实体和文档组成，即q = v$_1$,. . . ，v$_E$d$_1$,. . . ，d$_D$。 边缘是这些实体之间的所有关系，加上文档和实体之间的实体链接，即

$$E_q = \{(s, o, r) \in E : s, o \in V_q, r \in R\} \cup \{(v, d_p, r_L) : (v, d_p) \in L_d, d \in V_q\},$$

其中r$_L$表示特殊的“链接”关系。

R$^+$ = R ∪ {r$_L$}是子图中所有边类型的集合。

## 3 移植物网

问题q及其答案a$_q$诱导G$_q$中节点的标记：如果va$_q$ ∈ V$_q$和y$_v$= 1否则y$_v$= 0则我们让y$_v$= 1对于所有v ∈ {0, 1}。 QA的任务然后减少到在图的节点上执行二进制分类。 在文献中已经提出了几种基于图传播的模型，其学习节点表示然后执行节点的分类（基普夫和威灵，2016; Schlichtkrull等。，2017）。 这些模型遵循标准的聚集 - 应用 - 散布范式来学习具有同构更新的节点表示，即平等地处理所有邻居。

这些模型的基本配方如下：

1. 初始化节点表示h$^{(0)}_v$。
2. 对于l = 1,. . . ，L更新节点表示

$$h^{(l)}_v = \phi\left(h^{(l-1)}_v, \sum_{v^i \in N_r(v)} h^{(l-1)}_{v^i}\right)$$

其中N$_r$（v）表示沿着类型r的入射边缘的v的邻居，并且φ是神经网络层。

这里L是模型中的层数，并且对应于信息应该在图中传播的路径的最大长度。 一旦传播完成，最后的层表示h$^{(L)}_v$用于执行期望的任务，例如知识库中的链接预测（Schlichtkrull等。，2017）.

但是，我们的设置与之前研究的基于图形的分类任务有两点不同。 第一个区别是，在

在我们的例子中，图$G_q$由异构节点组成。 图中的一些节点对应于表示符号对象的KB实体，而其他节点表示作为可变长度的单词序列的文本文档。 第二个区别是我们想要在自然语言问题q上调节图中节点的表示。 在 3.2 我们引入异构更新来解决第一个差异，并在 3.3 我们引入了对第二个问题（及其实体）进行条件化的机制。

### 3.1 节点初始化

使用固定大小的向量$h^{(0)} = x_v \in R^n$初始化对应于实体的节点，其中$x_v$可以是预训练的KB嵌入或随机的，并且n是嵌入大小。 图中的文档节点描述了可变长度的文本序列。 由于多个实体可能链接到文档中的不同位置，因此我们在每个实体中维护文档的可变长度表示层。 这由$H^{(1)} \in R^{|d| \times n}$表示。 鉴于文件中的文字$(w_1, \ldots, w_{|d|})$，我们初始

将其隐藏的表示形式称为：

$$H_d^{(0)} = LSTM(w_1, w_2, \ldots),$$

其中LSTM指的是长期短期记忆单位。 我们表示第p行$H^{(1)}$，对应于文档中第p个字的嵌入

$d$在层1，如$H^{(1)}_d$。

### 3.2 异构更新

数字 2 显示实体和的更新规则
文件，我们在这里详细描述。

实体。 设$M(v) = \{(d, p)\}$是文件d中的位置p的集合，其对应于实体v的提及。实体节点的更新在 -

推动单层前馈网络（FFN）
在四个州的连接中：

$$h_v^{(l)} = \begin{cases} \text{FFN}\left(h_v^{(l-1)}, h_q^{(l-1)}\right), & \text{如果 } v \in V \\ h_v^{(l-1)}, & \text{如果 } v \in Q \end{cases} \quad (1)$$

在下一节），并应用关系特定的变换$\psi_r$。 以前关于关系图卷积网络的工作（Schlichtkrull等。，2017）使用线性投影$\psi_r$。对于批量实现，这会产生大小为$O(B|E|n)$的矩阵，其中B是批量大小，对于大型子图，这可能非常大[4]。 因此，在这项工作中，我们使用$r_q$的关系向量$x_r \in R^n$而不是矩阵，并沿边缘计算更新：

$$\psi_r(h_{v_i}^{(l-1)}) = \text{pr}_{v_i}^{(l-1)} \text{FFN}(x, r, h_{v_i}^{(l-1)})。 \quad (2)$$

这里$\text{pr}_{v_i}^{(l-1)}$是一个PageRank分数，用于控制沿路径开始的嵌入传播 -
来自种子节点，我们将在下一节中详细介绍。上述存储器复杂度为$O(B(|E_q|+|V_q|)n)$，其中$G_q$是子图$G_q$中的事实数，$|F| \ll |E| \ll |V|$

最后一项聚合了子图中文档中与实体v的提及相对应的所有标记的状态。 请注意，

日期取决于他们的实体的位置
包含文件。

文档。 令$L(d, p)$是与文档d中位置p处的单词链接的所有实体的集合。 文档更新分两步进行。 首先，我们分别对每个位置的实体状态进行汇总：

$$\tilde{H}_d^{(l)} = \text{FFN}\left(H_d^{(l-1)}, h_v^{(l-1)}\right)。 \quad (3a)$$
$$v \in L$$

这里$h_v^{(l-1)}$由v处的输出边数归一化。接下来我们在其中聚合状态
使用LSTM的文件：

$$H_d^{(l)} = \text{LSTM}(\tilde{H}_d^{(l)})。 \quad (3b)$$

### 3.3 调整问题

对于到目前为止描述的部分，图形学习者在很大程度上不了解该问题。 介绍

对问题的依赖有两种方式：关注

过度关系和个性化传播。
为了表示q，设$w_1, \ldots, w_{|q|}$是问题中的单词。 初始表示是com-

前两个术语对应于实体表示和问题表示（详情如下 – 低），分别来自前一层。

第三个术语汇总了州的状态

当前节点的实体邻居$N_v$（v），af– $_i$

用注意力量$\alpha^r$进行缩放（描述 $_r$

作为：

$$h^{(0)} = LSTM\,(w^q, \ldots, w^q)_{|q|} \in R^n, \quad (4)$$

$q$ 1 $|q|$

[4]这是因为我们必须使用大小的邻接矩阵

/ Rq | × | Eq | × | Eq | 聚合来自邻居的嵌入

同时所有节点。

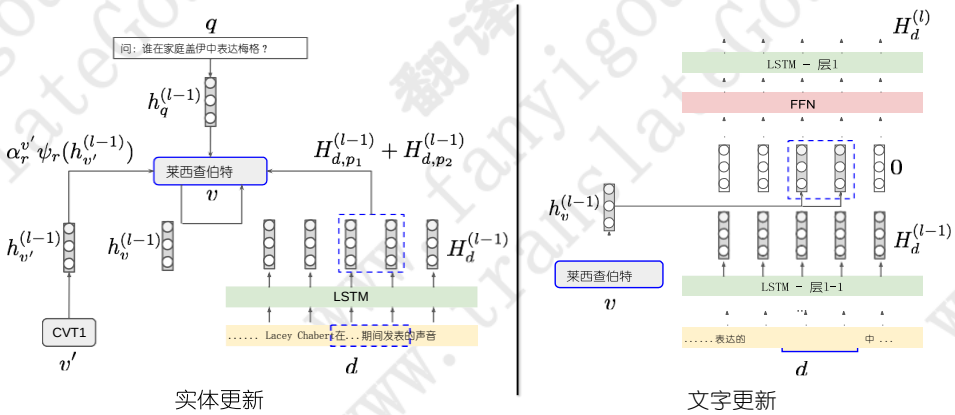
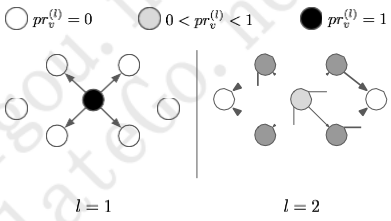



图2：实体（左）和文本文档（右）的异构更新规则示意图

我们从LSTM的输出中提取最终状态。 在随后的层中
问题表示更新为$h^{(1)}$ = (J, FFN $_{v\in S_q}$ $h^{(v)}$)，其中$S_q$ 表示种子




问题中提到的tities。

关注关系。 方程式第三项中的注意力。 (1)
使用问题和关系嵌入计算：

$$\alpha_r^{vi} = \text{softmax}(x_r^T h^{(l-1)}),$$

其中softmax归一化在$v^i$的所有输出边上，而$x_r$是关系r的关系向量。 这确保了嵌入是沿着与问题相关的边缘进行更多的调整。

定向传播。 许多问题需要多跳推理，其遵循从问题中提到的种子节点到目标应答节点的路径。 为了在传播嵌入时鼓励这种行为，我们开发了一种受IR个性化PageRank启发的技术（Haveliwala, 2002）。 传播开始于问题中提到的种子实体$S_q$。 除了节点处的矢量嵌入$h^{(l)}$之外，

我们还维护标量“PageRank”得分$pr_v$，它测量从种子实体到当前节点的路径总重量，如下所示：

$$r(v) = \begin{cases} \frac{1}{|S_q|} & \text{if } v\in S_q \\ 0 & \text{流} \end{cases},$$

$$pr^{(l)} = (1-\lambda)\, pr^{(l-1)} + \lambda \sum_{v^i\in N_r(v)} \alpha_r^{vi} pr^{(l-1)}.$$

请注意，我们重用了注意力量$\alpha^{vi}$ 在传播PageRank时，确保节点

图3：GRAFT-Net中嵌入的定向传播。 为跨越层的每个节点v维护标量PageRank得分$pr_v^{(l)}$，其从种子节点展开。 嵌入仅从$pr^{(l)}>0$的节点传播。

沿着与问题相关的路径获得高权重。 当在方程式中沿边缘传播嵌入时，PageRank分数用作缩放因子。 (2)。 对于l = 1，除种子实体之外的所有实体的PageRank分数将为0，因此传播将仅从这些节点向外发生。 对于l = 2，对于种子实体及其1跳邻居，它将不为零，并且仅沿着这些边缘发生传播。 数字 3 说明了这个过程。

### 3.4 答案选择

最终表示$h^{(L)}\in R^n$用于
二元分类来选择答案：

$$Pr(v\in\{a\}_q | G_q, q) = \sigma(w^T h_v^{(L)} + b), \quad (5)$$

其中σ是S形函数。 训练使用这些概率的二元交叉熵损失。

### 3.5 通过事实辍学进行正规化

为了鼓励模型学习一个利用所有可用信息源的强大分类器，我们在训练期间用概率$p_0$随机地从图中删除边。 我们称之为

其实差。 从KB中提取答案通常比从文档中提取答案更容易，因此模型往往依赖于前者，尤其是在KB完成时。 此方法类似于DropConnect（Wan等人。，2013).

## 4 相关工作

的工作 达斯等人。 (2017c）尝试针对KB事实和文本的QA的早期融合策略。 他们的方法基于键值存储网络（KV-MemNNs）（米勒等人。，2016）加上通用模式（Riedel等。，2013）独立地填充具有KB三元组和文本片段的表示的内存模块。 这个模型的关键限制是它忽略了事实和文本片段之间丰富的关系结构。 另一方面，我们基于图形的方法明确地使用这种结构来传播嵌入。 我们在实验中比较了两种方法（ 5），并表明GRAFT-Nets在所有任务中都胜过KV-MemNN。

对于文本断言和KB事实，还尝试了非深度学习方法来进行QA。 Gardner 和 Krishnamurthy (2017）为任务使用开放词汇语义解析的传统特征提取方法。 刘某 等。(2014）使用流水线系统汇总来自非结构化和半结构化来源的证据，用于开放域QA。

另一项工作是研究KB和关系提取和知识库完成（KBC）文本的组合表示（老挝等人。，2012; Riedel等。，2013; Toutanova等。，2015; Verga等人。，2016; 达斯 等。，2017b; 韩等人。，2016）。 与KBC相比，QA的主要差异在于，在QA中，知识源的推理过程必须以问题为条件，因此不同的问题会引起KB的不同表示并保证不同的推理过程。 此外，KBC在KB之前定义的固定模式下运行，而自然语言问题可能不遵循此模式。

GRAFT-Net模型本身就是图形表示学习的大量工作（Scarselli等。，2009; 李等人。，2016; 基普夫和威灵，2016; 阿特伍德和托斯利，2016; Schlichtkrull等。，2017）。 与大多数其他基于图形的模型一样，GRAFT-Nets也可以被视为消息传递的实例化神经网络（MPNN）框架 吉尔默 等。(2017）。 GRAFT-Nets也是像GraphSAGE这样的归纳表示学习者（汉密尔顿 等。，2017），但在异构的节点混合上运行，并使用检索来获取子图而不是随机抽样。 最近提出的Walk-Steered Convolution模型使用随机游走来学习图形表示（江等人。，2018）。 我们的个性化技术也借鉴了这种随机游走文献，但是用它来定位嵌入的传播。 通过基于深度学习的方法（如内存网络），在KB上实现了QA的巨大进步（Bordes等。，2015; 耆那教，2016）和强化学习（梁等人。，2017; 达斯等人。，2017a）。 但是，用文本扩展它们是我们的主要关注点，这是非常重要的。 在另一个方向，还有关于生成文本数据的简约图形表示的工作（克劳斯等人。，2016; Lu等人。，2017）；然而，在本文中，我们使用一个简单的顺序表示来增加实体链接到KB效果很好。

对于仅限文本的质量保证，主要关注的是阅读理解的任务（Seo等人。，2017; 龚和鲍曼，2017; 胡等人。，2017; 沉等人。，2017; Yu等人。，2018）自推出SQuAD（Rajpurkar等人。，2016）。 这些系统假设含有答案的段落是先验的，但是当这个假设放松时已经取得了进展（陈 等 人 。，2017; Raison 等 。，2018; Dhingra 等 。，2017; 王 等 。，2018，2017; Watanabe等。，2017）。 我们在后一种环境中工作，必须从大型信息源检索相关信息，但我们也将KB纳入此过程。

## 5 实验和结果

### 5.1 数据集

WikiMovies-10K由来自WikiMovies数据集的10K随机抽样训练问题组成（米勒等人。，2016），以及原始的测试和验证集。 我们对训练问题进行抽样以创建更难的设置，因为原始数据集仅在8种不同的关系类型上有100K问题，这在我们看来是不现实的。 在 5.4 我们还使用完整的训练集与现有的最新技术进行了比较。

我们使用由Wikipedia发布的KB和文本语料库米勒等人。 (2016）。 对于实体链接，我们使用简单的表面级匹配，

并检索种子周围的前50个实体以创建问题子图。 我们使用Lucene搜索文本语料库，将前50个句子（以及他们的文章标题）添加到子图中。 我们构建的子图中的整体回答是99.6%。

WebQuestionsSP （Yih 等。，2016）由Freebase实体提出的4737个自然语言问题组成，分为3098个培训和1639个测试题。 我们为模型开发和早期停止预留了250个培训问题。我们使用连接S-MART输出的实体[5]并从Freebase中的问题种子周围的邻域中检索500个实体以填充问题子图[6]。 我们进一步从维基百科中检索前50个句子，其中包含两个阶段的过程 2。 各子图的答案总体回忆率为94.0%。 §

表 1 显示每个数据集中所有检索到的子图的统计数据。 这两个数据集呈现出不同程度的难度。 虽然WikiMovies中的所有问题都对应于单个KB关系，但对于WebQues-tionsSP，模型需要针对30%的问题聚合超过两KB的事实，并且还需要推理超过7%问题的约束（梁等人。，2017）。 为了实现最大的可移植性，QA系统需要在多个KB可用性程度上具有可靠性，因为不同的域可能包含不同数量的结构化数据； KB完整性也可能随时间而变化。 因此，我们构建了另外3个数据集，每个数据集来自上面两个，KB事实数量下采样到原始数据的10%，30%和50%，以模拟KB不完整的设置。 我们重复每个采样KB的检索过程。

### 5.2 比较模型

KV-KB是来自的关键值存储器网络模型 米勒等人。 (2016); 达斯等人。 (2017c) 但只使用KB并忽略文本。 KV-EF（早期融合）是同一模型，可以访问KB和文本作为记忆。 对于文本，我们在整个句子上使用BiLSTM作为键，并将实体作为值提及。 这种重新实现在纯文本和仅KB的WikiMovies任务上表现出比结果更好的性能。

[5] http: / / / / github.com scottyih斯塔格

[6] 共有13个问题没有检测到实体。 这些在培训期间被忽略，在评估期间被视为不正确。

移植过[7]（见表 4）。 GN-KB是忽略文本的GRAFT-Net模型。 GN-LF是GRAFT-Net模型的后期融合版本：我们训练两个单独的模型，一个仅使用文本，另一个仅使用KB，然后将两者合并[8]。 GN-EF是我们早期融合的主要GRAFT-Net模型。 GN-EF + LF是GN-EF和GN-LF模型的集合体，采用与GN-LF相同的集成方法。 我们报告Hits @ 1，这是模型中最高预测答案的准确性，以及F1分数。 为了计算F1得分，我们调整开发集上的阈值，以根据子图中每个节点的二进制概率选择答案。

### 5.3 主要结果

表 2 提供了所有数据集中上述模型的比较。GRAFT-Nets（GN）在所有设置中都显示了对两个数据集上的KV-MemNN的一致改进，包括仅KB（-KB），仅文本（-EF，纯文本列）和早期融合（-EF）。 有趣的是，我们观察到KV模型的Hits和F1得分之间的相对差距比我们的GN模型更大。 我们认为这是因为对KV的关注是对存储器的标准化，这是KB事实（或文本句子）：因此模型不能同时为多个事实分配高概率。 另一方面，在GN中，我们将关注从节点传出的关系类型规范化，因此可以为所有正确答案分配高权重。

我们还看到早期融合与晚期融合（-LF）的持续改进，并且通过将它们组合在一起，我们看到所有模型中的最佳性能。 在表中 2（右），随着KB量的增加，我们进一步显示KV-EF相对于KV-KB，GN-LF和GN-EF相对于GN-KB的改善。这衡量了这些方法在利用文本和KB方面的有效性。 对于KV-EF，当KB非常不完整时，我们会看到改进，但在完整KB设置中，融合方法的性能更差。 类似的趋势适用于GN-LF。另一方面，带有文本的GN-EF比仅KB的更好

[7] 对于所有KV型号，我们调整了层数1,2,3，批量大小{0,30,}30，型号尺{寸50,80。}我们还在KB + Tex{t设置}中使用事实丢失正则化，调整范围为0,0.2,0.4。

[8] 对于合奏，我们采用模型产生的答案概率的加权组合，并在开发集上调整权重。 对于仅在文本中或仅以KB为单位的答案，我们按原样使用概率。

| 数据集 | # train / dev / test | #entity节点 | #edge类型 | #文档节点 | # 问题词汇 |
|---|---|---|---|---|---|
| WikiMovies-10K | 10K 10K 10K / / | 43,233 | 9 | 79,728 | 1759 |
| WebQuestionsSP | 2848 / 250 / 1639 | 528,617 | 513 | 235,567 | 3781 |

表1：WikiMovies-10K和WebQuestionsSP的所有检索子图$\bigcup_q G_q$的统计数据。

| 模型 | 纯文本 | KB +文字 | | | |
|---|---|---|---|---|---|
| | | 10 % | 30% | 50% | 100% |
| WikiMovies-10K | | | | | |
| kV KB | – | 15.8 / 9.8 | 44.7 / 30.4 | 63.8 / 46.4 | 94.3 / 76.1 |
| kV | 50.4 / 40.9 | 53.6 / 44.0 | 60.6 / 48.1 | 75.3 / 59.1 | 93.8 / 81.4 |
| GN KB | – | 19.7 / 17.3 | 48.4 / 37.1 | 67.7 / 58.1 | **97.0 / 97.6** |
| GN LF | | 74.5 / 65.4 | 78.7 / 68.5 | 83.3 / 74.2 | 96.5 / 92.0 |
| GN EF | **73.2 / 64.0** | 75.4 / 66.3 | 82.6 / 71.3 | 87.6 / 76.2 | 96.9 / 94.1 |
| GN-EF+LF | | **79.0 / 66.7** | **84.6 / 74.2** | **88.4 / 78.6** | **96.8 / 97.3** |
| WebQuestionsSP | | | | | |
| kV KB | – | 12.5 / 4.3 | 25.8 / 13.8 | 33.3 / 21.3 | 46.7 / 38.6 |
| kV | 23.2 / 13.0 | 24.6 / 14.4 | 27.0 / 17.7 | 32.5 / 23.6 | 40.5 / 30.9 |
| GN KB | – | 15.5 / 6.5 | 34.9 / 20.4 | 47.7 / 34.3 | 66.7 / 62.4 |
| GN GN EF LF | **25.3 / 15.3** | 31.5 / 17.7<br>29.8 / 17.0 | 40.7 / 25.2<br>39.1 / 25.9 | 49.9 / 34.7<br>46.2 / 35.6 | 67.8 / 60.4<br>65.4 / 56.8 |
| GN-EF+LF | | **33.3 / 19.3** | **42.5 / 26.7** | **52.3 / 37.4** | **68.7 / 62.3** |

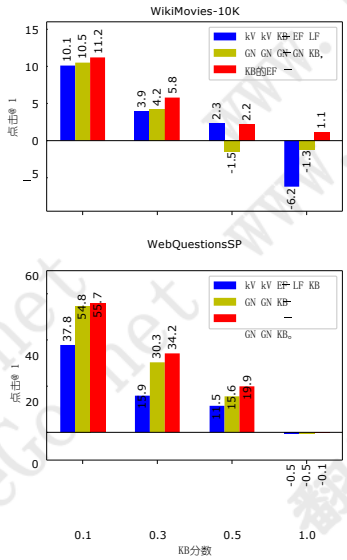



表2：左：与只有KB（-KB），早期融合（-EF）和晚期融合（-LF）设置的KV-MemNN（KV）相比，GRAFT-Nets（GN）的命中@ 1 / F1得分。 右图：随着KB完整性的增加，在仅KB（-KB）设置上改进早期融合（-EF）和晚期融合（-LF）。

在所有设置中提供。 然而，正如我们所期望的那样，随着KB变得越来越完整，添加文本的好处也会减少。

### 5.4 与专业方法的比较

在表中 4 我们将GRAFT-Nets与仅使用KB或仅使用文本进行QA专门设计和调整的最先进模型进行比较。 对于此实验，我们使用完整的WikiMovies数据集来实现与先前报告的数字的直接比较。 对于DrQA（陈等人。，2017），在原始论文之后，我们限制WebQuestionsSP的答案范围以匹配Freebase中的实体。 在每种情况下，我们还仅使用KB事实或仅使用文本句子训练GRAFT-Nets。 在四种情况中的三种中，我们发现GRAFT-Nets匹配或优于现有的theart-theart模型。 我们强调后者没有处理融合环境的机制。

唯一的例外是WebQuestionsSP的仅限KB的情况，其中GRAFT-Net比神经符号机器差6.2％的F1点（梁等人。，2017）。 分析提出了三种解释：（1）在仅KB设置中，子图检索的召回率仅为90.2％，这限制了整体性能。 在oracle环境中

我们确保答案是子图的一部分，

F1比分增加4.8％。 （2）我们对所有问题使用相同的概率阈值，即使答案的数量可能有很大差异。 将查询解析为符号形式的模型不会遇到此问题，因为以确定的方式检索答案。如果我们为每个问题调整单独的阈值，则F1得分提高7.6％。 （3）在选择答案时存在约束的少数情况下，GRAFT-Nets表现不佳（例如，“谁首先在家庭中表达梅格”）。 如果我们忽略这些约束，并且认为与种子具有相同的关系序列的所有实体都是正确的，那么F1的性能提高了3.8％。 诸如那些使用的启发式 Yu等人。 （2017）可用于改善这些情况。数字 3 显示了GRAFT-Net无法准确预测正确答案集的示例。

### 5.5 模型组件的影响

异构更新。 我们测试了我们模型的非异构版本，而不是使用细粒度实体链接信息来更新Eqs中的节点表示（M（v）和L（d，p）。 1，3a），我们汇总了文件

其所有职位均为 $J.P$ H和 $(l)_d$

将此组合状态用于所有更新。 没有

| 题 | 正确答案 | 预测的答案 |
|---|---|---|
| 大多数人在阿富汗说什么语言 | 普什图语，波斯语（东方语言） | 普什图语 |
| 什么大学做约翰斯托克顿去 | 冈萨加大学 | 冈萨加大学，Gonzaga预科学校 |

表3：WebQuestionsSP数据集中的示例。 上：模特错过了正确的答案。 底部：该模型预测了一个额外的错误答案。

| 方法 | WikiMovies（完整版） | | WebQuestionsSP | |
|---|---|---|---|---|
| | kb | 医生 | kb | 医生 |
| MINERVA 97.0 / - | – | – | – | – |
| R2的LCP | – | 85.8 / – | – | – |
| NSM | – | – | **– / 69.0** | – |
| DrQA* | – | – | – | 21.5 / |
| - R-GCN# | **96.5 / 97.4** | – | 37.2 / 30.5 | – |
| KV | 93.9 / – | 76.2 / – | – / – | – / – |
| KV# | 95.6 / 88.0 | 80.3 / 72.1 | / / 13.0 GN 38.6 | |
| 23.2 46.7 | **96.8 / 97.2** | **86.6 / 80.8** | 67.8 / 62.8 | **25.3 / 15.3** |

表4：与仅使用KB或文本的SOTA模型相比，命中@ 1 / F1得分：MINERVA（达斯等人。，2017a），R2-AsV（Watanabe等。，2017），神经符号机（NSM）（梁等。，2017），DrQA（陈等人。，2017），R-GCN（Schlichtkrull等。，2017）和KV-MemNN（米勒等人。，2016）。 * DrQA在SQuAD上进行了预训练。 #重新实现。

异构更新，所有实体v L（d，）将从文档d接收相同的更新。 因此，该模型不能消除同一文档中提到的不同实体的歧义。 结果见表 5 表明这个版本一直比异构模型差。

| 0 KB | 0.1 0.3 0.5 1.0 kb的片段22.7 / 13.6 kb的 |
|---|---|
| NH 4 KB / 15.8 35.6 / / / 66.5 59.8% 23.2 47.2 33.3 | |
| H 25.3 / 15.3 31.5 / 17.7 40.7 / 25.2 49.9 / 34.7 67.8 / 60.4 | |

表5：WebQuestionsSP上的非异构（NH）与异构（H）更新

**调**整问题。 我们对定向传播方法和对关系的关注进行了消融测试。 我们观察到这两个组件都会带来更好的性能。 在完整和不完整的KB场景中都会观察到这种效果，例如在WebQuestionsSP数据集上，如图所示 4 （剩下）。

**事实辍学。** **数字** 4 （右）比较早期融合模型的表现，因为我们改变了事实辍学率。 适度的事实

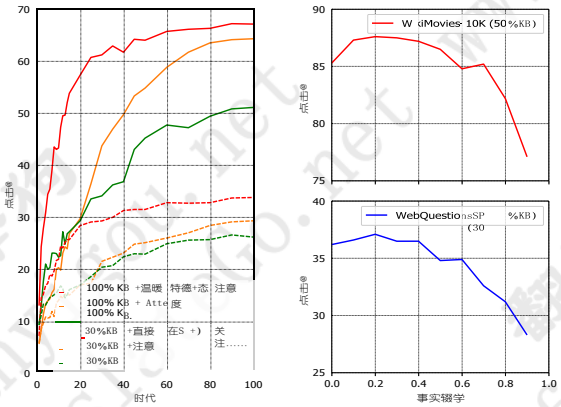



图4：左：定向传播的影响和对具有30％KB和100％KB的We-bQuestionsSP数据集的关系的基于查询的关注。 右：点击@ 1，不同的事实辍学率和WikiMovies和WebQuestionsSP。

dropout提高了两个数据集的性能。 随着事实辍学率的增加，性能会提高，直到模型无法从KB学习推理链。

## 6 结论

在本文中，我们使用文本结合不完整的KB来调查QA，这一任务在过去受到了有限的关注。 我们通过修改现有的问答数据集为这项任务介绍了几个基准问题，并讨论了解决这个问题的两种广泛方法 - “晚期融合”和“早期融合”。 我们表明早期融合方法表现更好。 我们还引入了一种新的早期融合模型，称为GRAFT-Net，用于对包含KB实体和文本文档的子图中的节点进行分类。 GRAFT-Net建立在图表表示学习的最新进展基础之上，但包括一些改进此任务性能的创新。 GRAFT-Nets是一个单一的模型，它在纯文本和仅KB的集合中实现了与theart-theart方法相比的性能 -

使用文本与不完整的KB结合使用时，并且优于基线模型。 目前未来工作的方向包括 - （1）扩展GRAFT-Nets选择文本范围作为答案，而不仅仅是实体和（2）改进子图检索过程。

## 致谢

Bhuwan Dhingra由NSF根据CCF-1414030和IIS-1250956以及Google的拨款提供支持。 Ruslan Salakhutdinov部分得到ONR拨款N000141812861，Apple和Nvidia NVAIL奖的支持。

## 参考

詹姆斯阿特伍德和唐托斯利。 2016.扩散 - 卷积神经网络。 “神经信息处理系统进展”，第1993-2001页。

Soerren Auer，Christian Bizer，Georgi Kobilarov，Jens Lehmann，Richard Cyganiak和Zachary Ives。 Dbpedia：开放数据网络的核心。 在语义网中，第722-735页。 斯普林格。

彼得鲍迪斯。 2015.Yodaqa：模块化问答系统管道。 在POSTER 2015-19th国际电气工程学生会议上，第1156-1165页。

Kurt Bollacker，Colin Evans，Praveen Paritosh，Tim Sturge和Jamie Taylor。 Freebase：一个用于构建人类知识的协作创建的图形数据库。 在2008年ACM SIGMOD数据管理国际会议论文集，第1247-1250页。 ACM。

Antoine Bordes，Nicolas Usunier，Sumit Chopra和Jason Weston。 2015.大规模简单的问题回答与内存网络。 arXiv preprint arXiv：1506.02075。

Danqi Chen，Adam Fisch，Jason Weston和Antoine Bordes。 2017.阅读维基百科以回答开放域名问题。 在计算语言学协会（ACL）。

Rajarshi Das，Shehzaad Dhuliawala，Manzil Zaheer，Luke Vilnis，Ishan Durugkar，Akshay Krishnamurthy，Alex Smola和Andrew McCallum。 2017A。 散步并得出答案：使用强化学习推理知识库中的路径。 arXiv preprint arXiv：1711.05851。

Rajarshi Das，Arvind Neelakantan，David Belanger和Andrew McCallum。 2017b。 使用递归神经网络推理实体，关系和文本的链。 EACL。

Rajarshi Das，Manzil Zaheer，Siva Reddy和Andrew McCallum。 2017c。 使用通用模式和内存网络回答知识库和文本的问题。 ACL。

Bhuwan Dhingra，Kathryn Mazaitis和William W Cohen。 2017. Quasar：通过搜索和阅读来回答问题的数据集。 arXiv preprint arXiv：1707.03904。

Bhuwan Dhingra，Danish Pruthi和Dheeraj Rajagopal。 2018.简单有效的半监督问答。 NAACL。

David Ferrucci，Eric Brown，Jennifer Chu-Carroll，James Fan，David Gondek，Aditya A Kalyanpur，Adam Lally，J William Murdock，Eric Nyberg，John Prager，et al。 2010.Watson建筑：deepqa项目概述。 AI杂志，31（3）：59-79。

Matt Gardner和Jayant Krishnamurthy。 2017.开放式词汇语义解析，包括分布统计和形式知识。 在AAAI，第3195-3201页。

Justin Gilmer，Samuel S Schoenholz，Patrick F Riley，Oriol Vinyals和George E Dahl。 2017.量子化学的神经信息传递。 ICML。

Yichen Gong和Samuel R Bowman。 2017.反刍读者：通过门控多跳注意推理。 arXiv preprint arXiv：1704.07415。

William L. Hamilton，Rex Ying和Jure Leskovec。 2017.大型图表上的归纳表示学习。 CoRR，abs / 1706.02216。

徐晗，刘志远，孙茂松。 2016年。知识图完成的文本和知识的联合代表学习。 arXiv preprint arXiv：1611.04125。

Taher H Haveliwala。 2002.主题敏感的pagerank。 在第11届万维网国际会议论文集，第517-526页。 ACM。

胡明浩，彭宇兴，邱鹏鹏。 2017年。机器理解的助记符读者。 arXiv preprint arXiv：1705.02798。

Mohit Iyyer，Wen-tau Yih和Chang-Wei Chang。 2017.基于搜索的神经结构学习，用于顺序问答。 “计算语言学协会第55届年会论文集（第1卷：长篇论文）”，第1卷，第1821-1831页。

萨尔塔克耆那教。 2016.使用事实记忆网络回答知识库的问题。 在NAACL学生研究讲习班的会议录，第109-115页。

江嘉涛，崔震，徐春燕，李成政，杨健。 2018.步进控制的图表分类。 arXiv preprint arXiv：1804.05837。

Thomas N Kipf和Max Welling。 2016.使用图卷积网络的半监督分类。 arXiv preprint arXiv: 1609.02907。

Sebastian Krause, Leonhard Hennig, Andrea Moro, Dirk Weissenborn, Feiyu Xu, Hans Uszkoreit和Roberto Navigli。 2016. Sar-graphs: 一种语言资源，将语言知识与知识图中的语义关系联系起来。 Web语义: 万维网上的科学，服务和代理，37: 112-131。

Ni Lao, Amarnag Subramanya, Fernando Pereira和William W Cohen。 2012.使用学习的语法 - 语义推理规则阅读网络。 在2012年自然语言处理和计算自然语言学习中的经验方法联合会议论文集，第1017-1026页。 计算语言学协会。

Yujia Li, Daniel Tarlow, Marc Brockschmidt和Richard Zemel。 2016.门控图序列神经网络。 ICLR。

陈亮，Jonathan Berant，Quoc Le，Kenneth D Forbus和Ni Lao。 2017.神经符号机器：在监控较弱的情况下在freebase上学习语义解析器。 ACL。

陆正东，崔浩天，刘祥根，严毓妍，郑大琦 2017.面向对象的神经编程（oonp）用于文档理解。 arXiv preprint arXiv: 1709.08853。

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes和Jason Weston。 2016.用于直接读取文档的键值存储网络。 EMNLP。

Bonan Min, Ralph Grishman, Li Wan, Chang Wang和David Gondek。 2013.对不完整知识库的关系提取的远程监督。 在2013年计算语言学协会北美分会会议记录：人类语言技术，第777-782页。

Martin Raison，Pierre-Emmanuel Mazare'，Rajarshi Das和Antoine Bordes。 2018. Weaver：深入共同编码机器阅读的问题和文件。 arXiv preprint arXiv: 1804.10490。

Pranav Rajpurkar，张健，Konstantin Lopyrev和Percy Liang。 2016.Squad：机器理解文本的100,000多个问题。 arXiv preprint arXiv: 1606.05250。

Sebastian Riedel, Limin Yao, Andrew McCallum和Benjamin M Marlin。 2013.使用矩阵分解和通用模式的关系提取。 在2013年计算语言学协会北美分会会议记录：人类语言技术，第74-84页。

Pum-Mo Ryu, Myung-Gil Jang和Hyun-Ki Kim。 2014.使用基于维基百科的知识模型开放域名问题解答。 信息处理与管理，50（5）：683 - 692。

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner和Gabriele Monfardini。 2009.图神经网络模型。 IEEE Transactions on Neural Networks，20（1）：61-80。

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov和Max Welling。 2017.使用图卷积网络建模关系数据。 arXiv preprint arXiv: 1703.06103。

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi和Hannaneh Hajishirzi。 2017.机器理解的双向注意力流程。 ICLR。

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen。 理智：学会停止阅读机器理解。 在第23届ACM SIGKDD知识发现和数据挖掘国际会议论文集，第1047-1055页。 ACM。

A. Talmor和J. Berant。 2018.网络作为回答复杂问题的知识库。 在北美计算语言学协会（NAACL）。

Kristina Toutanova，Danqi Chen，Patrick Pantel，Hoi-fung Poon，Pallavi Choudhury和Michael Gamon。 2015.代表联合嵌入文本和知识库的文本。 在2015年自然语言处理经验方法会议论文集，第1499-1509页。

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth和Andrew McCallum。 2016.使用组合通用模式的多语言关系提取。 NAACL。

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun和Rob Fergus。 2013.使用dropconnect进行神经网络的正则化。 在国际机器学习会议，第1058-1066页。

王硕航，莫宇，郭晓晓，王志国，蒂姆克林格，张伟，张世玉，杰拉尔特沙罗，周博文，荆江。 2018. $R^3$：增强的开放域问答的读者 - 排名。 AAAI。

王朔杭，莫宇，江江，张伟，郭晓晓，张世宇，王志国，蒂姆克林格，杰拉尔德特萨罗和默里坎贝尔。 2017.在开放域问题回答中回答重新排名的证据汇总。 arXiv preprint arXiv: 1711.05116。

Yusuke Watanabe，Bhuwan Dhingra和Ruslan Salakhutdinov。 2017.通过检索和理解从非结构化文本回答的问题。 arXiv preprint arXiv: 1703.08885。

Johannes Welbl, Pontus Stenetorp和Sebastian Riedel。 2018.构建跨文档的多跳阅读理解的数据集。 TACL。

Georg Wiese, Dirk Weissenborn和Mariana Neves。2017.生物医学问答的神经域适应。 在第21届计算机自然语言学习会议论文集（CoNLL 2017），第281-289页，加拿大温哥华。 计算语言学协会。

Wen-tau Yih, Matthew Richardson, Chris Meek, MingWei Chang和Jina Suh。 2016.知识库问答的语义解析标签的价值。 “计算语言学协会第54届年会论文集（第2卷：短文）”，第2卷，第201-206页。

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi和Quoc V Le。 2018. Qanet：将局部卷积与全球自我关注相结合，以便阅读理解。 ICLR。

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang和Bowen Zhou。2017.改进的知识库问答的神经关系检测。 ACL。