

2023-2024 学年冬季学期

《数据科学与工程 2-知识工程》(3XS081017)

课程报告

成绩	
----	--

学号	23724590	学院	材料基因组工程研究院
姓名	郎志远	手工签名	
报告题目	基于知识图谱的问答系统		
课程 报 告 成 绩	1. 选题：有前沿性和应用前景。（10）		
	2. 报告主要部分： 2.1. 对所研讨领域的相关研究工作进行较全面的综述，相关研究要做到查新、查全，有论述有评价，具体介绍的方法不少于 4 种；（30） 2.2.报告主要内容叙述条理清楚、内容详细（正文部分不少于 8 页）；（20） 2.3 有相关技术的案例分析，鼓励对研讨内容中的模型和算法进行实验复现或应用实践；（20） 2.4 鼓励原创性工作，包括提出新的思路、模型、算法或者实验。（10） 书写格式：书写规范、用词正确、无明显的错别字，图表、代码清晰规范，格式协调、效果好，参考文献书写（不少于 10 篇参考文献，英文文献不少于 2 篇）、引用规范合理。（10）		
报告评语：			
教师签名：			
日期： 年 月 日			

基于知识图谱的问答系统

郎志远

(23724590, 材料基因组工程研究院)

摘要

本报告探讨了基于知识图谱的问答系统，报告首先介绍了构建模板的问答方法和语义解析的问答方法，包括基于词典-文法的语义解析、基于语义图的问句解析以及基于神经网络的方法。接着，报告详细阐述了构建知识图谱问答系统的主要方法，如三元组匹配、知识图谱嵌入、记忆网络和信息抽取。特别地，报告使用基于微调大语言模型的知识图谱问答框架进行了微调，通过在服务器上部署的 Yi-6B-Chat 模型的微调前后问答对比，展示了该框架在基于知识图谱的问答系统领域的适用性。

关键词：知识图谱；问答系统；大语言模型

1 前言

问答系统作为信息检索和自然语言处理领域的交叉研究方向，其核心目标是使计算机能够理解并自动回答用户以自然语言形式提出的问题。知识库作为一种存储和管理结构化及非结构化知识的数据管理系统，为问答系统提供了丰富的信息源^[1]。基于知识图谱的问答系统仅能够理解用户的自然语言输入，还能够庞大的知识库中快速定位到准确的答案，极大地提高了信息检索的效率和准确性。然而，随着知识图谱数据规模的不断扩大和用户直接获取答案需求的增长，现有的开放领域知识图谱问答研究面临着一系列挑战。首先，系统需要从自然语言问题中准确提取语义信息，以便在庞大的知识图谱中定位到与问题高度相关的子图谱，从而避免不必要的搜索空间。其次，如何训练系统充分理解和利用知识图谱中的信息，是需要解决的另一个难题。本报告将深入介绍构建基于知识图谱的问答系统的关键技术。

2 基于知识图谱的问答系统方法基本原理和关键技术

2.1 构建模板的问答方法

构建模板的问答方法是一种基于规则的自然语言处理技术，它通过预定义的模板来处理和回答用户提出的问题^[2]。这种方法通常用于构建问答系统，尤其是在处理特定领域或简单问题时。构建模板问答方法的基本步骤如下：首先设计一系列预定义的问题模板，这些模板覆盖了系统能够回答的问题类型，模板通常包含占位符，用于在实际问题中插入具体的信息。接着使用自然语言处理技术（如分词、词性标注、依存句法分析等）来理解用户问题的意图，确定用户问题与预定义模板之间的匹配关系。然后根据用户问题的解析结果，将问题中的关键词或短语填充到模板的占位符中。最后使用填充后的模板来生成答案。这可能涉及到从知识库中检索信息，或者执行一些简单的逻辑操作。在某些情况下，可能需要对生成的答案进行后处理，以确保其自然流畅。

构建模板的问答方法的优点是实现简单，对于特定领域的问题处理效率高，由于模板是预定义的，系统可以快速生成查询并返回结果，模板方法可以确保查询的准确性。然而，这种方法的缺点是灵活性有限，模板可能无法完全适应所有类型的查询，特别是那些需要复杂逻辑或新信息的查询，对于复杂或未定义的问题

题可能无法有效处理。另外模板库维护成本高，为了覆盖各种用户查询，需要构建和维护一个庞大的模板库，这需要大量的人力和时间。在未来可以考虑利用机器学习技术，自动从知识库中学习并生成新的模板，或者开发能够根据用户查询动态调整的模板，以提高系统的灵活性和适应性。通过这些优化，基于模板的基于知识图谱的问答系统可以在保持高效率 and 准确性的同时，提高对新查询类型的适应能力，减少人工维护的工作量。

2.2 语义解析的问答方法

语义解析涉及对自然语言问句的深入分析，将其转化为机器可以理解的逻辑表达式^[3]。这一过程的关键在于捕捉问句中的语义成分，并将这些成分映射为知识图谱中的结构化查询，类似于 SQL 语句。尽管这种方法在查询效率上具有优势，但它对普通用户来说可能存在一定的使用门槛，因为设计规范的查询语句需要专业知识。为了实现基于知识图谱的语义解析问答，系统必须完成两个关键步骤：首先，利用语义解析器将用户的自然语言问题转换为机器可理解的语义表示；其次，基于这个语义表示生成结构化查询语言，对知识图谱进行查询。

2.3.1 基于词典-文法的语义解析方法

基于词典-文法的语义解析方法是一种结构化的语言理解技术，它依赖于精心构建的词典和一套详尽的文法规则来分析自然语言句子。这种方法的核心在于创建一个能够将自然语言转换为机器可理解的语义表示的形式化语法系统。首先，需要构建一个包含丰富词汇及其语法类别和语义角色的词典，这是理解句子结构的基础。随后，定义一系列文法规则，这些规则规定了词汇如何合法地组合成句子。在给定一个自然语言句子时，系统会应用这些词典和文法规则进行句法分析，最终从解析过程中提取出句子的语义结构，这种结构可能是逻辑形式、框架表示或其他形式的语义表示。通过这种方法，系统能够深入理解句子的深层含义，并为后续的问答、信息检索等任务提供精确的语义信息。

基于词典-文法的语义解析方法以其高度的可解释性和清晰的结构，在特定领域的问答系统中表现出色。然而，这种方法依赖于人工编写的词汇表和规则集，这不仅耗时费力，而且在训练数据有限的情况下，其性能会显著下降。此外，这种方法在语义表示与知识库的紧密联系上存在不足，导致在解析过程中难以利用知识库的约束来提高准确性。更进一步，面对大规模知识库的开放域特性，文本歧义问题尤为突出，这进一步加了解析的复杂性和难度。因此，尽管基于词典-文法的方法在某些应用场景下有效，但在更广泛的自然语言处理任务中，需要考虑其局限性并探索更灵活、自动化的方法。

2.3.2 基于语义图的问句解析方法

与基于词典-文法的语义解析方法不同，基于语义图的问句解析方法通过对问句分析，构建相对应的语义图，语义图由节点（实体、变量或者类型）、边（关系）、操作符（count、argmax 等）构成，被看作知识图谱子图，实现将问句映射到知识图谱中，再通过图匹配完成问题回答。这种方法特别适用于需要多跳推理和复杂逻辑结构的问题。比如在处理问句“谁是美国总统？”时，系统首先识别出关键实体“总统”和“美国”，然后确定它们之间的关系，即总统是与美国相关的职位。接着，系统构建一个语义图，其中“总统”作为一个未知实体（用 X 表示），并通过“是美国总统”的关系与“美国”相连。随后，系统利用这个语义图在知识库中搜索，找到与“美国”相关且当前担任总统职位的实体。最后，这个实体被用来填充语义图中的 X 位置，从而得出答案：“乔·拜登”。

基于语义图的问句解析方法关键在于语义图表示以及语义图构建，优点是语义图结构和自然语言句子结构具有相似性，语义解析过程充分利用知识库的知识约束，由组合文法转换为语义图构建，减少了搜索空间。缺点是依赖一些启发式方法构建语义图，导致缺乏通用性。

2.3.3 基于神经网络的方法

基于神经网络的方法将自然语言及对应的语义看作是两种不同语言，语义分析任务被看做类似于机器翻译任务，利用端到端模型，实现将问句翻译成对应语义的表示序列。比如在处理用户提出的“谁是爱因斯坦？”这个问题时，问答系统首先将问题转换成一系列词向量，这些向量代表了问题的语义内容。随后，这些词向量序列被输入到一个深度神经网络中。网络通过逐个处理这些向量，利用其结构来捕捉序列中的长期依赖关系，从而理解问题的深层含义。接下来，网络输出一个向量，这个向量是对问题语义的精炼表示。系统使用这个向量在知识库中进行检索，知识库中的答案同样被编码为向量形式。通过计算问题向量与答案向量之间的相似度，系统能够找到与问题最匹配的答案。最后，系统从知识库中选取与问题向量相似度最高的答案向量，并将其解码回自然语言，给出最终答案：“阿尔伯特·爱因斯坦”。这个过程体现了问答系统如何通过深度学习和知识库的结合，实现对复杂问题的理解和准确回答。

基于神经网络方法，相比其他两种方法，模型简单，在处理开放领域的问题时，基于神经网络的方法显示出强大的能力，尤其是在处理那些没有明确结构或规则的问题时。这种方法在大规模数据集上训练时，能够生成准确且流畅的逻辑形式，适用于各种类型的问答系统。但是它的可解释性差，还需要预先准备相应的训练语料，对模型进行训练时，训练过程较长，还需要调整参数。

3 构建知识图谱问答系统主要方法

3.1 三元组匹配

三元组匹配方法起源于早期的 RDF 问答系统，通过将自然语言问题转换为三元组形式，利用相似性度量在 RDF 数据中检索相似的三元组来找到答案。这种方法已经演变为两种主要的流水线策略：问题拆分和问题复述。问题拆分侧重于将复杂问题分解为更小的子问题，而问题复述则侧重于改写问题以更好地匹配知识图谱中的结构。Bao 等人提出了一种三元组匹配方法，该方法通过三个主要步骤来处理知识图谱问答：问题拆分、翻译和三元组排序。首先，在问题拆分阶段，利用依存解析树模板将问题分解为子问题，然后通过合并子问题的答案来确定最终答案。接着，在翻译阶段，采用 CYK 算法结合预定义的问题模式模板和关系表达式，将自然语言问题转换为对应的三元组。最后，在三元组排序阶段，根据一系列特征（如答案中包含的原问题单词数量、疑问词数量、转换成功的跨度数量以及翻译得到的三元组总数）对三元组进行排序。排序过程使用最小错误率训练算法来调整特征权重，以优化排序结果。这种方法通过细化问题处理流程，提高了对复杂问题的解析能力。也可以将问题拆分改为问题复述，使用人工预定义模板对自然语言问题进行改写，表述为与三元组具有相似词汇、语法结构的多个新问题。

代码逻辑如下：

```
1 knowledge_graph = {#以字典的形式表示一个简化的知识图谱
2     'Person': {'name': 'John Doe', 'age': '30'},
3     'City': {'name': 'New York', 'country': 'USA'},}
4 def split_question(question): # 问题拆分：将问题分解为更小的子问题
5     return question.split()# 这里我们简单地通过空格分割问题
6 def translate_to_triples(question, knowledge_graph):#将子问题转换为三元组
7     triples = []#假设每个子问题对应一个三元组
8     for sub_question in split_question(question):
9         triple = ('Person', 'name', sub_question)
10        triples.append(triple)
```

```

11     return triples
12 def sort_triples(triples, features): # 三元组排序：根据特征对三元组进行排序
13     sorted_triples = sorted(triples, key=lambda x: sum([features.get(f, 0)
14                                                         for f in x]))
14     return sorted_triples

```

3.2 知识图谱嵌入

知识图谱嵌入允许问题和候选答案被映射到一个共同的向量空间中，通过计算它们之间的相似度来确定最终答案。这种方法不需要依赖于额外的规则、词汇表或语法工具，而是直接利用向量表示来捕捉实体和关系之间的语义关系^[4]。由 Bordes 等人提出的 TransE^[5]是知识图谱嵌入的基石，它的改进版本通过将问题和答案的路径及其周围子图编码为嵌入向量来实现。在 TransE 模型中，问题和答案的嵌入向量通过点积计算得分，训练的目标是确保正确答案的得分最高。这种方法简化了知识图谱问答的处理流程，并为后续的神经网络模型提供了基础。

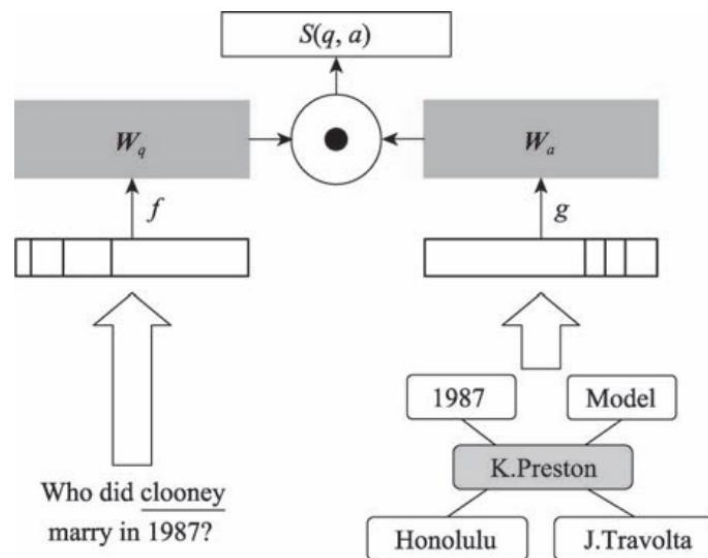


图 1 TransE 的改进版本

代码逻辑如下：

```

1 import numpy as np
2 knowledge_graph = {# 假设的知识图谱，实体和关系以列表形式表示
3     'Person': ['John Doe', 'Jane Smith'],
4     'City': ['New York', 'Los Angeles'],
5     'WorksIn': [('John Doe', 'New York'), ('Jane Smith', 'Los Angeles')]}
6 # 初始化嵌入向量，我们为每个实体和关系分配一个随机的向量
7 embedding_dim = 10 # 嵌入向量的维度
8 entity_embeddings = {entity: np.random.rand(embedding_dim) for entity in
9     knowledge_graph['Person'] + knowledge_graph['City']}
9 relation_embeddings = {relation: np.random.rand(embedding_dim) for relation in
10     knowledge_graph['WorksIn']}

```



```

10 # TransE 模型的核心是计算头实体、关系和尾实体之间的向量差
11 Def transe_score(head,relation,tail,entity_embeddings,relation_embeddings):
12     h = entity_embeddings[head]
13     r = relation_embeddings[relation]
14     t = entity_embeddings[tail]
15     return np.dot(h + r, t)
16 #计算 John Doe 在 New York 工作的概率
17 john_doe = 'John Doe'
18 new_york = 'New York'
19 score=transe_score(john_doe,'WorksIn',new_york,entity_embeddings,
20                     relation_embeddings)
21 print(f"Score for John Doe working in New York: {score}")

```

3.3 记忆网络

为了赋予模型类似于人类的记忆能力，研究人员开发了记忆网络和循环神经网络（RNN）的变体，如长短时记忆模型（LSTM）。这些模型能够长期存储和检索大量信息。记忆网络（MemNN）是一种学习系统，由 Weston 等人提出^[6]，它通过一个可读写的记忆组件来存储知识图谱的所有内容，并确定自然语言输入与 KG 查询语言之间的关系。MemNNs 系统由五个主要模块组成：记忆模块（M）、输入模块（I）、输出模块（O）、响应模块（R）和泛化模块（G）。记忆模块负责存储 KG 的三元组，并允许其他模块进行读写操作。输入模块处理外部输入，包括 KG 三元组和自然语言问题的预处理。输出模块在记忆中选择与输入最相关的内容，并对其进行评分和排序。响应模块将选定的记忆以适当的格式输出为最终结果。泛化模块则用于将外部知识整合到记忆模块中，以更新和扩展记忆内容。

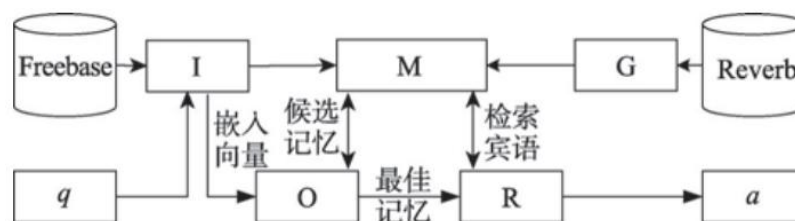


图 2 MemNN 框架流程图

3.4 信息抽取

信息抽取的定义为是指从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。文本数据是由一些具体的单位构成的，例如句子、段落、篇章，文本信息正是由一些小的具体的单位构成的，例如字、词、词组、句子、段落或是这些具体的单位的组合。抽取文本数据中的名词短语、人名、地名等都是文本信息抽取，文本信息抽取技术所抽取的信息可以是各种类型的信息^[7]。

3.4.1 命名实体识别

信息提取的首要任务是识别文本中的命名实体，这些实体可以是人名、地名、组织名，也可以是日期、时间、价格等专有名词。命名实体识别（NER）的目标是精确地定位这些实体在文本中的位置，并为其分配正确的类型标签。这一过程面临的挑战包括实体边界的模糊性和类型分类的不确定性。为了解决这些问题，NER 通常采用序列标注的方法，其中最常用的编码方式有 IO 编码和 IOB 编码。IO 编码将实体之外的

词汇标记为 O (Other)，而 IOB 编码则通过 B (Begin)、I (Inside) 等标记来指示实体的边界和内部。在基于特征的 NER 系统中，特征的提取至关重要。这些特征可能包括词性、形态、局部词汇信息以及句法结构等。对于汉语 NER，汉字层面的特征如单字提示和常用尾字也非常重要。此外，全局信息和外部知识，如维基百科等，也被广泛应用于提高 NER 模型的性能。在实际应用中，NER 系统会利用这些特征和外部知识库来训练序列分类器，如 MEMM 或 CRF，以标记新文本中的命名实体。通过这种方式，NER 不仅能够提高信息提取的准确性，还能够为其他自然语言处理任务提供基础，如问答系统、文本摘要和知识图谱构建。

3.4.2 关系抽取

知识图谱是语义关联的实体，它将人们对物理世界的认知转化为计算机能够以结构化的方式理解的语义信息。关系抽取就是从非结构化文本中提取实体之间的关系。依据实体是否在文本中被标记，关系抽取方法可分为联合抽取和流水线式抽取。在现实世界中，关系的提取要比实体提取复杂得多，自然句子的形式也多种多样，所以关系的提取比实体提取困难得多。关系抽取的方法可以分为两大类：联合抽取和流水线式抽取。联合抽取是一种端到端的方法，它试图在一个统一的框架中同时解决实体识别和关系分类的问题。这种方法的核心思想是，实体和关系之间存在内在的联系，通过同时处理这两个任务，可以更好地利用这种联系来提高整体的抽取性能。联合抽取可以共享实体和关系之间的信息，这有助于提高抽取的准确性，因为实体的上下文信息对于理解关系至关重要。整个模型可以作为一个整体进行训练，这简化了模型设计和优化过程。由于实体和关系是同时抽取的，减少了后续步骤中可能需要的后处理和调整。但是联合模型通常比流水线式模型更复杂，需要更多的计算资源和调优，为了训练有效的联合模型，需要大量同时标注实体和关系的标注数据。流水线式抽取是指先使用实体识别模型识别文本中的实体对，然后判断实体对的关系。

关系抽取经典算法有基于规则的关系提取、有监督关系抽取、远程监督模型、联合关系抽取。

基于规则的关系抽取方法依赖于预定义的模式和规则来识别文本中的实体间关系，主要分为基于触发词模式和基于依存关系（语法树）两种类型。触发词模式通过识别特定的词汇和结构来提取关系，如“X is in Y”模式用于识别位置关系，但这种方法受限于模式的固定性和对复杂句子结构的处理能力。依存关系分析则通过理解句子的语法结构来揭示实体间的深层联系，能够更准确地捕捉关系，但需要复杂的规则定义和大量的人工劳动。尽管这些方法在特定领域内表现出高准确率，但它们在多语言环境下的应用受限，且难以适应语言的多样性。

有监督神经网络方法在关系抽取领域中占据主导地位，它通过在大规模标注数据集上训练深度学习模型来实现。这种方法通常涉及构建一个层叠的二分类器，该分类器能够判断两个给定实体之间是否存在特定关系。为了训练这样的模型，文本数据需要先经过其他自然语言处理模型进行标注，以提取关键特征，如上下文单词、词性标注、实体间的依存关系等。训练过程包括手动标注正负样本，学习分类器以识别关系类型，并最终在新文本中应用这些分类器来抽取关系。尽管有监督方法能够提供高质量的监督信号和明确的负样本，但它也存在一些局限性，如标注成本高昂，难以扩展到新的关系类型，以及对新领域的泛化能力有限。此外，模型通常由特征提取器和关系分类器组成，其中特征提取器可以是 CNN、LSTM、GNN、Transformer 或 BERT 等先进的深度学习架构。

远程监督方法旨在解决有监督关系抽取任务中高质量标注数据稀缺的问题。这种方法基于一个核心假设：如果两个实体在知识库中存在某种关系，那么任何同时提及这两个实体的句子都可能表达这种关系。例如，如果实体 e_1 和 e_2 在知识库中被标记为具有关系 R_a ，那么任何包含 e_1 和 e_2 的句子都可以被视为 R_a 关系的正样本。这种方法通过结合知识库和文本库，自动生成标注数据，极大地简化了数据准备过程。

然而，远程监督也带来了一系列挑战，包括数据噪声、多关系实体对的处理以及 False negative 问题。为了应对这些挑战，研究者们提出了多种策略。多实例学习通过从多个实例中选择最具代表性的样本来提高训练效率；利用外部信息，如领域知识，来筛选出更准确的实例；以及采用先进的模型和训练方法，如 soft label、增强学习或对抗学习，来提高模型对噪声数据的鲁棒性，这些方法共同致力于提高远程监督模型的性能，使其在关系抽取任务中更加可靠和有效。

参数共享的联合模型采用了一种分步但同步的方法来抽取实体（S）、谓词（P）和对象（O）组成的三元组。这种模型的设计允许在不同步的多个步骤中完成抽取过程，同时将各个步骤的损失函数合并，以整体优化模型。在训练过程中，模型的参数会根据这些合并的损失进行更新，而后续步骤可以利用前序步骤的输出作为特征，从而实现子任务间的信息共享。相比之下，联合解码模型则更深入地体现了“联合”的概念，它在一个统一的步骤中同时识别实体和关系，从而更有效地共享子任务间的信息，然而这种模型可能无法准确识别那些在文本中重叠出现的实体关系。

3.5 向量建模

向量建模方法，借鉴了信息抽取的思想，通过将问题和答案映射到低维空间中的分布式表达来实现问题和答案之间的关联性学习，这种方法的核心在于，通过训练数据，使得问题向量和正确答案向量在低维空间中的关联得分最大化，从而筛选出最可能的答案。为了将问题和答案映射到低维空间的问题，可以采用了一种多热编码的稀疏表示方法。对于问题，通过计算问题中每个单词（或实体、关系）的出现次数来构建问题向量；而对于答案，由于答案通常对应知识库中的实体，可以采用一种简化的词袋模型，或者通过问题中的主题词在知识库图中定位到的路径来构建答案向量。在训练分布式表达时，通过最小化基于边界的排名损失函数，确保正确答案的得分高于任意错误答案。对于规模数据集，可以通过自动化方式构造了额外的数据集，并采用多任务学习策略，使得同一类簇中的问题得分较高，以此来提高模型的泛化能力。尽管向量建模方法在基于知识图谱的问答系统任务中取得了不错的效果，但它也存在一些局限性，如缺乏解释性和先验知识的应用。此外，问题向量的表示方法没有考虑语言顺序，而训练模型的复杂度相对较低。这些问题可以通过深度学习方法来解决，例如使用更复杂的神经网络结构来捕捉语言的顺序信息，以及整合更多的先验知识来提升模型的推理能力。图 3^[8]为向量建模的过程：

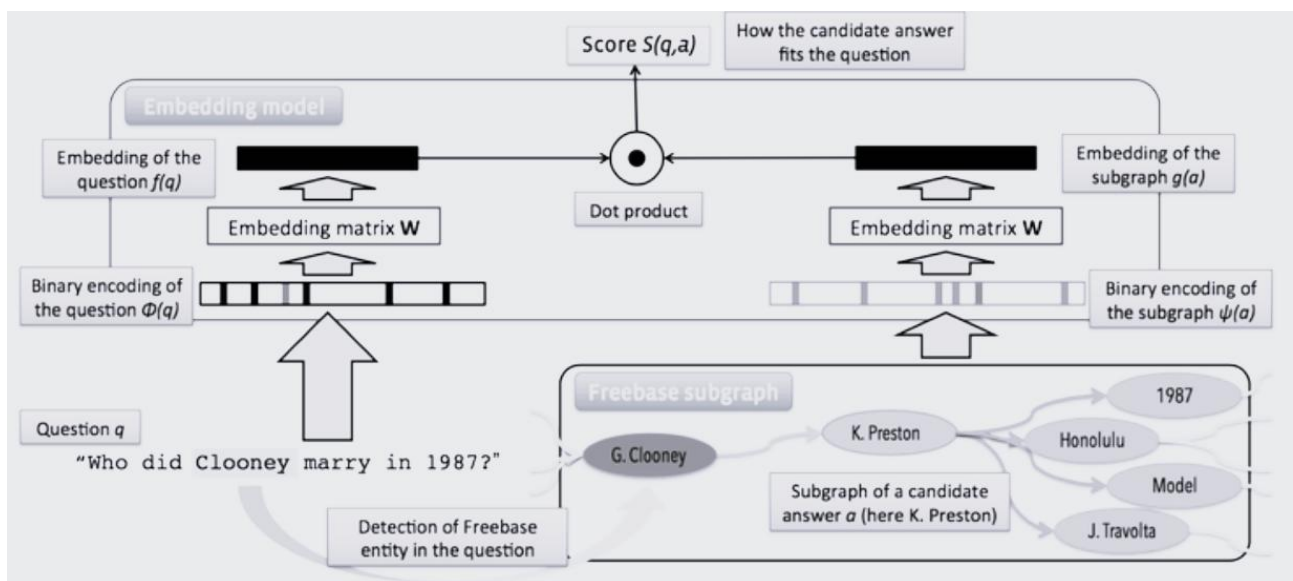


图 3 向量建模

4 基于微调大语言模型的知识图谱问答

在知识库问答领域，传统的研究方法主要关注于知识检索和语义解析两大核心问题，但这些方法面临着检索效率低下、检索错误对语义解析的误导以及任务处理流程复杂等挑战。大型语言模型的出现，如 ChatGPT 和 GPT-4，以其强大的文本生成和深度语义理解能力，为重塑基于知识图谱的问答系统任务提供了新的可能性。然而，大型语言模型的决策过程不透明，可能导致无法准确呈现知识体系，甚至产生错误答案。与此同时，知识图谱以其结构化的知识和高可解释性，为推理和问答任务提供了清晰的知识基础，但它们在自然语言处理和维持方面存在挑战。ChatChaKBQA 应运而生，旨在结合大型语言模型的学习能力和知识图谱的结构化知识，以实现更高效、可解释的问答系统^[9]。该框架通过微调大型语言模型生成逻辑形式，然后利用无监督检索方法在知识图谱中精确检索实体和关系，最后执行可解释的 SPARQL 查询。这种方法不仅提高了检索效率，减少了错误检索的影响，还简化了基于知识图谱的问答系统任务的处理流程。ChatChaKBQA 的生成-检索框架，将大型语言模型的自然语言处理能力与知识图谱的结构化知识相结合，为医疗、法律等专业领域的可解释推理问答提供了新的解决方案，展现了 LLMs 与知识图谱互补优势的完美融合。

ChaKBQA 是一个使用经过微调的开源大模型的生成-然后检索的知识库问答框架。首先，ChatChaKBQA 框架需要通过指令调整在基于知识图谱的问答系统数据集集中的（自然语言问题，逻辑形式）对来高效微调开源大模型。微调后的大模型用于通过语义解析将新的自然语言问题转化为相应的候选逻辑形式。然后，ChatChaKBQA 在短语级检索这些逻辑形式中的实体和关系，并在将其转化为 SPARQL 后搜索可以针对知识库执行的逻辑形式。最后，通过转化获得的 SPARQL 用于获取最终的答案集，并实现对自然语言问题的可解释和需要知识的回答。

为了构建高效的指令微调训练数据，ChatChaKBQA 采取了一种创新的方法：首先，它将基于知识图谱的问答系统测试集中的自然语言问题对应的 SPARQL 查询转换为逻辑形式，并用实体标签替换其中的实体 ID 以便大型语言模型能够更清晰地识别和处理实体。接着，将这些自然语言问题和对应的逻辑形式配对，作为“输入”和“输出”，并附上指令“生成一个逻辑形式查询，以检索与给定问题相关的信息”，从而构建出用于微调大型语言模型的训练数据集。为了降低对参数众多的 LLMs 进行微调的成本，ChatChaKBQA 采用了参数高效微调技术。这些技术包括 LoRA，它通过低秩近似在微调过程中调整模型权重，有效减少了内存占用；QLoRA 则通过将梯度反向传播到一个冻结的 4 位量化模型，同时保持与全 16 位微调相当的性能；P-tuning v2 则通过在输入前的每一层添加可微调参数，实现了前缀微调；Freeze 技术则通过仅微调模型最后几层的全连接层参数，同时冻结其他参数，加速了模型的收敛过程。经过微调，大模型已经拥有了一定的自然语言问题转化为逻辑形式的语义解析能力。在检索阶段，ChatChaKBQA 采用无监督检索方法，通过短语级语义检索和替换，将候选逻辑形式转化为可在 KB 上执行的 SPARQL 查询。这一过程首先进行实体检索，然后是关系检索，每一步都通过计算相似度和排序来优化结果。最终，ChatChaKBQA 通过迭代候选逻辑形式列表并将其转换为 SPARQL 查询，执行查询以获取答案。这种方法不仅提供了自然语言问题的完整推理路径，还确保了答案的可解释性。

基于 ChatChaKBQA 框架，我在服务器上部署了零一万物 Yi-6B-Chat 模型，在微调过程中使用的数据集来自 Chinese Open Instruction Generalist (COIG) 相比其他的中文指令数据集，COIG 数据集在领域适应性、多样性、数据质量等方面具有一定的优势。模型为微调前后问答对比如图 4 所示。

=====Baseline: Greedy=====

Human: 你知道哪些中国传统节日和习俗吗? Assistant: 作为一个大语言模型, 我了解一些中国传统节日和习俗, 例如: 春节、清明节、端午节、中秋节、重阳节等。在这些节日中, 人们通常会举行庆祝活动, 如家庭聚会、祭祀祖先、吃传统美食等。此外, 还有许多其他的习俗和活动, 如舞龙舞狮、放孔明灯等, 这些活动都是为了庆祝节日并传承传统文化。

=====finetune: Greedy=====

Human: 你知道哪些中国传统节日和习俗吗? Assistant: 中国的传统节日和习俗众多, 以下是一些典型的例子:

1. 春节 (农历正月初一): 这是中国最重要的节日, 象征着新的一年的开始, 人们会举行各种庆祝活动, 如贴春联、放鞭炮、吃饺子、穿新衣、拜年等。
2. 清明节 (农历三月三日): 又称寒食节, 是中国传统的扫墓节日, 人们会去墓地祭扫, 缅怀先人。
3. 端午节 (农历五月初五): 端午节的主要习俗是赛龙舟和吃粽子, 相传是为了纪念战国时期楚国大夫屈原。
4. 中秋节 (农历八月十五): 中秋节的主要习俗是赏月、吃月饼、品酒、玩花灯等, 象征着阖家团圆。
5. 重阳节 (农历九月初九): 重阳节的主要习俗是登高

图 4 模型为微调前后问答对比

5 总结

本报告深入探讨了基于知识图谱的问答系统中的关键技术, 详细介绍了构建模板的问答方法、语义解析的问答方法, 以及基于神经网络的方法。这些方法各有优势和局限, 但共同推动了基于知识图谱的问答系统在特定领域和开放领域问题处理能力的提升。报告还探讨了构建知识图谱问答系统的主要技术, 包括三元组匹配、知识图谱嵌入、记忆网络和信息抽取, 这些技术为基于知识图谱的问答系统提供了强大的支持。基于知识图谱的问答系统面临着深化语义理解、知识图谱的动态管理、增强答案可解释性、实现多模态和跨领域问答、提升个性化和上下文感知、强化鲁棒性和泛化能力以及保障隐私和安全性等挑战^[10]。未来的研究应聚焦于这些关键领域, 以期实现系统的实时交互、多轮对话能力, 并拓展其在多样化应用场景中的适应性, 为用户提供更智能、高效和安全的信息服务。通过持续的技术革新, 特别是在深度学习与知识图谱融合方面的突破, 基于知识图谱的问答系统系统有望进一步发展。

参 考 文 献

- [1] 胡译文,孙建军,武夷山.国内知识图谱应用研究综述[J].图书情报工作, 2013, 57(3).DOI:10.7536/j.jssn.0252-3116.2013.03.024.
- [2] 王智悦,于清,王楠,等.基于知识图谱的智能问答研究综述[J].计算机工程与应用, 2020, 56(23):11.DOI:10.3778/j.issn.1002-8331.2004-0370.
- [3] Berant J, Chou A, Frostig R, et al.Semantic Parsing on Freebase from Question-Answer Pairs[C]//Empirical Methods in Natural Language Processing.Association for Computational Linguistics, 2013.
- [4] 张天成,田雪,孙相会,等.知识图谱嵌入技术研究综述[J].软件学报, 2023, 34(1):277-311.
- [5] Bordes A, Usunier N, Garcia-Duran A, et al.Translating Embeddings for Modeling Multi-relational Data[C]//Neural Information Processing Systems.Curran Associates Inc. 2013.
- [6] Kapashi D, Shah P .Answering Reading Comprehension Using Memory Networks[J].[2024-01-29].
- [7] Yao X, Durme B V .Information Extraction over Structured Data: Question Answering with Freebase[C]//Meeting of the Association for Computational Linguistics.2014.DOI:10.3115/v1/P14-1090.
- [8] Bordes A, Chopra S, Weston J .Question Answering with Subgraph Embeddings[J].Computer Science, 2014.DOI:10.3115/v1/D14-1067.
- [9] Luo, H., Haihong, E., Tang, Z., Peng, S., Guo, Y., Zhang, W., Ma, C., Dong, G., Song, M., & Lin, W. (2023). ChatKBQA: A Generate-then-Retrieve Framework for Knowledge Base Question Answering with Fine-tuned Large Language Models. ArXiv, abs/2310.0897
- [10] 闫悦,郭晓然,王铁君等.问答系统研究综述[J].计算机系统应用,2023,32(08):1-18.DOI:10.15888/j.cnki.csa.009208.