

申请上海交通大学博士学位论文

基于知识库的自然语言理解

论文作者 罗康琦  
学 号 0120339027  
导 师 朱其立教授  
专 业 计算机科学与技术  
答辩日期 2019 年 2 月 21 日



Submitted in total fulfillment of the requirements for the degree of Doctor  
in Computer Science and Technology Major

# Knowledge Base Empowered Natural Language Understanding

KANGQI LUO

Advisor

Prof. KENNY QILI ZHU

DEPART OF COMPUTER SCIENCE & ENGINEERING

SHANGHAI JIAO TONG UNIVERSITY

SHANGHAI, P.R.CHINA

Feb. 21st, 2019



## 基于知识库的自然语言理解

### 摘要

自然语言是人类进行信息交流和知识保存的重要工具，同时也是人机交互过程中最主要的形式。因此，让机器实现对自然语言的理解，是现阶段人工智能的重要发展方向，同时也是学术界的热门研究课题。自然界中存在的不同事物，以及事物之间的联系已是海量级别，随着互联网中以维基百科、IMDB 等数据库为首的结构化信息的大量积累，用于组织和维护开放领域中海量知识的大规模结构化知识库应运而生。它们以标准化的符号存储了千万以上的实体、以及十亿以上实体之间具有的关系，成为了语义表达的有效载体，同时也引出了一系列基于知识库的自然语言理解研究。因此，本文针对描述客观事实的自然语言文本，利用知识库实现多个维度的语义理解。

根据语义所体现的不同层次，本文从实体、关系和句子这三个层面研究自然语言理解问题。实体是语义中不可再分的元素，多个实体由关系互相连接构成基本事实，而句子往往包含着多个关系，具有更加复杂的整体语义。具体而言：实体层面的理解体现为直接匹配，将文本中代表实体的短语链接至知识库中的特定实体；关系层面的理解体现为结构匹配，将自然语言关系转换为由知识库关系（谓词）所构建的特定语义结构；句子层面的理解则对单一关系的结构匹配进行深入扩展，对于问句而言还体现为推理匹配，即根据语义结构，从知识库中寻找问句的正确答案。对于这些粒度的自然语言理解问题，需要使用不同的方法进行语义建模。

对于实体理解问题，其核心为计算实体短语的上下文信息与候选知识库实体间的匹配程度。经典的实体链接任务具有以下特点：候选实体数量庞大，实体短语普遍存在的一词多义性，以及候选实体之间存在相互依赖关系。本文中，我们关注对表格文本进行跨语言的实体链接任务，除了上述特点以外，表格文本所具有的半结构性，以及文本和知识库由不同的语言所描述，这给此任务带来了新的挑战。为此，我们提出了基于神经网络和跨语言词向量的链接模型，其优势在于：降低翻译过程带来的信息损失，学习表格行列方向的上下文和一致性特征，并通过联合训练框架提升整体链接质量。在跨语言和单语言两个场景上的实验表明，我们的模型有效捕捉表格中实体之间的特殊联系，同时在跨语言场景中具有稳定而良好的效果。

对于关系理解问题，其核心为用知识库中的结构描述自然语言中，一个二元关系的语义。该问题主要具有以下两个特点：首先自然语言关系同样存在多义性，其次关系和

知识库中的谓词存在语义间隔，难以实现简单的一一对应。基于这两个特点，我们对自然语言关系进行了两个粒度的语义建模。粗粒度的建模聚焦于关系的多义性，我们通过对知识库构建更加丰富的类型层次结构，挖掘一个二元关系的主语和宾语所具有的不同类型搭配，实验结果表明我们的模型效果优于传统的选择偏好模型。细粒度的建模旨在利用知识库实现对关系语义的精确表达，我们致力于使用人类能理解的图结构描述关系语义，提出了基于规则推导的模式图推理模型，以挖掘关系可能的复杂结构表示，并将其运用于知识库补全任务。实验结果显示，我们的模式图推理模型不仅具有高度可解释性，而且效果优于其它规则推导模型和新兴的知识库向量模型。

对于问句理解问题，我们着眼于基于知识库的自动问答任务，即在知识库中寻找代表答案的实体集合。由于问句包含了未知答案与其它实体的一个甚至多个关系，其语义变得更加复杂的同时，带来了如下挑战：如何描述问句的复杂语义，以及如何有效度量问句和语义结构之间的相似度。基于深度学习的语义匹配模型得到了广泛的研究，但这些模型所适用的语义结构存在限制，对复杂问题的回答存在瓶颈。为此，我们提出了针对复杂问题的深度学习语义匹配模型。该模型沿用关系理解中的图结构表示，首先生成问句可能对应的候选查询图，然后利用深度神经网络学习这些查询结构的整体语义表示，以此捕捉问句中不同语义成分的有机结合。实验结果表明，基于复杂查询图的深度学习模型在多个复杂问题和简单问题数据集上都具有良好的性能。

综上所述，本文从实体、关系、问句三个粒度出发，研究自然语言和知识库之间的语义理解与匹配问题。在实体理解中，我们提出了基于神经网络、跨语言词向量以及联合训练的链接模型，并用于解决跨语言场景中对表格文本进行的实体链接问题。对关系和问句的语义理解，我们始终贯彻语义建模的可解释性，使用主宾语类型搭配描述关系具有的多义性，以及使用基于知识库的图结构描述关系或问句的精确语义。对于自动问答任务，我们提出的深度学习模型实现了对复杂图结构的整体建模，得以充分体现其特征学习能力，更有效地度量问句与复杂结构的语义匹配程度。最后，希望本文的一系列工作能够对该领域今后的学术研究有所帮助。

**关键词：**知识库，自然语言理解，实体链接，知识库补全，自动问答，深度学习

# KNOWLEDGE BASE EMPOWERED NATURAL LANGUAGE UNDERSTANDING

## ABSTRACT

Natural language is an important tool for human information exchange and knowledge preservation, and also the most important form used in human-computer interaction. To make the machine better understand natural language (NL) becomes the main direction of artificial intelligence at this stage, and also the hot research topic in academic fields. There exists massive things and relations between things in the world. With the growing number of structural knowledge in the World Wide Web, such as Wikipedia and IMDB, researchers had developed large-scale structured databases to store, organize and maintain those massive facts in open domains, and we call them Knowledge Bases (KB). Knowledge bases make use of standardized symbols to store more than tens of millions of entities and more than one billion facts that exist between entities. Therefore, the KB becomes an effective carrier of semantic representation, leading to a series of KB-based semantic understanding research. In this dissertation, we use the KB to realize the semantic understanding of multiple dimensions for natural language texts that describe objective facts.

Considering different granularities of semantics, we conduct our research of the semantic understanding problem from three levels: entity, relation and sentence level. Entities are semantically indivisible elements, and multiple entities are connected by a relation, describing a single fact. While a sentence may contain several relations, thus holds a more complex semantics. For the understanding at the entity level, we directly mapping phrases in the text to specific entities in the KB. For the relation level, we attempt to represent natural language predicates by using specific structures made up of predicates in the KB. The understanding at the sentence level goes deeper than relation level, especially for questions, where we aim at automatically retrieving answers by the inference over the KB. Different methods are needed for the semantic modeling at different levels.

For entity understanding, the kernel part is to calculate the degree of matching between the phrase and the KB entity. The classical entity linking task has the following characteristics: large number of candidate entities, ambiguity of phrases, and the interdependence of entities

of different phrases. Moreover, we focus on the cross-lingual entity linking task for the text from web tables. In addition to the above features, how to leverage the semi-structure of table texts, and how to bridge the linguistic gap between texts and KBs, become the new challenges of the task. Thus, we propose a linking model based on neural networks and cross-lingual word vectors, which has the advantages of reducing the information loss caused by the translation process, learning the context and coherence features of the table row and column direction, and improving the overall link quality through the joint training framework. Experiments on both monolingual and cross-lingual scenarios show that our model effectively captures the special connections between entities in the table, and keeps a stable and good result.

For relational understanding, the core is to describe the semantics of a NL relation with structures in the KB. It has two characteristics: First, the NL predicate has ambiguity; second and different from entity understanding, there exists semantic gaps between NL and KB predicates, hence it is difficult to achieve a simple one-to-one mapping. Based on these two points, we attempt to model the semantics of a relation via two granularities. The coarse-grained modeling focuses on the ambiguity of NL predicates. By constructing a richer hierarchical structure for types in the KB, we mine the different type combinations of the subject and object that a NL predicate holds. Experimental results show that our model outperforms the traditional selectional preference model. The fine-grained modeling aims to precisely express the semantics of relations using the KB. We use the human-understandable graph structure to describe the NL predicate, and propose a rule induction based inference model, which is able to express the complex semantics of NL predicates via schema graphs. We apply the structural representations to the knowledge base completion task, and the experimental results show that our schema graph inference model is not only highly interpretable, but also outperforms other rule induction model and the emerging knowledge base embedding model.

For question understanding, we focus on the task of knowledge base question answering, that is to retrieve the answer entity set of the question from the KB. There exists one or more relations between the unknown answer and the other entities in the question, which brings a more complex semantics as well as the following challenges: how to describe such complex semantics, and how to effectively measure the similarity of the question and candidate semantic structures. The semantic matching model based on deep learning has been widely studied, but usually these target structures are limited, hence answering complex questions becomes a bottleneck of previous work. To this end, we propose the deep semantic matching model for answering complex questions, which follows the idea of graph based semantic representation

used in relation understanding. We first generate candidate query graphs of the question, then encode such complex query structure into a uniform vector representation via deep neural networks, thus successfully capture the interactions between individual semantic components within a complex question. Experiments on multiple QA datasets show that our approach consistently outperforms existing methods on complex questions while staying competitive on simple questions.

In summary, starting from the three granularities of entity, relation and question, this paper studies the problem of semantic understanding and matching between natural languages and knowledge bases. For entity understanding, we propose the link model based on deep neural networks, cross-language word vectors and joint learning scheme, for solving the entity linking task of tabular texts in cross-lingual scenarios. For the understanding of both relations and questions, to keep the interpretability of semantic modeling, we use the type combination of subjects and objects to describe the ambiguity of the relation, and use the graph structure based on the KB to describe the exact semantics of the relation or the question. For the task of question answering, our proposed deep learning model aims at the encoding of the entire query structure, which makes better use of the ability of feature learning, and more effectively measures the matching level between questions and complex semantic structures. Finally, hoping our work in this paper can help future academic researches in this field.

**KEY WORDS:** knowledge bases, natural language understanding, entity linking, knowledge base completion, question answering, deep learning



# 目 录

<b>插图索引</b>	<b>XI</b>
<b>表格索引</b>	<b>XIII</b>
<b>算法索引</b>	<b>XV</b>
<b>第一章 引言</b>	<b>1</b>
1.1 研究背景 . . . . .	1
1.2 研究现状 . . . . .	6
1.3 主要工作和贡献 . . . . .	10
1.3.1 实体理解问题 . . . . .	10
1.3.2 关系理解问题 . . . . .	10
1.3.3 问句理解问题 . . . . .	11
1.4 论文结构 . . . . .	12
<b>第二章 国内外相关研究综述</b>	<b>13</b>
2.1 实体理解：实体链接任务 . . . . .	13
2.1.1 基于特征工程的实体链接 . . . . .	15
2.1.2 基于深度学习的实体链接 . . . . .	16
2.1.3 跨语言词向量 . . . . .	18
2.2 关系理解：知识库补全任务 . . . . .	20
2.2.1 基于规则推导的模型 . . . . .	21
2.2.2 基于知识库向量的模型 . . . . .	22
2.3 问句理解：知识库自动问答任务 . . . . .	25
2.3.1 基于语义解析的问答模型 . . . . .	26
2.3.2 基于信息抽取的问答模型 . . . . .	30
2.4 本章小结 . . . . .	32
<b>第三章 跨语言的表格实体链接研究</b>	<b>35</b>
3.1 概述 . . . . .	35
3.2 相关工作 . . . . .	37

3.3	任务规范定义 . . . . .	38
3.4	我们的方法 . . . . .	39
3.4.1	候选实体生成 . . . . .	39
3.4.2	向量表示及跨语言模块 . . . . .	40
3.4.3	指示特征与上下文特征 . . . . .	41
3.4.4	一致性特征 . . . . .	41
3.4.5	训练及测试 . . . . .	42
3.4.6	模型实现细节 . . . . .	44
3.5	实验 . . . . .	45
3.5.1	实验设置 . . . . .	45
3.5.2	基线模型 . . . . .	45
3.5.3	实验结果 . . . . .	46
3.5.4	模型分析测试 . . . . .	49
3.6	本章小结 . . . . .	51
<b>第四章 自然语言关系的语义理解研究</b>		<b>53</b>
4.1	关系的主宾语类型搭配挖掘 . . . . .	53
4.1.1	引言 . . . . .	53
4.1.2	我们的方法 . . . . .	55
4.1.3	实验 . . . . .	59
4.2	关系的结构化语义挖掘 . . . . .	61
4.2.1	概述 . . . . .	62
4.2.2	相关工作 . . . . .	64
4.2.3	任务定义 . . . . .	66
4.2.4	我们的方法 . . . . .	67
4.2.5	实验 . . . . .	70
4.3	本章小结 . . . . .	77
<b>第五章 面向复杂语义的知识库自动问答研究</b>		<b>79</b>
5.1	概述 . . . . .	79
5.2	相关工作 . . . . .	82
5.3	我们的方法 . . . . .	83
5.3.1	分阶段查询图生成 . . . . .	83
5.3.2	基于神经网络的语义匹配模型 . . . . .	86

5.3.3 实体链接扩充 . . . . .	88
5.3.4 问答系统整体训练及预测 . . . . .	89
5.4 实验 . . . . .	90
5.4.1 实验设置 . . . . .	90
5.4.2 端对端实验比较 . . . . .	90
5.4.3 模型分析 . . . . .	91
5.5 小结 . . . . .	95
<b>第六章 总结与展望</b>	<b>97</b>
6.1 论文工作总结与主要贡献 . . . . .	97
6.2 未来工作展望 . . . . .	99
<b>参考文献</b>	<b>101</b>
<b>致 谢</b>	<b>113</b>
<b>攻读学位期间发表的学术论文</b>	<b>115</b>
<b>攻读学位期间参与的项目</b>	<b>117</b>



## 插图索引

1–1 Freebase 缩略图。 . . . . .	4
1–2 维基百科中的实体与表格链接。 . . . . .	5
1–3 搜索结果页面的右侧显示了当前实体的信息框。 . . . . .	7
1–4 搜索引擎精确返回复杂问题的答案。 . . . . .	8
1–5 实体、关系、句子语义理解之间的级联关系。 . . . . .	8
2–1 Limaye 等人提出的表格链接任务。 <sup>[51]</sup> . . . . .	14
2–2 基于多粒度卷积神经网络的实体链接模型。 <sup>[17]</sup> . . . . .	17
2–3 基于神经张量层的链接模型。 <sup>[18]</sup> . . . . .	18
2–4 英语和德语间的跨语言词向量例子。 <sup>[21]</sup> . . . . .	19
2–5 多种知识库向量模型示意图。 <sup>[30, 80]</sup> . . . . .	23
2–6 例句“what state borders texas”的多种解析结构。 . . . . .	27
2–7 HR-BiLSTM 模型。 <sup>[49]</sup> . . . . .	29
2–8 CDSSM 模型。 <sup>[94]</sup> . . . . .	30
2–9 QASE 模型。 <sup>[95]</sup> . . . . .	31
2–10 MCCNN 模型。 <sup>[96]</sup> . . . . .	32
3–1 中文表格到英文知识库的跨语言链接示例。 . . . . .	36
3–2 基于神经网络的联合训练模型示意图。 . . . . .	40
3–3 微观准确率随候选实体数量 $N_{cand}$ 的变化情况。 . . . . .	48
4–1 二元关系模式挖掘的流程框图。 . . . . .	55
4–2 二元关系“has grandfather”的语义表示。 . . . . .	63
4–3 模式图的一般形式。 . . . . .	66
4–4 “has father”模式图挖掘示例。 . . . . .	68
4–5 不同的规则推导系统对四个复杂关系生成的代表性结构。 . . . . .	72
5–1 一个具有复杂语义的问句示例。 . . . . .	80
5–2 分阶段候选图生成的具体例子。 . . . . .	84
5–3 语义匹配模型的整体结构 . . . . .	87



## 表格索引

3–1 候选生成步骤的 Hits@ $n$ 测评结果。 . . . . .	46
3–2 跨语言表格链接的测试结果，基线模型仅使用百度翻译工具。 . . . . .	47
3–3 中文环境下的表格链接准确率。 . . . . .	49
3–4 不同特征组合在验证集上的跨语言链接准确率。 . . . . .	50
3–5 不同模型训练方式在测试集上的跨语言链接准确率。 . . . . .	50
4–1 ReVerb 三元组的实体链接实验结果。 . . . . .	60
4–2 二元关系模式推理的评测结果。 . . . . .	61
4–3 生成的二元关系模式举例。 . . . . .	61
4–4 模式图列表的 AvgSc@ $n$ 测评结果。 . . . . .	73
4–5 在 PATTY-100 上进行主宾语预测的测评结果。 . . . . .	74
4–6 在 FB15k-37 上进行主宾语预测任务的测评结果。 . . . . .	75
4–7 三元组分类任务的 MAP 测评结果。 . . . . .	76
5–1 预测最佳查询图所使用的特征。 . . . . .	89
5–2 CompQ 和 WebQ 数据集上的实验结果，评价指标为平均 $F_1$ 分数 . . . . .	92
5–3 SimpQ 数据集上的语义匹配测试结果 . . . . .	93
5–4 对谓词表示的分析结果。 . . . . .	93
5–5 问句表示和语义组合的分析测试。 . . . . .	94



## 算法索引

3-1 基于局部搜索下降的预测过程 . . . . .	43
4-1 复杂模式图搜索 . . . . .	69



## 第一章 引言

本文主要的研究对象是自然语言中的单词、短语以及句子。为了使机器可以像人类一样，理解一个句子背后的含义，而不是仅仅停留在识别字面上的不同词汇，我们需要让机器挖掘出句子中的两类信息，即句子里的实体（人物、地点、组织、事件等），以及实体之间的关系。考虑到自然语言描述方式的多样性，我们需要使用标准化的数据库作为语义理解的载体，其包含自然界中所有已知的实体，以及它们之间存在的关系，这样的结构化数据库被称为知识库。本文研究的问题，就是将自然语言中的实体、关系以及句子的语义，映射到知识库的过程。

### 1.1 研究背景

人类的进化总是伴随着知识的不断积累。传统的知识存储媒介为纸质书籍，随着互联网的发展以及分布式存储系统的成熟，越来越多人类知识以电子文档的形式进行存储。截至 2018 年 6 月，Google 搜索引擎已索引大约 470 亿个网页<sup>1</sup>，考虑到大量未被索引的网页以及未被电子归档的书籍资料，人类所拥有的知识远不止统计数据所显示的规模。并且随着新事件的发生，新的知识也还在不断涌人。身处信息化时代，人类和计算机的交互变得频繁，且交互方式呈现多样化。而在这其中，自然语言一直是最重要的交互方式，主要体现为文本和语音的形式，与人类之间的正常交流最为接近。人工智能成为当今科研的热门方向，由于计算机已经拥有海量的非结构化文本数据，因此人类期望智能化的机器能够感知并掌握人类的知识，从而更好地与人类进行自然语言交互。科幻电影中经常安排了这样的机器人角色：回答人类的提问，分析人类的情感，进行持续的聊天对话等等。这些诉求支撑了自然语言处理领域的蓬勃发展。

自然语言理解是人工智能的一个重要分支。为了衡量机器在自然语言上的智能化水平，早在 1950 年，英国科学家阿兰·图灵就提出了著名的思想实验“图灵测试”，即让机器作为被测试者，仅通过文本与人类测试者对话，并说服人类自己是人而不是机器。若机器通过了图灵测试，则意味着其智能化已达到接近人类的水平。图灵测试属于开放领域的通用对话场景，目前人工智能水平还远远不能达到这样的高度，通过含糊其辞的策略进行对话可以做到欺骗人类测试者，但这并非真正理解对话含义。

之所以机器难以理解人类的语言，是因为自然语言本身具有很高的复杂性，主要体现为两个方面。一方面，自然语言中的不同词汇和语义之间不具有一对一关系。人类语

<sup>1</sup><http://www.worldwidewebsize.com/>

言并非遵循某种确定的规则产生，而是随着时间在不停演变，例如“苹果”一词原本仅指代一种水果，而苹果公司的出现，使得自然语言文本中的“苹果”有了明显的歧义。可见，词汇和语义的对应来自约定俗成，并不需要满足严谨的区分度。这使得同义词和一词多义成为了自然语言中的普遍现象。另一方面，词汇的排列顺序也在影响着语义。自然语言中存在着层次关系，多个词汇组成短语，多个短语组成句子等等。组合而成的语义并不等于各个部分的简单叠加，例如“深度学习”与“学习深度”两个短语具有较大的语义差别。

人工智能离完美的语义理解还有很长的距离，现阶段也很难设计一套语言理解系统，使其适用于各种不同的自然语言场景。在不同类型的文本中，描述客观事实的自然语言文本数量庞大，以维基百科为代表的语料库凝结了人类在各种领域所拥有的知识。与此同时，相比于带有主观信息的句子，客观事实所具有的语义更加明确，不受个人感情色彩的影响，因此能够更加准确地评判机器的理解能力好坏。鉴于以上两点原因，如何让机器更好地理解客观事实紧密相关的自然语言信息，成为了我们的研究重心。

对于机器而言，怎样才算理解自然语言中的客观事实？我们以例句“苹果于 1997 年收购了 NeXT 公司，乔布斯回归并担任临时 CEO。”进行阐述，这段文字讲述了与苹果公司相关的一些客观事实。机器理解的前提，在于能从非结构化的文本中抽取出语义信息。计算机对它的信息抽取主要包含以下两个层次。

浅层次的抽取，体现在识别句子描述的是关于**哪些事物**的客观事实，不仅需要找出代表它们的短语，而且要正确对应到客观存在的事物。例如识别出短语“苹果”指的是苹果公司，而不是水果。通常需要识别的事物为命名实体，即具体的人名、地名、组织名等等，有时也包括一些抽象概念，比如“森林”、“CEO”、“物理学家”等等。我们将这些客观存在的具体或抽象事物统称为**实体**。

更深层的抽取在于，识别不同实体之间具有怎样的**关系**。例如苹果和 NeXT 之间存在着收购关系，苹果和乔布斯之间存在任命关系等。在一个句子中，描述两个实体之间关系通常会采用主谓宾的形式，也存在着其它形式（主系表，介宾结构等）。为了方便论述，我们将联系两个实体的二元关系表示为（主语，谓语，宾语）三元组的形式，其中关系在三元组中充当谓语成分，主语和宾语则是关系的两个参数实体。

判断一个人的智力高低，不仅在于他见过多少实体或了解多少关系，而且在于能否将已知的事实再大脑中进行整合为知识，从而能够举一反三，在遇到问题时，寻找出和问题匹配的事实，对面对复杂问题的情况，还能结合多个事实进行推理回答。问句“NeXT 公司被谁收购？”可以直接与例句中的关系匹配，而要回答“NeXT 公司被收购时，美国总统是谁？”，则需要一定的推理能力。对于机器而言，这样的自动问答任务既是人机交互中的重要场景，同时也是衡量句子语义理解是否智能化的有力参照。

综上所述，针对以客观事实为主的自然语言理解，我们从实体、关系、句子的语义理解这三个角度出发，展开一系列研究。它们的共同点是，需要一种手段来整合并维护人类的客观知识，包括实体、概念、类型，以及联系它们的关系等。计算机虽然存储了海量文本，但杂乱的非结构化文本显然无法胜任。

信息抽取技术直接针对非结构化文本，从中提炼出有价值的客观事实，组成三元组形式( $arg_1, relation, arg_2$ )，本文中称之为关系三元组，或关系实例。其中 $arg_{1,2}$ 分别代表关系 $relation$ 的主语和宾语。例如可以从句子“*Mozart was born in Salzburg, but moved to Vienna in 1781*”中抽取出两个三元组(Mozart, was born in, Salzburg)以及(Mozart, moved to, Vienna)。早期的信息抽取主要针对特定领域的文本数据，而近年来提出的开放领域信息抽取(Open Information Extraction, OpenIE)系统<sup>[1-4]</sup>则从不限定领域的海量文本中提取不同的二元关系。然而，关系三元组中的每一个成分为字符串，同义词和一词多义现象依然存在。为了消除歧义，计算机需要通过更加规范的方式，表示不同的实体、关系和具体的三元组，而不是只停留在字符串层面。

对于实体的表示，维基百科是一个优秀的载体，英文维基百科具有一千万以上的实体(页面)，同时实体命名遵循特定规则，利用括号信息区分名字相同的实体。同时维基百科还维护了分类信息(Category)，具有相似特点的实体会被归为同一分类。普林斯顿大学设计的WordNet<sup>[5]</sup>数据库更加关注概念上的区分，其中同义词集(Synset)是用来表示概念的基本单位，具有多义的单词(或词组)指向多个同义词集，而每个同义词集由一系列近义词构成，并配有简短文字解释概念词义，并且同义词集之间包含着丰富的上下位关系。在关系方面，以PropBank<sup>[6]</sup>为代表的数据集以动词分析为主，对它们的词义进行了归类。与WordNet类似，PropBank使用同义动词集作为词义的基本单位，并且对动词的参数在语境中所扮演的角色进行标记，以此表示这些同义动词的用法，因此PropBank常用于语义角色标注任务中。

以上的数据集，除维基百科以外，其余都还停留在单独的实体、概念、关系层面，没有维护实体的属性值，以及实体之间的具体关系。维基百科页面中存在大量具有特定模板的信息表格(Infobox)，以半结构化的形式描述了实体的属性，以及和其它实体的关系。由于维基百科面向人类读者，半结构化信息较难直接被机器理解。为此，一系列研究工作旨在将维基百科中的半结构化知识进行组织，形成的结构化数据库包含了与实体相关的大量事实，我们称之为知识库(Knowledge Base)，或知识图谱(Knowledge Graph)。

具有代表性的知识库包括DBpedia<sup>[7]</sup>，YAGO<sup>[8]</sup>和Freebase<sup>[9]</sup>，它们皆诞生于2007年。DBpedia的构建过程基于对维基中的信息表格、页面分类、外部链接等半结构化信息进行自动抽取，可以看作是最纯粹的结构化维基百科；YAGO的主要信息同样来自对

Infobox 内容的自动抽取，同时将页面分类信息与 WordNet 严谨的概念层次关系进行融合，构建出代表实体类型的层次关系；由 MetaWeb 公司开发的 Freebase 集成了维基百科、IMDb<sup>1</sup>、MusicBrainz<sup>2</sup>等多个数据库的知识，并提供接口，允许用户对 Freebase 的结构化内容进行编辑或添加新的知识，因此相比其余知识库，Freebase 具有更大的规模。这些结构化知识库的共同点在于使用资源描述框架（Resource Description Framework），每一条知识都由 SPO 三元组表示，即  $(subject, predicate, object)$  形式，其中 *subject* 和 *object* 为知识库中的节点，代表着不同的实体、类型或属性值，*predicate* 为连接不同节点的边，代表实体间的关系，或实体的属性，统称为谓词。在知识库中，不同的节点和边具有独立的编号，因此不具有歧义。为了与 OpenIE 的关系三元组区分，我们将知识库里的 SPO 结构称为事实三元组。这些知识库内的所有事实三元组构成了庞大的图结构，即“知识图谱”名称的由来。

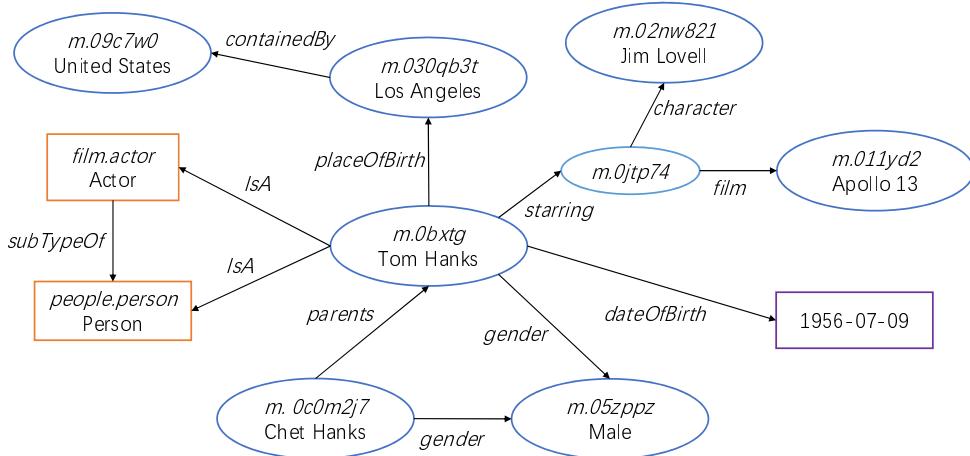


图 1-1 Freebase 缩略图。

Figure 1-1 A snapshot of Freebase.

由于上述知识库结构相似，研究方法具有普适性，因此我们的工作主要基于资源最丰富的 Freebase，包含至少四千万个不同实体，三千种以上的常用类型，六千种以上的常用谓词，以及十亿以上的关系三元组。按功能划分，Freebase 中的事实三元组可归为三类：描述实体和类型间层次结构的（实体，*IsA*，类型）和少量（类型，*subTypeOf*，类型）三元组；描述实体间的关系的（实体，关系，实体）三元组；描述实体自身属性的（实体，属性，属性值）三元组，其中属性值为整数、浮点数、时间或字符串。Freebase 的一个缩略图如图1-1所示，不同的实体、类型、谓词都有独立的编号，例如谓词编号

<sup>1</sup><https://www.imdb.com><sup>2</sup><https://musicbrainz.org>

*type.object.type* 代表 *IsA* 关系, *type.object.name* 代表名称属性。同时它们都具有唯一的名称属性值, 部分实体还具有多个别名属性 (*common.topic.alias*) 值。例如实体 *m.02\_286* 的名称为 “New York City”, 具有别名 “The Big Apple”、“Empire City” 等; 谓词 *location.location.contained\_by* 名称为 “Contained By”, 描述了地点实体间的包含关系。Freebase 支持使用 SPARQL 语句进行结构化查询, 以类似 SQL 的语法从图结构知识库中筛选出满足查询条件的实体。

接下来, 我们将关注在实体、关系、句子这三个递进层面上, 以知识库作为语义载体的一些自然语言理解问题。

首先, 实体层面的语义理解体现为实体链接任务, 即从自然语言文本中寻找出代表实体的短语, 并匹配到知识库中的特定实体。如同维基百科编辑者在页面中会添加一些超链接文本, 并指向其它实体页面, 使得读者可以快速了解与当前页面相关的实体信息, 自动化的实体链接可以应用于开放领域的自然语言文本, 从而实现消歧义的目标。文本输入可以是非结构化的文本, 也可以是半结构化形式, 例如互联网中存在的表格。图1-2展示了维基百科中关于赛车的一个页面, 其中纯文本和表格中的内容均被添加了超链接, 指向了特定的车手、车队等实体页面。



The screenshot shows a Wikipedia article page. At the top, there's a sidebar with sections like "Results" and "Grands P...". The main content area has several paragraphs of text with blue links underlined. In the center, there's a large image of Sebastian Vettel. To the right of the image is a detailed sidebar with information about Scuderia Ferrari S.p.A. Below the sidebar is a table titled "Grand Prix" with columns for Round, Grand Prix, Pole position, Fastest lap, Constructor, and Report.

Round	Grand Prix	Pole position	Fastest lap	Constructor	Report	
1	Australian Grand Prix	Lewis Hamilton	Daniel Ricciardo	Sebastian Vettel	Ferrari	Report
2	Bahrain Grand Prix	Sebastian Vettel	Valterri Bottas	Sebastian Vettel	Ferrari	Report
3	Chinese Grand Prix	Sebastian Vettel	Daniel Ricciardo	Daniel Ricciardo	Red Bull Racing-TAG Heuer	Report
4	Azerbaijan Grand Prix	Sebastian Vettel	Valterri Bottas	Lewis Hamilton	Mercedes	Report
5	Spanish Grand Prix	Lewis Hamilton	Daniel Ricciardo	Lewis Hamilton	Mercedes	Report
6	Monaco Grand Prix	Daniel Ricciardo	Max Verstappen	Daniel Ricciardo	Red Bull Racing-TAG Heuer	Report
7	Canadian Grand Prix	Sebastian Vettel	Max Verstappen <sup>[note 6]</sup>	Sebastian Vettel	Ferrari	Report

图 1-2 维基百科中的实体与表格链接。

Figure 1-2 Entity and table linking in Wikipedia.

其次, 对于关系层面的语义理解, 一个传统任务为关系分类任务: 给定包含两个实体的句子, 将它们之间的关系归类至预定义的关系类别中。然而关系分类任务较难扩展

到开放领域，一方面，关系类别增多的同时，任务数据集的标注代价显著提升，另一方面，不同于实体匹配，知识库谓词和自然语言关系可能存在一定偏差，难以直接分类。在这样的前提下，关系理解的实质在于，如何利用知识库的已有谓词信息，学习目标关系语义。这引出了知识库补全（Knowledge Base Completion）任务：由于知识库中的事实可能存在缺失，对于其中的目标谓词，能否推理出它和其它谓词之间的联系，从而自动添加新的事实。例如知识库中，某人的国籍缺失，那么可以根据其出生地所在的国家进行推测。由于知识库中的事实三元组和自然语言的关系三元组具有类似形式，因此目标谓词可以从知识库扩展到自然语言中。

最后，对于句子的语义理解，为了能充分利用知识库的信息，我们关注描述客观事实的问句，通过自动问答任务衡量其语义理解能力。自动问答具有广泛的应用场景，搜索引擎便是其中之一。传统的搜索引擎工作方式依靠信息抽取相关算法，比较用户查询与网页的相似度，主要利用词级别的共现模型，如 TF-IDF<sup>[10]</sup>，或词级别的语义匹配模型，如 LSI<sup>[11]</sup>, pLSA<sup>[12]</sup> 等。但对于用户输入的复杂问题，词级别的匹配难以直接定位到用户想要的答案或网页。更加智能化的搜索引擎，则尝试对问题进行推理，并根据已有的知识库，将答案定位至已知的实体中。Google 建立了以 Freebase 为基础的知识库（Google Knowledge Graph），当用户搜索一些名词时，如图1-3中，用户搜索“george w bush”，结果页面右侧会显示实体的信息框，包括其属性，以及与其关联的其它实体等。而如图1-4所示，当用户输入较为复杂的问题“who was president of US when beijing olympics was held”，搜索引擎能在结果页面上方显示出自动回答的结果。对比下方的传统页面搜索结果可见，精确返回答案能大大提升查询过程的用户体验。

实体、关系、问句的语义理解研究之间，存在着紧密的内在联系。如图1-5所示，一个具体的问句通常包括一至多个关系，因此问句理解可以看做关系理解的扩展，即在问句中寻找疑问词和多个已链接实体间存在的不同语义。而作为自然语言理解的底层部分，实体理解是另外两者的基础，它为关系理解提供了消歧义的三元组，也为问句理解划定了大致的语义范围，由此可见，实体、关系、句子的语义理解具有级联关系，也是本文从这三个方向进行展开研究的动机。

## 1.2 研究现状

上一节介绍了本文关心的语义理解问题，本节中，我们延续实体、关系、问句语义理解这三部分，回顾已有的一些相关工作。

实体链接的研究开始较早，Mihalcea 等人<sup>[13]</sup>于 2007 年的研究是以维基百科为载体进行链接的鼻祖工作。对自然语言文本进行实体链接，主要分为两个步骤：挖掘所有代表实体的短语，以及将短语映射至知识库中的特定实体。短语挖掘可以通过字符串模糊

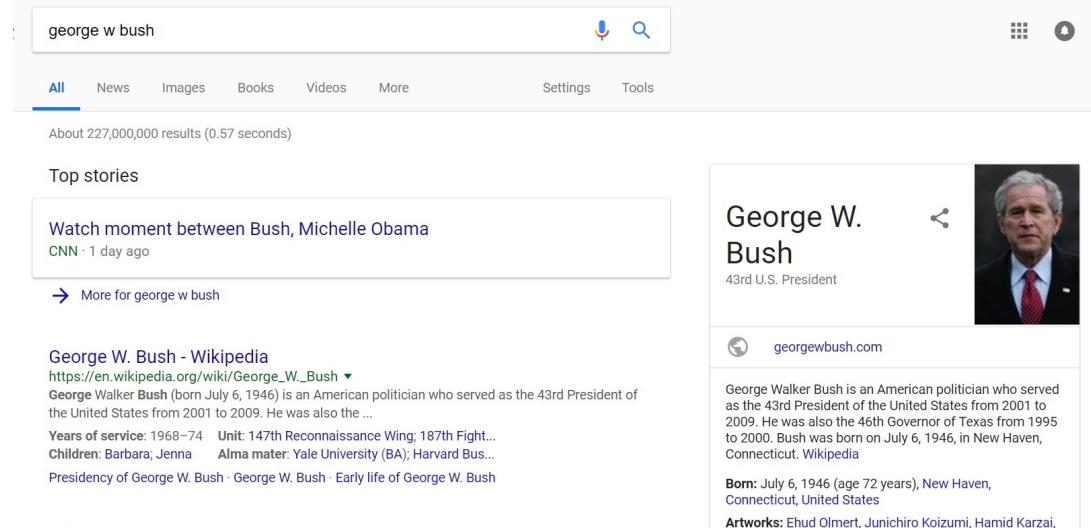


图 1-3 搜索结果页面的右侧显示了当前实体的信息框。

Figure 1-3 The infobox at the right side of search result pages.

匹配的方式进行收集，以保证较高的召回率。之后的映射步骤则是实体链接模型的重点，为了实现消歧义，需要利用短语所在文本的上下文特征，以及多个短语所映射的实体之间的关联程度，对于表格形式的文本输入，行列间实体所具有的特性也不可忽略。根据以上观察，以特征工程为核心的机器学习模型被运用于此，涵盖的特征主要包括基于维基内部超链接统计的先验概率，基于 TF-IDF 模型<sup>[14]</sup>的短语和实体的上下文相似度，基于 PMI<sup>[15]</sup>、WLN<sup>[16]</sup>等以维基共现频率衡量的不同实体间的相关度，等等。考虑到特征设计耗费人力，且与特定任务高度相关，更新的工作对基于深度学习的实体链接模型进行了研究，模型依赖神经网络建立实体和短语上下文的特征表达，并计算向量表达之间的相似度衡量短语和实体的匹配程度。以文献 [17] 为代表，对于输入文本中的短语和维基百科中的实体，模型可以关注不同粒度的上下文，利用卷积神经网络或循环神经网络对文本进行建模。同时模型可以学习维基百科或知识库中，实体分类、类型等信息的向量表达，以此丰富实体的语义特征，例如文献 [18-20]。此外，若文本和知识库的语言不同，则为跨语言场景的实体链接。通过翻译工具可以转化为单语言的实体链接，但受制于翻译步骤的准确率，因此主要的模型使用了跨语言的词向量技术 [21]，将不同语言下的单词映射至同一连续语义空间。

关系语义学习的研究，主要针对三元组级别，给定目标关系或谓词，根据它所已知的三元组信息，对其语义进行建模。按照关系语义的表示方法进行划分，主要研究可以分为规则推导和知识库向量学习两类。基于规则推导的模型中，二元关系或谓词  $p$  等同于布尔函数  $p(x, y)$ ，对于给定目标，模型旨在推导出由其它谓词构成的一阶逻辑表



图 1-4 搜索引擎精确返回复杂问题的答案。

Figure 1-4 The search engine precisely returns the answer of the complex question.

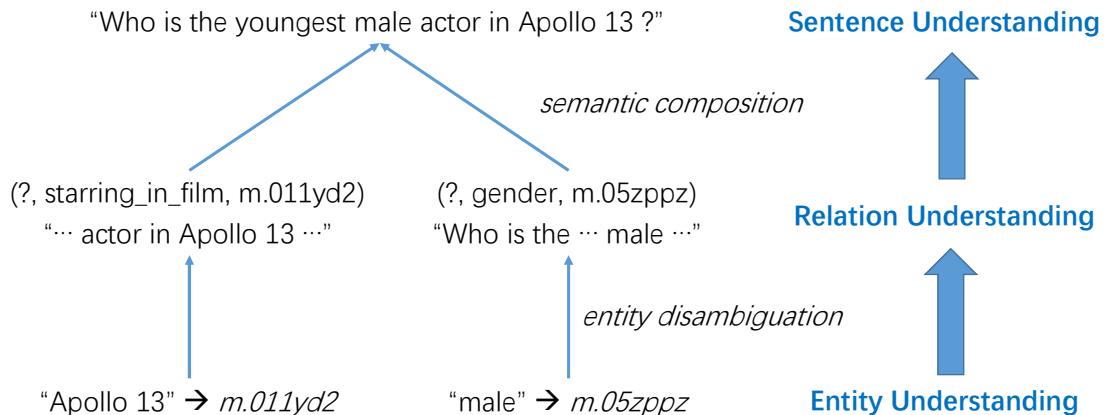


图 1-5 实体、关系、句子语义理解之间的级联关系。

Figure 1-5 The cascaded relationship between entity, relation and sentence understanding.

达式，单个表达式的语义具有确定性，同时人类可直接理解其语义表示，具有很高的可解释性。早期研究以 AMIE 模型<sup>[22]</sup> 为代表，挖掘具有高置信度的规则，后续的改进研究着眼于挖掘多种可能的规则，并赋予不同权重或概率，丰富语义表达能力，例如基于 MLN 模型的文献 [23, 24]，以及生成负样本，对大量路径形式的规则进行特征学习的 PRA 模型<sup>[25]</sup> 和 SFE 模型<sup>[26]</sup>。另一个分支为知识库向量模型，则依据已有的大量三元组信息，学习每一个实体和谓词的连续向量（或矩阵）表示，并通过实体和谓词表示之间

的代数运算，判断任意一个三元组事实的置信度。具体置信度定义方式不同，ER-MLP 模型<sup>[27]</sup> 基于简单的多层神经网络；RESCAL 模型<sup>[28]</sup> 基于实体向量和谓词矩阵表示的双线性运算；TransE 模型<sup>[29]</sup> 基于主谓宾向量之间的距离度量，并衍生出一系列改进模型 [30-32]。这些方法能够更充分地利用海量三元组信息进行建模，但是对谓词的语义缺乏直观解释。

对于问句理解与推理的研究，根据知识来源和答案形式的不同，可以分为两类任务：基于结构化知识，答案通常为实体、数值等简单形式的知识库问答（KBQA）；以及基于非结构化文档，答案表示为自然语言文本片段的检索式问答（IRQA）。后者不是我们的研究重点，现有的研究工作主要使用阅读理解模型 [33, 34]，从已有文档中抽取出最适合的文本片段作为答案。与关系学习类似，根据对问句语义的表示形式，解决知识库问答任务的方法也可分为两类。第一类方法在相关文献 [27, 35-37] 中称为“基于信息抽取的方法”<sup>1</sup>，它们遵循端到端问答的思路：对每个问句生成一定量候选答案后，机器学习模型用来解决二分类问题，即判断每个 < 问句，候选答案 > 对是否正确。模型使用的特征来自于候选实体在知识库中的信息（包括名称，具有的类型，直接相连的谓词，相邻实体等）与问句中不同单词的交互，早期模型使用特征工程，最新的研究则聚焦于利用神经网络进行特征表示学习。这类模型的优点在于实现简单、训练数据容易获取，但缺点在于可解释性较差，大量抽象特征的存在，使得人类难以理解某个候选答案被分类为正确或错误的根本原因。第二类方法称为语义解析（Semantic Parsing），答案实体的选取依赖于对问句语义的结构化建模，将其表示为知识库中的实体、类型、谓词组成的一阶逻辑表达式。只需要将其转换为知识库上的结构化查询语言，例如 SPARQL，即可查询出所有满足语义的答案，因此具有很明确的可解释性。这类方法通常先根据问句信息，生成少量候选查询结构，再利用机器学习算法训练问句和查询结构间的语义匹配模型，因此整体效果取决于这两个步骤的质量。查询结构生成方面，可依靠 CCG<sup>[38, 39]</sup>、DCS<sup>[40, 41]</sup> 等语法进行自底向上构建，或者依赖深度优先搜索由简到繁生成查询结构<sup>[42, 43]</sup>，还可以基于固定模板进行生成<sup>[44, 45]</sup>，以牺牲灵活性为代价，换取更有针对性的语义表达。语义匹配计算方面，传统机器学习方法从候选结构的生成步骤中提取特征，并与问句进行特征组合，以实现模型训练；与此同时，深度学习方法通过学习知识库中不同实体、谓词等向量表示，遵循“编码—比较”框架，将问句和查询结构的语义特征映射至同一空间，并计算向量间的相似度。基于深度学习的问答模型在近几年研究异常火热，相关文献 [46-50] 等在简单语义的问答场景中不断取得突破。但对于更加复杂的问题，这些模型并不能直接学习一个复杂查询结构的整体语义表达，难以充分体现深度学习强大的特征表示能力。

<sup>1</sup>该名称容易与检索式问答产生混淆，这里讨论的是 KBQA 的一类方法，而不是 IRQA。

## 1.3 主要工作和贡献

本节中，我们具体介绍本文的主要工作，包括实体、关系、问句理解三方面的研究内容，技术难点，以及我们的贡献。

### 1.3.1 实体理解问题

对于实体的理解，我们聚焦于跨语言场景下的表格实体链接任务，即以非英文编写的表格作为输入，将每个单元格对应的实体链接至英文知识库。我们以该任务作为研究对象，主要有两个动机：首先，英文知识库规模最为庞大而全面，有助于让不同语言的人类共同理解知识；其次，表格行列间通常存在着特定的关系，因此相比纯文本中的实体链接，表格链接能更有效地帮助知识库进行知识扩充。本文是学术界首次对该任务进行具体研究，具有以下两个主要特点：1) 短语和目标实体处于不同语言中，必须在字面相似之外，寻找候选实体收集和匹配度量的方式；2) 每个单元格和目标实体间的匹配可以体现在多个粒度，例如单元格自身，以及其所在的行列上下文，同行列的实体之间还存在明显的关联性。

为了解决跨语言表格链接问题，我们提出了基于神经网络和跨语言词向量的链接模型。该模型首先通过已有翻译工具进行候选实体的收集，但匹配过程并不依赖唯一的翻译结果；其次，模型利用多语言词向量的训练，将不同语言中的实体和短语映射至相同维度空间，实现最基本的匹配度量；再次，模型利用深度神经网络学习表格与目标实体间不同粒度的匹配特征，并提出了基于方差计算的一致性特征，以捕捉同列实体所具有的同质性；最后，模型基于联合训练思路，以优化整张表格的匹配程度为目标，进一步提升整体链接质量。在跨语言和单语言两个场景上的实验表明，我们的模型有效捕捉表格中实体之间的特殊联系，同时在跨语言场景中具有稳定而良好的效果。

### 1.3.2 关系理解问题

对于关系的理解，我们旨在利用知识库构建人类可理解的结构化表达，以描述二元关系语义，并用于下游任务中，例如关系分类和知识库补全。对关系的建模问题，有以下两个特点：1) 与实体类似，自然语言中关系存在着多义性，即对应多种描述方式；2) 自然语言关系和知识库谓词存在语义间隔，由于知识库的构建避免了信息冗余，一个关系未必能直接映射到知识库中的单个谓词。

为了学习关系的语义，我们根据开放式信息抽取系统在海量文本中抓取的结果为输入，获取单个关系的多个主谓宾三元组，从而以数据驱动的方式进行语义建模，并由粗细两个粒度进行分析。在粗粒度方向，我们关注于关系的多义性，并通过其主宾语的类型搭配来区分其不同的语义。为此，我们对知识库中不同类型间的包含关系进行挖掘，

构建出更加丰富的类型层次结构，并根据主宾语类型对的覆盖率和精细度进行筛选，生成最能够代表关系语义的类型搭配。实验结果表明我们的模型效果优于传统的选择偏好模型。在细粒度的方向，我们关注关系语义的精确表达，使用人类能理解的图结构作为语义的表达形式。我们提出了基于规则推导的模式图推理模型，该模型从已有的关系三元组出发，用知识库中的谓词序列连接主宾语，并在此基础上搜索额外的限制，生成一系列具有“路径 + 分支”形式的复杂模式图。随后，模型再次利用已有实例，学习候选模式图上的概率分布，在描述关系多义性的同时，也完成了具体和宽泛模式图之间的平衡。我们将生成的模式图概率分布运用于知识库补全任务中，实验结果显示，我们的模式图推理模型不仅具有高度可解释性，而且效果优于其它规则推导模型和新兴的知识库向量模型，

### 1.3.3 问句理解问题

对于问句的理解，我们着眼于基于知识库的自动问答任务，即对描述客观事实的问句，从知识库中寻找出其对应的一个或多个答案实体。对于只包含简单语义的问题，自动问答的过程等价于将问题转换为知识库上的一个事实三元组。然而，人类提出的问题并不总以简单形式呈现，而是会在其中加入更多限制。例如问句中存在多个与答案相关的实体、类型，或是包含时间、顺序信息等，对应着多条和目标答案相关的三元组，因此面向简单问句的模型便不再奏效。在复杂语义场景中，知识库问答具有以下挑战：1) 如何从问句中发现存在的多个关系，并组合成为一个候选语义结构；2) 如何计算自然语言问句和复杂语义结构之间的匹配程度。

我们提出了针对复杂问句的深度学习语义匹配模型。该模型的思路是细粒度关系理解模型的延续，使用知识库中的实体、谓词、类型等信息作为基本元素组成图结构，称其为查询图，用于表示问句语义。它在具有良好可解释性的同时，可以通过结构化查询语句（如 SPARQL）在知识库中找出所有满足语义的答案实体。模型首先利用多阶段候选生成方式，由简到繁生成不同的候选查询图，我们已有工作进行优化，更快速和准确地表示类型和时间限制。之后，语义匹配模型的核心是利用深度神经网络学习问句和查询图整体的匹配程度，我们的创新点在于对复杂的查询图进行整体编码，生成其唯一的向量表示，从而捕捉问句中不同语义成分的组合特征。同时，我们利用依存语法分析作为对问句字面信息的补充，使模型能更有效地将问句和不同的语义成分对齐。实验结果表明，基于复杂查询图的深度学习模型在多个复杂问题和简单问题数据集上都具有良好的性能。

## 1.4 论文结构

本文组织结构如下：

第一章主要介绍了知识库的发展，以及语义理解的研究背景，并从实体、关系、问句三个角度出发，对研究现状进行概述，最后介绍了本文在这三个方面的语义理解的研究内容和贡献。

第二章主要对实体、关系、问句的语义理解问题进行综述，围绕具体任务介绍这些问题的基础知识，以及已有的相关工作。

第三章围绕实体理解问题，主要介绍本文在跨语言场景中，面向表格文本进行实体链接的深度学习模型。

第四章围绕关系理解问题，主要介绍本文对自然语言关系的粗粒度和细粒度建模，前者面向关系的多义性，后者面向关系的准确语义表达，并运用于知识库补全任务。

第五章围绕问句理解问题，主要介绍本文在知识库问答任务中，面向复杂问句的深度语义匹配模型。

第六章主要对论文工作进行总结，并展望未来可以继续的研究工作。

## 第二章 国内外相关研究综述

本章中，我们将围绕具体任务，介绍实体、关系、问句语义理解的基础知识和研究综述。实体理解部分，我们从传统的实体链接任务出发，介绍在文本和表格中的特征工程和深度学习模型，以及用于跨语言实体链接任务中的跨语言词向量模型；关系理解部分，我们关注知识库补全任务，并介绍两类主要的模型，分别是规则推导模型和知识库向量模型；问句理解部分，我们将注意力放在面向客观事实类问题的知识库自动问答任务，同样介绍两类模型，即基于语义解析和基于信息抽取的模型，前者与我们的研究更加密切，本节也会重点阐述语义解析模型的多个组成部分。

### 2.1 实体理解：实体链接任务

实体链接任务是一类从自然语言文本中识别出代表实体的字符串，并将其映射到知识库中特定实体的任务。人工进行的实体链接体现在维基百科的页面编辑过程中，页面作者会手动为部分代表实体的短语添加超链接，指向对应实体的维基页面。这种带有维基内部超链接的短语被称为锚文本（Anchor Text），本文中也称为实体短语。基于机器学习的实体链接可以应用于不同场合的文本输入，背后所使用的目标知识库也不局限于维基百科，其它常用的知识库包括 DBPedia，Yago 以及 Freebase。考虑到这些知识库均基于维基百科信息构建而成，实体链接任务又被称为“维基化”（Wikification）<sup>[13]</sup>。

以英文维基百科为例，一个典型的实体链接任务见下例：

Michael Jordan, also known by his initials, MJ, is a former professional basketball player. He played 15 seasons in the National Basketball Association for the Chicago Bulls and Washington Wizards.

实体链接任务首先需要提取出句中存在的实体短语，即下划线对应的部分。该步骤与命名实体识别任务类似，不同之处在于我们关注的实体短语除了命名实体（具体的人名、地名、组织名、书名、电影名等）之外，还包含了维基百科中存在的概念实体，用于指代一组相似实体。下一个步骤对每个实体短语从维基百科中抽取出候选实体集，并定义短语和候选实体之间的匹配分数，从而将短语链接至最相关的候选实体，例如句中的 5 个实体短语分别对应维基百科中的实体 *Michael Jordan*<sup>1</sup>, *basketball*<sup>2</sup>, *National*

<sup>1</sup>[https://en.wikipedia.org/wiki/Michael\\_Jordan](https://en.wikipedia.org/wiki/Michael_Jordan)

<sup>2</sup><https://en.wikipedia.org/wiki/Basketball>

*Basketball Association<sup>1</sup>, Chicago Bulls<sup>2</sup> 以及 Washington Wizards<sup>3</sup>*。

除了无结构的纯文本以外，互联网语料中的表格也蕴含了大量与实体相关的知识。对表格进行实体链接的研究起源于 Limaye 等人<sup>[51]</sup>的工作，如图2-1所示，除了每个单元格所对应的实体之外，得益于半结构化的组织形式，同一表列内的实体通常具有相同的类型，而且两列实体之间描述了同一种关系的不同实例，这些是非结构化文本所不具备的优势。因为表格带来了丰富关系知识，表格上的实体链接在近年来受到了更多的关注。

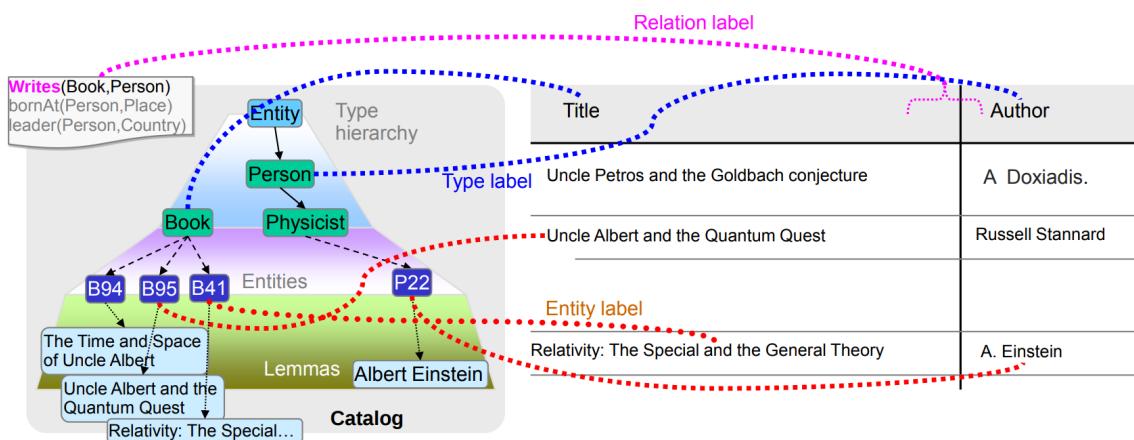


图 2-1 Limaye 等人提出的表格链接任务。<sup>[51]</sup>

Figure 2-1 Entity linking task on web tables proposed by Limaye et al.

作为自然语言理解中的基本任务，实体链接是一系列下游任务的前置步骤。首先，开放式信息抽取抽取的主谓宾三元组均为文本表示，通常具有歧义，一些研究工作旨在对三元组中的实体短语进行链接，代表文献包括 [4, 52]，在实现三元组消歧的同时，结合知识库推理出主语和宾语所代表的类型，有助于挖掘不同谓词关系之间的语义联系。其次，实体链接与知识库补全任务密切相关，该任务的目的是向已有知识库中补充新的事实三元组。这些新添加的三元组主要来源于两方面：随时间发展所产生的全新事实，或基于已有知识的归纳推理。基于前者的补全依赖于信息抽取系统的不断挖掘，因此对主宾语的链接的准确率决定了知识库补全的效果。知识库补全的相关内容将在2.2节中论述。最后，对于自动问答任务，尤其是我们关注的基于知识库的事实类自动问答任务，其描述的是与问句中实体相关的事，因此不管以何种方式对答案进行建模，都依赖于对已有实体的准确定位，以限定问句对应语义的搜索范围。因此，实体链接的结果好坏，

<sup>1</sup>[https://en.wikipedia.org/wiki/National\\_Basketball\\_Association](https://en.wikipedia.org/wiki/National_Basketball_Association)

<sup>2</sup>[https://en.wikipedia.org/wiki/Chicago\\_Bulls](https://en.wikipedia.org/wiki/Chicago_Bulls)

<sup>3</sup>[https://en.wikipedia.org/wiki/Washington\\_Wizards](https://en.wikipedia.org/wiki/Washington_Wizards)

对这些任务的效果均有着很大程度的影响。

实体链接的重点在于从多个候选中找出正确的那个实体，其本质为消除实体级别存在的一词多义性，例如短语“Michael Jordan”具有多个可能的候选，在维基百科中可能代表篮球明星、足球运动员、著名的机器学习教授，甚至更多不那么有名的人。在缺乏上下文信息的情况下，很难进行准确的链接。因此，一个良好的实体链接模型，相关性分数需要考虑多个因素，包括实体本身的先验知识，实体与短语的匹配程度，以及实体与短语所在上下文的契合度。

基于特征工程的方式，其信息来源主要为维基百科上的统计数据。随着深度学习的发展，实体链接的研究将重心放在用表示学习替代或改良传统的特征工程方法。得益于文本的向量表示技术，以及神经网络的特征学习能力，深度学习方法通过计算文本、实体的向量作为其语义表达，并利用高维空间的相似度衡量短语和候选实体之间的相关性，在效果上也取得了不错的提升。接下来的几个小节主要介绍基于特征工程和深度学习的实体链接模型，并单独介绍用于跨语言链接场景中的跨语言词向量技术。

### 2.1.1 基于特征工程的实体链接

基于特征工程的实体链接方法，较为经典的工作包括文献 [15, 16, 53-55]。这类方法的共性在于用预定义好的函数或概率值，描述候选实体与实体短语及其上下文之间的特征，通过学习特征的带权相加计算最终的匹配度。

在这类方法中，一个候选实体所具有的特征可以分为三类：先验特征，上下文语义特征，以及与句中其它实体的关联特征。先验特征与短语所处的文本无关，仅基于短语与候选实体在维基百科中的统计信息，主要体现为短语所在锚文本链接至该实体的概率，以及实体出现在不同维基页面中的概率。上下文语义特征关注文本及目标实体的语义近似程度，利用词袋模型（Bag-of-Words）将短语的上下文以及目标实体的维基页面分别转化为向量形式，通过 TF-IDF 得到不同词语的重要性，并用向量间的相似度代表语义匹配程度。实体间的关联特征体现在不同短语所对应的实体之间，它们维基百科页面的相关性有助于实体链接的判断。在维基百科中，若两个实体同时指向的页面较多，或同时指向这两个实体页面的其它实体较多，则它们具有较高的相关度。常用的计算方式主要为点对点互信息（PMI）<sup>[15]</sup> 或基于谷歌距离的维基链接度量（WLM）<sup>[16]</sup>。其它的可选方式包括 Jaccard 相似度，以及具有非对称形式的条件概率。

对于实体链接模型的训练方式，若不同的实体短语之间互相独立，那么训练过程是一个简单的监督学习问题，即判断每一个 < 短语，实体 > 对的正确性。考虑到每个短语仅对应唯一的正样本，其余所有候选实体均为负样本，为了保持正负样本的平衡，一般采用 Ranking SVM<sup>[56]</sup> 或最大间隔（Max Margin）模型进行训练。这样的做法称为局部优

化，显然忽略了候选实体之间的关联。为了能捕捉这一特征，Ratinov 等人<sup>[15]</sup>首先利用局部优化方案，对每个实体短语进行链接，得到具有一定质量的次优解，然后根据次优解计算候选实体与其它实体的关联特征，用于完整的模型训练。Shen 等人<sup>[16]</sup>对两个实体之间的相关度计算进行了类似的简化，将单个实体与其余短语的所有候选分别进行相关度计算，并选取最大值作为特征。Bhagavatula 等人<sup>[57]</sup>通过迭代的方式不断对每个短语预测的实体进行更改，实体间的相关度特征也随着不断变化，更加靠近真实情况。以上这些方法通过简化特征或迭代预测的方式，使得训练过程得以保持对每一个 < 短语，实体 > 进行打分的形式。

与之相对的全局优化方法则将整个文本以及所有不同短语的候选实体整合在一个目标函数中，因此可以对更加复杂的相关性进行建模。Hoffart 等人<sup>[53]</sup>在模型中构建了一个包含所有实体短语以及候选实体的无向图，不同的特征值体现为图中具有不同权重的边，该模型通过贪心算法寻找图中具有最大权重的稠密子图，使得每一个短语在子图中与唯一的一个实体相连。Luo 等人<sup>[55]</sup>利用条件随机场 (CRF) 对实体链接与命名实体识别任务 (NER) 同时建模，使实体链接过程能够利用 NER 的一系列特征。Yang 等人<sup>[54]</sup>提出的 S-MART 模型考虑到了多个实体短语不能重叠的限制，属于全局优化的范畴，利用前向后向算法对所有短语进行链接，保证在多个短语重叠的情况下，最多一个短语指向具体的实体，其余均指向空实体。同时，该模型通过迭代决策树 (MART，即 GBDT) 对匹配度进行建模，MART 在工业界被广泛使用，具有非常良好的效果。

### 2.1.2 基于深度学习的实体链接

传统的特征工程方法需要人工干预，寻找更有效的特征还需要花费更多的时间。最新的深度学习研究更加关注利用神经网络的表示学习能力，自动挖掘文本和实体的隐藏特征，形成各自的抽象表示，在避免特征工程耗费人力的同时，还能学习人类难以直接描述的高层特征。在自然语言处理领域中，词向量技术<sup>[58-60]</sup>为深度学习模型的基础。以 Skip-Gram 和 CBOW<sup>[60]</sup>为代表的词向量模型，通过半监督方式从纯文本语料中构建训练数据，根据上下文单词预测、词序列正确性预测等任务，学习每个单词的向量表达，对应连续语义空间中的不同坐标点。相似单词在语料库中具有接近的上下文，因此在连续空间中位置更加接近。由于句子和段落都是不同单词的特定组合，因此它们的抽象表示主要通过神经网络对词向量进行计算而得，常见的方法包括卷积神经网络<sup>[61]</sup>，循环神经网络<sup>[62]</sup>以及具有注意力机制<sup>[63]</sup>的模型变种。

基于深度学习的实体链接模型包括文献 [17-19, 64]，它们首先通过特定的神经网络结构，计算出实体短语所在的上下文表示，以及候选实体的表示。之后通过定义向量之间的相似度函数作为匹配分数。而这些模型之间的区别，主要在于表示实体或短语的信

息来源和编码粒度。

Francis-Landau 等人<sup>[17]</sup> 使用了以卷积神经网络为主体的神经网络进行实体链接。如图2-2所示，文档中的一个实体短语对应着三种上下文：短语自身、短语所在句子、短语所在段落。相似地，一个候选实体也对应两种上下文：实体的名称，以及对应维基页面中的所有段落。将它们输入至不同参数的卷积神经网络，便可得到短语和实体在多个不同粒度的上下文向量表示。通过两两计算相似度的方式，即可得到 6 个不同粒度组合的语义相似度。最后结合已有的人工特征，通过逻辑回归层得到最终匹配度。

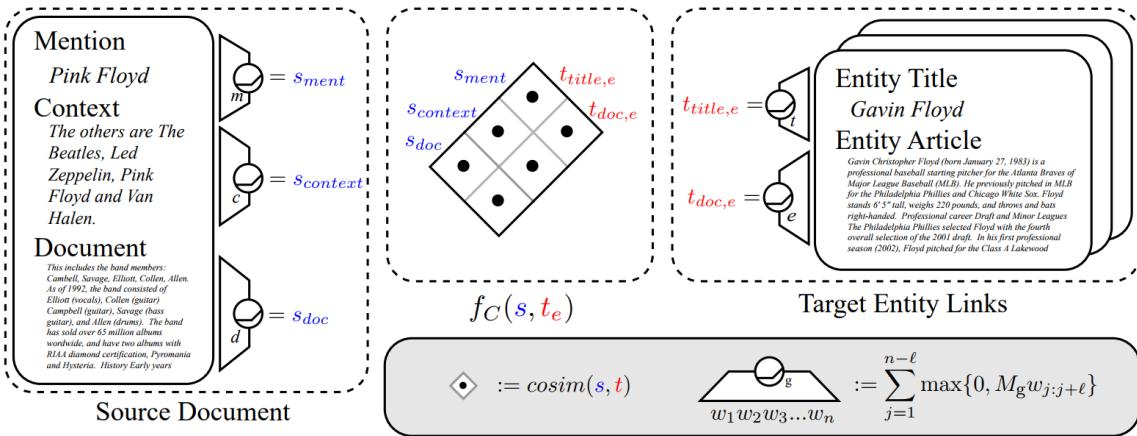


图 2-2 基于多粒度卷积神经网络的实体链接模型。<sup>[17]</sup>

Figure 2-2 Entity linking model with CNN at multiple granularities.

Sun 等人<sup>[18]</sup> 的模型对不同的上下文信息采用了不一样的网络结构。如图2-3所示，对短语建模的信息包括短语自身，以及去除自身后的句子两部分，而实体方面，除了利用本身名称之外，还使用了它在维基百科中的分类信息，用这类人工提炼的知识补充实体的表示。对句子的表示学习依然使用卷积神经网络，其余三种信息由于长度较短，均直接使用了词向量平均的方式得到向量表达。进一步，该模型利用较为复杂的神经张量层将各部分向量结合，分别得到实体和短语的整体表达。类似的方法还有文献 [19]，对实体的维基分类信息进行表示学习，通过双向循环神经网络对短语所在句子进行编码。同时模型定义了不同信息之间的多种损失函数，对训练数据的利用更加充分。

此外，知识库向量学习技术<sup>[29]</sup> 也被用于实体链接任务中。知识库向量学习与词向量学习类似，以大量事实三元组作为训练数据，学习每个实体的向量表示，使得相近语义的实体具有相近的向量。Fang 等人<sup>[64]</sup> 提出的链接模型基于知识库向量与词向量的融合：通过实体与短语互相替代的方式，定义了基于三元组以及共现词对的目标函数，促使实体与其短语的向量尽可能一致，因此所有向量表示被映射到同一个高维语义空间中。融合的优势在于实体和单词之间直接可比，通过距离度量函数计算候选实体与短语

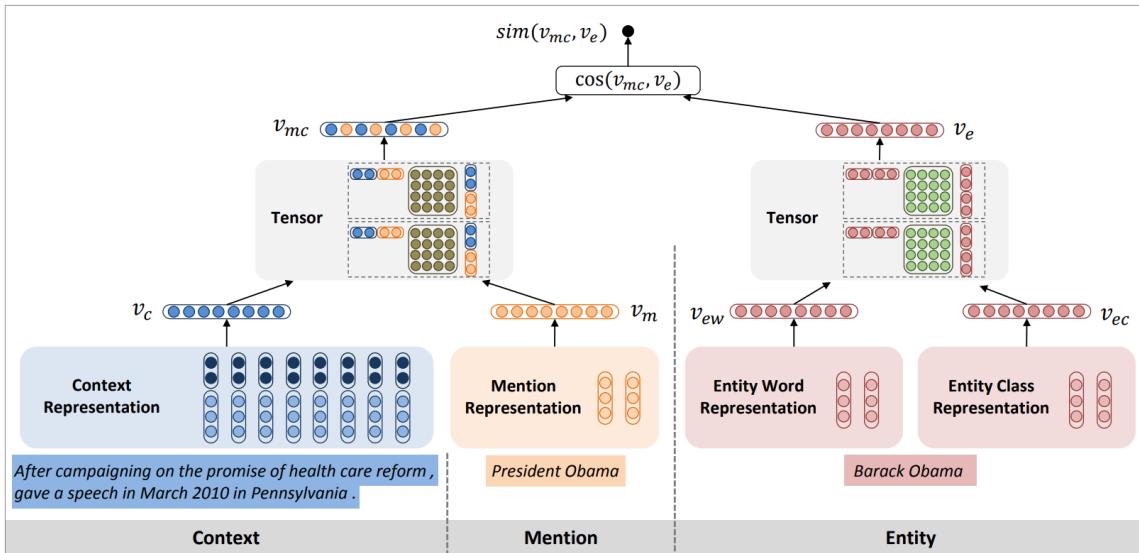
图 2-3 基于神经张量层的链接模型。<sup>[18]</sup>

Figure 2-3 NTN based entity linking model proposed by Sun et al.

上下文中不同词的距离，并以此作为链接模型的特征。

### 2.1.3 跨语言词向量

上一节的论述中提到了词向量模型，用于学习词汇的连续语义表示，但仅局限于单一语言。对于涉及多个语言的任务，难点在于如何实现语义的跨语言过渡。为了解决此问题，跨语言词向量模型（Cross-Lingual Embedding Model）旨在消除单词语义表示对语言的依赖，将不同语言的向量表示映射至同一连续空间，并依旧保持相似语义单词更加接近的特性，以此实现语义迁移。例如图2-4展示了一个英语和德语之间的共享语义空间，可以清晰地识别出两种语言间的许多组翻译词对。

跨语言词向量的训练，需要依赖平行语料库用于给模型提供语义对齐信号，不同的训练方式区别在于平行语料的类型不同，例如单词级别<sup>[60, 65, 66]</sup>、句子级别<sup>[67, 68]</sup>、文档级别<sup>[69]</sup>的对齐。利用单词级别对齐进行跨语言词向量训练的工作最为普遍，训练数据主要来自于双语或多语词典中抽取出的高质量翻译词对。以此为例，对于源语言  $s$  和目标语言  $t$ ，模型的损失函数  $J$  由三部分组成：

$$J = \mathcal{L}_{mono}^s + \mathcal{L}_{mono}^t + \Omega^{s \rightarrow t}, \quad (2-1)$$

其中， $\mathcal{L}_{mono}$  代表各自语言上进行单语言词向量训练的损失，可直接使用 CBOW 等已有模型进行计算，而  $\Omega^{s \rightarrow t}$  为正则项，对应单词对齐的损失。

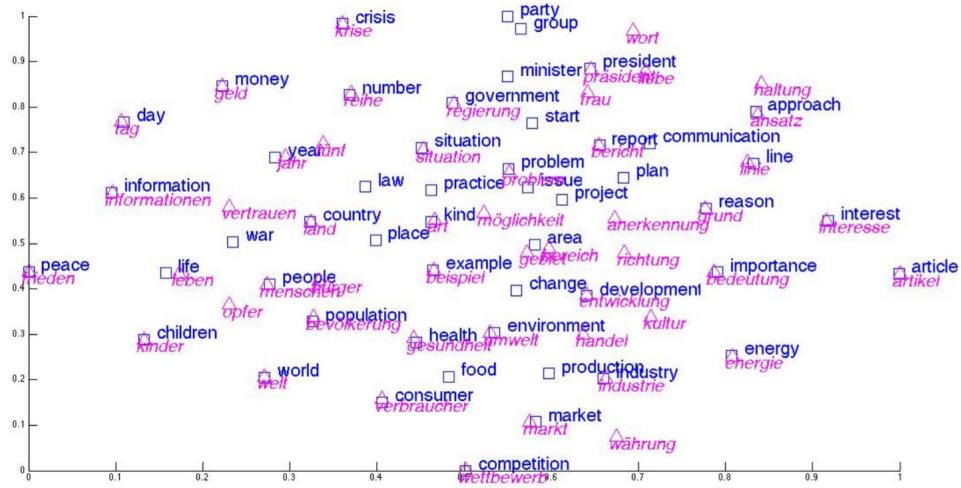
图 2-4 英语和德语间的跨语言词向量例子。<sup>[21]</sup>

Figure 2-4 An example of cross-lingual word embeddings between English and German.

对于  $\Omega^{s \rightarrow t}$  的定义，相关文献进行了不同的尝试。Mikolov 等人<sup>[60]</sup>发现，在不同语言中，多个单词的向量表达之间，几何关系较为相似，例如英语和西班牙语中，表示数字的单词之间的相对位置几乎一致，表示动物的单词也有类似特性。基于以上观察，该工作提出的模型使用了线性变换的方案，训练转移矩阵  $\mathbf{W}$ （或称投影矩阵），使得源语言词向量  $\mathbf{x}^s$  经过  $\mathbf{W}$  投影后，和对齐的目标语言词向量  $\mathbf{x}^t$  的欧氏距离平方（即均方误差）尽可能小：

$$\Omega^{s \rightarrow t} = \sum_i \|\mathbf{W}\mathbf{x}_i^s - \mathbf{x}_i^t\|^2, \quad (2-2)$$

基于线性映射的跨语言词向量模型具有一些变种，例如文献 [70, 71] 限制转移矩阵  $\mathbf{W}$  为单位正交阵，以保证映射后的词向量维持单位长度，Artetxe 等人<sup>[72]</sup>指出，对于模型效果而言，转移矩阵正交化比向量正则化更加重要。Lazaridou 等人<sup>[66]</sup>对  $\Omega^{s \rightarrow t}$  的定义使用了最大间隔（Max Margin）损失来代替均方误差损失，即不追求  $\mathbf{W}\mathbf{x}^s$  与  $\mathbf{x}^t$  绝对距离尽可能小，而是让  $\mathbf{x}^t$  比其它任何不相关单词都更加接近  $\mathbf{W}\mathbf{x}^s$ ，从而避免跨语言词向量出现过多中枢词的现象。Faruqui 等人<sup>[73]</sup>利用典型相关分析（Canonical Correlation Analysis, CCA)<sup>[74]</sup>进行词向量训练。CCA 同为线性投影方式，不同之处在于，CCA 对两个语言分别学习一个线性变换矩阵，目标是尽可能降低映射后每个翻译词对的互协方差分值。

此外，跨语言词向量的训练还可对于通过句子或文档级别的平行语料进行跨语言词向量的训练，由于不是本文的研究重点，故不展开论述。跨语言词向量能够应用在多种任务中，例如文档分类<sup>[65]</sup>、词性标注<sup>[71]</sup>、命名实体识别<sup>[75]</sup>、机器翻译<sup>[76]</sup>等，其带来的

知识迁移具有很高的实用性。对于训练集和测试集为不同语言的任务，跨语言词向量能实现知识在不同语言上的迁移；对于类似机器翻译、跨语言实体链接等输入和输出为不同语言的任务，预训练好的跨语言词向量能够作为特定任务模型的训练起点，消除语义的间隔，从而提升整体效果。

## 2.2 关系理解：知识库补全任务

现有的知识库具包含了庞大数量的实体和事实，但其中的内容仍然不够完全，尤其是存在大量的长尾实体，并没有多少事实与之相关。因此，知识库补全任务（Knowledge Base Completion, KBC）的目的，是向已有知识库中添加缺失的事实三元组  $(e_1, p, e_2)$ 。这些新增的三元组中，无论是谓词  $p$  还是它的两个参数实体  $e_1$  和  $e_2$ ，都已存在于知识库中。换言之，新增的事实并不会给知识库带来额外的节点，或是从没见过的边，而是让知识库的图结构更加稠密。

新增事实三元组（尤其是参数实体）的获取方式通常有两种。第一种来源为经过了实体链接过后的外部文本，纯文本或表格文本均可称为来源。对于纯文本，当在句中定位两个参数实体后，句子剩下的部分成为了描述实体间关系的上下文。此时即可依据上下文信息来预测对应的知识库谓词，即等价于关系分类问题，相关研究包括词级别卷积神经网络<sup>[61]</sup> 以及依存语法路径上循环神经网络<sup>[62]</sup>。对于表格文本，知识库补全关注于表格的两列之间，利用表列所包含的实体，判断两列之间的关系是否与知识库中特定谓词对应，实现可能的大批量事实补全。例如图2-1，Title 和 Author 列之间的关系被映射到 YAGO 中的 *Writes* 谓词。

第二种来源不涉及到任何知识库外的信息，新增三元组来自于知识库的内部挖掘，通过寻找不同谓词之间的关联，推理出可能缺失的事实。例如1.1节提到的例子，“某人的国籍缺失，可以根据其出生地所在的国家进行推测”，人类可以通过这样的方式进行知识的手动补充，正是因为掌握了国籍、出生地、地点被包含这三个谓词之间的关联。这条路线不受知识库外信息的干扰，因此是我们关注的研究点。

学术界已有工作在此场景上研究知识库补全的解决方案，并提出了相关的 KBC 数据集，例如 FB15k<sup>[29]</sup> 和 WN18<sup>[77]</sup>，分别是 Freebase 和 WordNet 的子集。对于知识库补全模型的测评，主要通过主宾语预测（Link Prediction）以及三元组分类（Triple Classification）这两个子任务来进行。前者为三元组  $(e_1, p, ?)$  或  $(?, p, e_2)$  预测缺失的主语或宾语，后者则判断给定的  $(e_1, p, e_2)$  是否为正确事实。两者虽然形式不同，但都需要计算三元组的置信分： $S(e_1, p, e_2; KB)$ 。

解决知识库补全的方法主要分为两类。第一类基于规则推导，用逻辑表达式描述实体间存在特定谓词时，所需要满足的特定规则；第二类基于知识库向量，学习所有实体、

谓词的连续特征表示，并挖掘三元组各部分特征表示之间存在的深层代数关系。下面将分别介绍这两种方法。

### 2.2.1 基于规则推导的模型

基于规则推导的模型旨在使用人类可以直接理解的规则形式，来描述不同谓词之间的联系，即如果  $e_1$  和  $e_2$  之间满足特定的条件，则推理出三元组  $(e_1, p, e_2)$  成立。为方便论述，我们将三元组以布尔表达式  $p(e_1, e_2)$  表示，若对应三元组存在于知识库中，则表达式为真，否则表达式为假。文献 [25] 以一个简单的推导规则为例：若已知某运动员为球队效力，以及球队所处联盟，则可以推导出运动员所参与的体育联盟。该规则可以通过一阶逻辑表达式进行形式化描述：

$$\begin{aligned} \exists b \in E, & \text{athletesPlaysForTeam}(a, b) \wedge \text{teamPlaysInLeague}(b, c) \\ & \implies \text{athletesPlaysInLeague}(a, c), \end{aligned} \quad (2-3)$$

其中  $E$  表示知识库实体集合。规则的左侧部分，使用的两个谓词实现了主语  $a$  到宾语  $c$  的连接，且不包含多余的布尔表达式，因此这条规则表现为由谓词序列构成的路径。对于“依靠出生地推测国籍”的规则，我们可以形式化为以下逻辑表达式：

$$\begin{aligned} \exists b \in E, & \text{placeOfBirth}(a, b) \wedge \text{containedBy}(b, c) \\ & \wedge \text{isA}(c, \text{country}) \implies \text{nationality}(a, c), \end{aligned} \quad (2-4)$$

其中谓词序列  $\{\text{placeOfBirth}, \text{containedBy}\}$  为连接主宾语的路径，而谓词  $\text{isA}$  使得宾语  $c$  还需要满足额外的限制条件，因此这条规则具有比路径更加复杂的结构。

利用单个推导规则进行知识库补全存在两个局限：首先，规则本身不一定完全正确，限制条件过于宽泛的规则可能打来错误的事实；其次，单个规则的覆盖率较低，能够补全的知识有限。针对这两个局限，已有的规则推导工作都致力于从知识库中挖掘目标谓词的多条规则，并学习不同规则的重要性，以实现更健壮的知识库补全。

Lao 等人提出的 PRA 模型<sup>[78]</sup> 实现了规则的挖掘和学习。对于目标谓词  $p$ ，模型首先利用谓词已有的三元组  $(e_1, p, e_2)$  作为训练数据，在知识库中寻找所有能连通  $e_1$  和  $e_2$  的谓词序列，形成了多种路径形式规则。模型利用逻辑回归实现规则权重的训练，以及对新三元组  $(e'_1, p, e'_2)$  是否为真的预测。每一条挖掘的规则类比为一个特征，对应的特征值为路径随机游走概率，即从主语  $e'_1$  出发，沿规则的谓词序列随意跳转，最后到达  $e'_2$  的概率。由于知识库中的谓词可能代表一对多关系，因此随机游走概率值会小于 1，甚至为 0（无法通过当前规则连通）。基于此，PRA 模型能够快速抽取大量路径规则，并利用已知三元组在规则上随机游走概率实现权重训练。

一些研究工作对 PRA 模型进行了扩展。文献 [25] 对挖掘的路径规则进行了限制，要求规则至少要适用于训练数据中一定比例的主语，并对计算随机游走概率的采样方式进行了优化，这两个扩展都是基于模型性能优化为考量。Gardner 等人提出了 SFE 模型<sup>[26]</sup>，它在 PRA 模型基础上进行了两项改进：首先，模型将作为特征值的随机游走概率替换为 0/1 特征，即只关心是否连通，在大幅度提升运行速度的同时，对结果没有显著影响；其次，模型引入不同于路径规则的其它特征，例如提取路径中的谓词 bigram，或由主语出发却不指向宾语的单边特征等，扩充特征集合以提升知识库补全效果。Wang 等人提出了 CPRA 模型<sup>[79]</sup>，主要针对具有相似谓词的规则推导优化。该模型通过层次聚类识别出具有相似语义的知识库谓词集合，然后对每个集合内的谓词采用多任务学习框架，共享挖掘的规则特征和部分耦合参数，使模型能够捕捉相似谓词之间的共性，实现隐式的训练数据共享。

## 2.2.2 基于知识库向量的模型

与词向量模型类似，知识库向量的模型旨在学习一个知识库中的实体、谓词等元素在连续空间的特征表达，以完成一系列下游任务。具体在知识库补全任务中，三元组置信分的计算并不依赖自动挖掘的推导规则，而是来自主谓宾三者的连续特征表达在不同维度上的交互。词向量模型依靠大规模的纯文本语料，构建上下文单词预测任务来完成训练<sup>[60]</sup>，相比之下，知识库向量模型的训练过程更为直观：利用知识库已有三元组作为训练数据正样本，并自动生成不存在的三元组作为负样本，计算三元组置信分进行训练，而这也恰好与知识库补全任务的目标一致。

知识库向量模型的优点在于所有谓词共同训练，更有效地利用训练数据，并且能在连续空间中体现不同谓词的相似性，但相应地，其缺点在于可解释性较弱。不同的知识库向量模型对实体和谓词的特征表示具有不同的形式，特征交互的方式也不尽相同，下面主要介绍几个具有代表性的模型。

由 Nickel 等人提出的 RESCAL 模型<sup>[28]</sup> 是一个基础的知识库向量模型，模型对三元组  $(e_i, r_k, e_j)$  置信分（简写为  $S$ ）的定义，基于主宾语实体特征表示的在不同维度间的两两交互：

$$S^{RESCAL} = \mathbf{e}_i^\top \mathbf{W}_k \mathbf{e}_j = \sum_{a=1}^d \sum_{b=1}^d w_{kab} e_{ia} e_{jb}, \quad (2-5)$$

其中  $\mathbf{e}_i$  和  $\mathbf{e}_j$  为对应实体向量，维度为  $d$ 。 $\mathbf{W}_k \in \mathbb{R}^{d \times d}$  为谓词  $r_k$  的权重矩阵，其中  $w_{kab}$  体现了实体向量的第  $i$  和  $j$  维特征对于第  $k$  个谓词的交互重要性。由于 RESCAL 使用矩阵连乘形式捕捉实体向量之间的交互，因此也被称为双线性模型。如图2-5a所示，RESCAL 模型可表示为双层神经网络结构，首先通过张量积（Tensor Product）构建实体

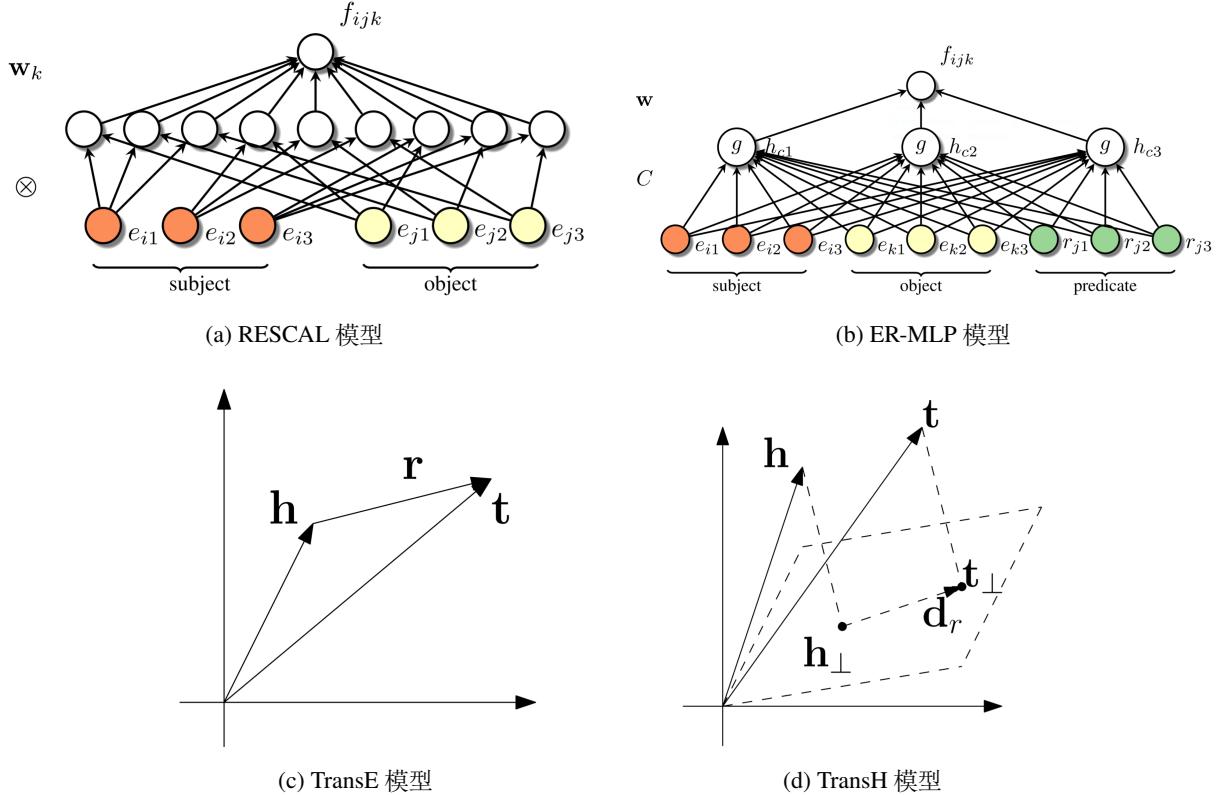
图 2-5 多种知识库向量模型示意图。<sup>[30, 80]</sup>

Figure 2-5 Examples of knowledge base embedding models.

对  $(e_i, e_j)$  的组合特征，再利用与特定谓词相关的参数  $\mathbf{W}_k$  作为权重，得到三元组最终的置信分。RESCAL 模型学习的实体特征表示能够捕捉不同实体间的语义相似性，换言之，若两个实体可以通过相似的谓词连接至相似的其它实体，那么它们的特征表达也更加相近。

RESCAL 模型存在的一个问题在于参数量过大，每一个谓词对应参数量为  $d \times d$ ，对于拥有大量谓词的知识库而言，会带来可扩展性的问题。一些后续研究对此进行了改进。Socher 等人以及 Dong 等人分别提出了 E-MLP<sup>[81]</sup> 和 ER-MLP 模型<sup>[27]</sup>。图2-5b为 ER-MLP 的示意图，均由两层前向网络构成，第一层用于学习三元组的组合特征表示，第二层则通过组合特征输出置信分。E-MLP 结构较为类似，故此处不专门画图。两个模型的置信分计算如下：

$$\begin{aligned} S^{E-MLP} &= \mathbf{w}_k^\top g(C_k[e_i; e_j]), \\ S^{ER-MLP} &= \mathbf{w}^\top g(C[e_i; e_j; r_k]), \end{aligned} \quad (2-6)$$

其中， $g$  为非线性激活函数。对比 RESCAL 模型，E-MLP 的最大不同在于可以通过调整矩阵  $C_k$  来学习实体间不同维度特征的交互，从而优化实体对  $(e_i, e_j)$  的组合特征表示，

并大幅度减少参数数量。E-MLP 模型中，不同的谓词依然对应不同的参数，而 ER-MLP 模型将谓词也映射为向量表示，与两实体共同作为第一层的输入，因此模型的参数  $C$  与  $w$  均与特定谓词无关。两个模型依然能够让语义相似的实体映射至连续空间的相近位置。

Nickel 等人提出了 HOLE 模型<sup>[82]</sup>，该模型利用循环相关运算（Circular Correlation）巧妙地代替了 RESCAL 中的张量积操作：

$$S^{HOLE} = \mathbf{w}_k^\top (\mathbf{e}_i \star \mathbf{e}_j) = \sum_{a=1}^d \sum_{b=0}^{d-1} w_{ka} e_{ia} e_{j,(a+b)\%d}, \quad (2-7)$$

可以从公式中看出，循环相关运算等同于将张量积的结果进行了分组，模型通过对实体特征表示的学习，让具有相似语义的特征交互归为同一组，共享同一个权重。因此 HOLE 的优势在于将 RESCAL 中的二维矩阵参数降低至一维，同时尽可能保留了特征交互的表示能力，并且在实验中效果优于 RESCAL 和 ER-MLP 模型。

此外，Socher 等人还提出了较复杂的神经张量网络（Neural Tensor Networks，NTN）模型<sup>[81]</sup>，可以看做是 RESCAL 和 E-MLP 的组合体，对实体对  $(\mathbf{e}_i, \mathbf{e}_j)$  构建的组合特征表达同时包括双线性和前向网络特征，但模型参数量也因此更加庞大，在小数据集上更容易出现过拟合。

另一类知识库向量模型是以 TransE 为典型的被称作隐距离模型。相比 RESCAL 等模型依照神经网络构建的置信分函数，隐距离模型对置信分的计算则与距离度量直接相关。这与 2.1.3 节介绍的跨语言词向量训练有着相似之处，源语言词向量经过转换后，与翻译后的词向量尽可能相近。而对于知识库向量模型，由于三元组中还有谓词的存在，因此模型训练的实质，是学习主宾语实体向量在特定谓词下的变换方式，对变换之后的向量表示进行距离度量，距离越近，则置信分越高。为了和相关工作统一，此处用  $(h, r, t)$  表示一个事实三元组。

Bordes 等人提出的 SE 模型<sup>[83]</sup> 较为基本，度量实体  $h$  和  $t$  经矩阵变换后的距离：

$$S^{SE} = -dist(\mathbf{A}_r^s \mathbf{h}, \mathbf{A}_r^o \mathbf{t}), \quad (2-8)$$

其中  $dist(\cdot)$  为距离度量函数，例如 L1 距离或欧氏距离。谓词  $r$  对应两个参数矩阵，分别映射主语和宾语实体的向量表示至同一空间。为了降低参数个数，Bordes 等人提出了 TransE 模型<sup>[29]</sup>，这也是后续很多改进模型的起点。受到词向量之间代数运算的启发，例如  $queen \simeq king - man + woman$ <sup>[84]</sup>，不同词之间的关系体现在了它们词向量的位置偏移中，因此如图 2-5c 所示，TransE 模型共享了实体与谓词的特征表示，利用在主语实体在同一空间中的平移变换，代替更加复杂的矩阵变换：

$$S^{TransE} = -dist(\mathbf{h} + \mathbf{r}, \mathbf{t}). \quad (2-9)$$

TransE 模型设计简单、容易实现，并且具有训练速度快、可扩展性高等优点，但是对于知识库中存在的一对多或多对一的谓词不友好，例如固定主谓，TransE 无法有效区分出多个匹配的宾语实体。为此，TransH 模型<sup>[30]</sup>尝试通过超平面投影来解决此问题，公式定义如下：

$$\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r, \quad \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r, \quad S^{TransH} = -dist(\mathbf{h}_\perp + \mathbf{d}_r, \mathbf{t}_\perp), \quad (2-10)$$

如图2-5d所示，TransH 首先把  $\mathbf{h}$  和  $\mathbf{t}$  的向量表示均投影到连续空间中，谓词  $r$  对应的超平面上（ $\mathbf{w}_r$  为单位法向量），并学习谓词向量表示  $\mathbf{d}_r$ ，在超平面上沿用 TransE 的度量。因此，TransH 具有更高的灵活度来应对一对多或多对一谓词，同时依然保有 TransE 可扩展性高的特点。TransR 模型<sup>[31]</sup>同样尝试解决一对多谓词的问题，类似于 SE 和 TransE 的组合体，利用投影矩阵  $\mathbf{M}_r$  将实体表示转移至新的空间后，再进行基于平移的距离度量。因此 TransR 模型中，实体和谓词的特征表达并不共享同一个语义空间，这与 TransE 和 TransH 模型均不同：

$$S^{TransR} = -dist(\mathbf{M}_r \mathbf{h} + \mathbf{r}, \mathbf{M}_r \mathbf{t}). \quad (2-11)$$

此外，还有其它 TransE 模型的改进工作，包括体现距离度量在不同特征维度间差异的 TransA<sup>[85]</sup>，生成谓词多个表示以解决一对多问题的 TransG<sup>[32]</sup> 等，这里不再展开讨论。

### 2.3 问句理解：知识库自动问答任务

自动问答任务是一类以自然语言问句为输入，并自动给出对应答案的任务。基于知识库的自动问答（Knowledge Base Question Answering, KBQA）是其中的一个热门研究方向，也是本文重点关注的问题。在此问答任务中，输入问句为来自开放领域的事实类问句（Factoid Question），即问句本身描述的是与某些特定实体相关的客观事实，对应的答案通常表示为知识库中的实体、时间、数值等简单形式，因此类似“how”“why”等以完整句子作为答案，或事实涉及到主观判断的问题，不在任务的考虑范围之内。以一个简单的英文问句为例，问句“What state borders Texas?” 描述了与德克萨斯州相关的事，其答案有多个，包括 New Mexico, Oklahoma, Arkansas, 以及 Louisiana 四个实体。

对于知识库问答任务，使用的外部信息显然为结构化知识库。正确答案的获取依赖于问答模型对问句整体语义的理解：一方面准确定位问句中出现的相关实体，并链接至知识库；另一方面根据问句信息，推理出未知答案与相关实体在知识库中具有的关系。前者涉及到实体链接技术，后者体现了问句与知识库的语义匹配，也是问答模型的核心。为了衡量问答模型对不同类型问题的效果，学术界已提出了大量知识库问答数据

集，例如对问句进行结构化语义标注的 QALD<sup>[86]</sup> 和 Free917<sup>[38]</sup>，以及具有更大规模问答数据量的 WebQuestions<sup>[40]</sup> 和 SimpleQuestions<sup>[47]</sup> 等。

与知识库补全任务类似，根据问句语义的表示形式进行划分，知识库问答模型大致可以分为两类，即基于语义解析（Semantic Parsing）和基于信息抽取（Information Retrieval）的模型，下面将分别介绍研究。

### 2.3.1 基于语义解析的问答模型

解释语义解析技术之前，我们先讨论人类对问题的思考方式。对于人类来说，问句“what state borders Texas”包含了两个与正确答案相关的线索：1) 答案是一个（美国的）州；2) 答案与德克萨斯州相邻接。由于答案未知，因此每一个线索都对应着一个具有变量参数的事实三元组。语义解析技术的目的，就是用存在于知识库上的实体和谓词，对这些线索进行结构化表示。根据这两条线索，原问题的答案集合可表示为一阶逻辑表达式：

$$\text{AnswerSet}(q) = \{x \mid \text{IsA}(x, \text{US\_State}) \wedge \text{adjoin}(x, \text{Texas})\} \quad (2-12)$$

其中  $x$  代表未知答案实体，表达式中的  $p(x, y)$  为真，当且仅当三元组  $(x, p, y)$  存在于知识库中。对于机器而言，得到逻辑表达式之后，将其翻译为知识库上的查询语句，即可直接得到所有满足语义的答案，这些答案彼此具有完全一致的特征。

由此可见，基于语义解析的自动问答模型，实质是寻找正确的语义结构化表示，即判断<问题，结构化语义>的匹配程度，而不仅仅寻找一个答案实体。相关工作 [39, 40, 42, 43] 的研究重点在于，如何由句子生成知识库上的结构化语义表示，以及如何对问题和语义结构的匹配程度进行建模。

#### 2.3.1.1 结构化语义生成方法

仍以“what state borders Texas”为例，不同研究工作中的结构化语义形式并不相同，但本质都为公式2-12所描述的逻辑表达式。图2-6列出了一些典型工作生成的解析结构。

早期的语义解析模型<sup>[38, 87, 88]</sup> 使用概率化组合文法（Probabilistic Combinatory Categorial Grammar, PCCG）生成语义解析树。该方法与语法解析中的概率化上下文无关文法（Probabilistic Context-Free Grammar, PCFG）相似，根据训练数据学习文法中不同生成式规则的概率值，并自底向上推理出每个句子最可能生成的成分解析树（Constituency Parsing Tree）。图2-6a为例句的成分解析树，描述了整句的语法结构，并标出了不同短语在句中的成分。通过 PCCG 生成的语义解析树如图2-6b所示，PCCG 的生成式规则中不仅具有代表语法的成分信息，同时还包含代表语义的  $\lambda$  表达式，因此不同成分按照语法规则组合的过程中，各自  $\lambda$  表达式也在进行拼接，从而得到对应整句话语义的逻辑

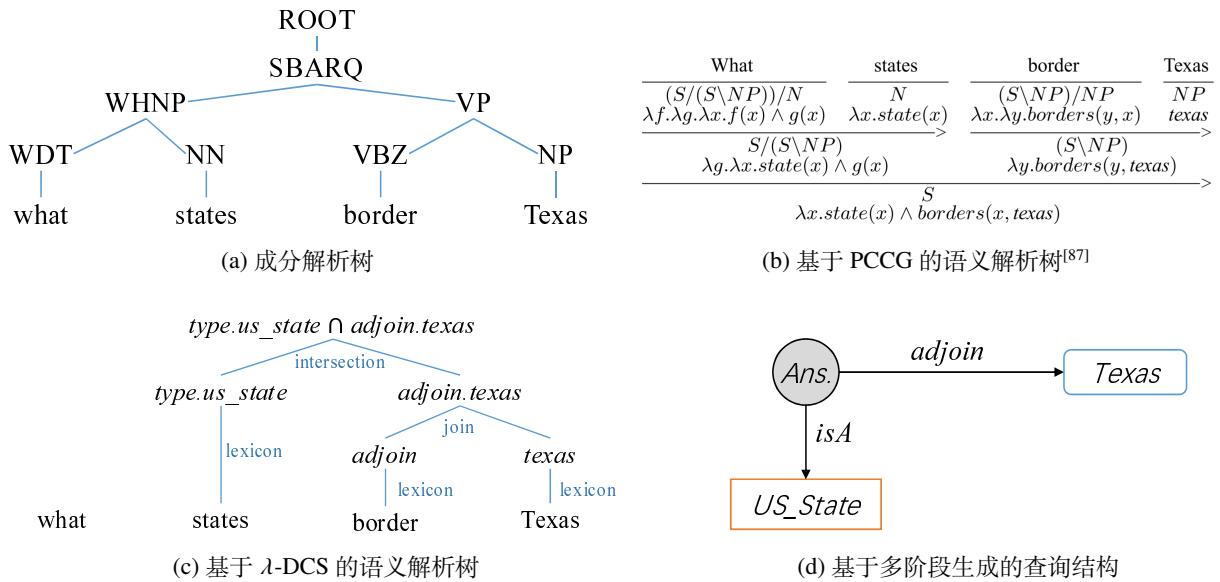


图 2-6 例句 “what state borders texas”的多种解析结构。

Figure 2-6 Several parsing structures for the question “what state borders texas”.

表示。PCCG 语法具有很强的语义表示能力，但由于生成式规则中涉及到不同的  $\lambda$  表达式，同时训练数据匮乏，使得模型的训练具有难度。

Liang 在 2013 年提出的  $\lambda$ -DCS<sup>[89]</sup> 旨在以更加简单的概念和流程，将问句转换为 Freebase 上的语义解析树。如图2-6c所示，生成过程依然是自底向上模式，叶节点（单词或词组）对应 Freebase 中的实体、类型或谓词，但不再具有显式且复杂的  $\lambda$  表达式。 $\lambda$ -DCS 定义了节点组合过程的有限种语义合并方式，包括连接、交集、并集甚至更加高阶的最值、计数等操作，使得与 PCCG 相比，生成的语义解析树在结构更加简单的同时，牺牲了一定表达能力，但对于事实类问题的理解来说依然足够。

Yih 等人<sup>[42]</sup> 提出了一种多阶段的语义结构生成方法，如图2-6d所示，语义解析树被表示为有向图形式，称为查询图，图中的每一条边以及连接的两个节点，都对应公式2-12中的三元组。与之前两种方法的自底向上生成不同，多阶段语义结构生成基于由简到繁，逐步生成查询图的思路。最简单的查询图为答案节点通过谓词（或多个谓词构成的序列）连接至问句中的某一实体，形成仅有一条有向路径构成的查询图。问句中抽取的其它实体、类型、时间等信息，则通过多个不同的阶段，逐步连接至已有的路径上，构成更加复杂的查询图。该方法不受限与问句中词的先后顺序，查询图的生成更加灵活，在多个问答数据集上均有良好的效果。

Cui 等人<sup>[44]</sup> 提出了一种基于模板的方式，对问句生成谓词序列形式的语义结构。模板是对问句抽象表示，它将问句中的实体替换成类型，指代了一组具有相同语法和语义

描述的问句，“what states border \$location”是一个具体的模板例子。模板的提取依靠外部的大规模问答数据，作者对 Yahoo! Answers 中大约 41M 问答对进行实体与答案识别后，生成了约 27M 不同的模板，并通过 EM 算法学习其指向谓词序列的条件概率。对于每一个问句的语义结构生成，则通过生成模型，由模板进行过渡得到不同谓词序列的概率。这样的方法，优点在于利用大量外部数据获取准确率高的模板以及和语义的匹配，但模型的召回率可能成为短板，当问句语法不规范时，简单的模板匹配容易失效。此外，一些文献 [90, 91] 使用了基于依存语法树转换的方式，利用结构相似性实现语义解析结构的生成，这里不再一一介绍。

### 2.3.1.2 语义匹配模型构建

由于自然语言的多义性，语义解析结构的生成结果通常都不唯一，因此需要对 < 问句，语义解析结构 > 的匹配度进行建模，选择最高匹配度的解析结构进行知识库上的答案查询。传统的语义解析模型主要基于特征工程，Berant 等人<sup>[40]</sup> 的研究工作为一个典型例子。语义解析树由  $\lambda$ -DCS 生成，抽取出的特征包含三类：问句中的短语与对应知识库谓词的对齐特征，不同谓词参与合并的特征，以及解析树的总体结构特征。前两类特征来自于解析树的自底向上生成过程，用于捕捉每一个操作，后一类特征则统计解析树中不同类型操作的数量，以及最终返回的答案数量。

为了弥补特征工程耗费人力的缺陷，同时获取更高层面的语义匹配信号，Berant 等人在后续的工作 [41] 中引入了转述特征（Paraphrasing Feature），通过简单的规则将解析树翻译成自然语言问句，并衡量原问句和生成问句之间是否具有转述关系，将其作为额外一组特征。转述关系涉及到自然语言文本匹配问题，作者使用基于词对应的关联模型和词向量的维度空间模型两种方式进行建模，使得问答系统可以得到解析树的整体语义，是传统特征工程的有力补充。

最新的自动问答模型广泛使用了深度学习技术。相关研究的共同点在于遵循一种“编码—比较”框架，其重点在于，通过神经网络的特征学习能力，对问句和解析结构分别进行编码，得到各自向量表示，最后计算向量之间的相似度，代表问句与解析结构的匹配程度。以简单问题数据集 SimpQuestions 为代表的问答模型几乎完全属于这一范畴，由于在 SimpleQuestions 中，问句的语义解析结构均为单一谓词序列，因此这些模型本质上都是对文本序列和谓词序列之间的匹配进行建模。Yu 等人<sup>[49]</sup> 提出的 HR-BiLSTM 模型利用循环神经网络进行建模，如图2-7所示，谓词序列输入分为两个粒度：以唯一编号表示的编号序列，以及将谓词名称相连的单词序列，分别通过双向 LSTM 层进行编码，问句文本的编码也利用了双向 LSTM 层，并使用多层次间的残差连接方式进行编码，旨在让模型能同时捕捉单词粒度和问句整体粒度的语义信号。SimpleQuestions 上的

其它类似模型还包括文献 [48, 50, 92, 93]。

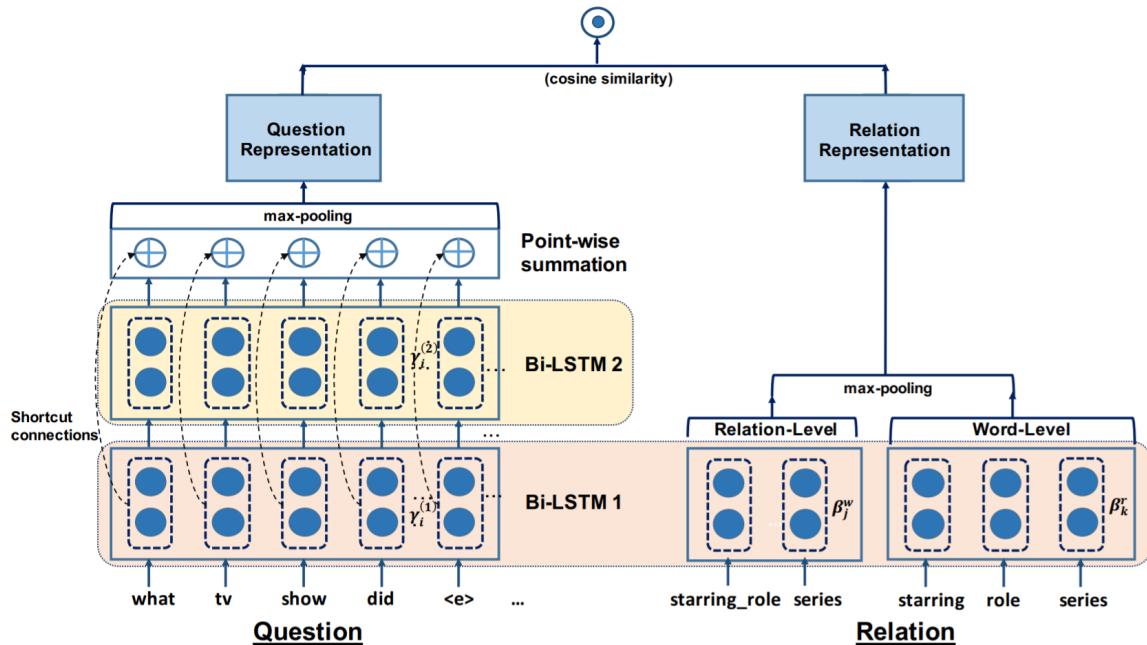


图 2-7 HR-BiLSTM 模型。<sup>[49]</sup>

Figure 2-7 The HR-BiLSTM model.

对于 WebQuestions 等数据集上的复杂问题，如同图2-6d的查询图，虽包含多条路径，但也可以选择其中最重要的路径作为主体与问句计算匹配程度。微软的两个自动问答的研究工作 [42, 43] 利用了基于卷积神经网络的 CDSSM 匹配模型<sup>[94]</sup>，对问句和谓词路径的特征学习更多关注局部的词序信息，见图2-8。其中，前一个研究工作由 Yih 等人<sup>[42]</sup> 提出，深度学习模型仅关注问句和最重要谓词路径的匹配度，对于查询图的其它分支路径，依然使用特征工程的方式寻找问句和谓词的字面匹配。Bao 等人<sup>[43]</sup> 的改进在于同样利用 CDSSM 模型，对分支路径与问句中的特定上下文进行匹配，替代了繁琐的特征工程。然而这些模型并没有能够学习到查询图整体在连续语义空间的特征表达，不同路径的语义互相独立，因此面对复杂问题仍存在缺陷，这也是我们的研究重点。

### 2.3.1.3 训练方式

训练方式的不同，主要取决于训练集中是否包含已标注的语义解析结构。已有的问答数据集中，Free-917 人工标注了每个问题的逻辑表达式，QALD 系列数据集则标注了 SPARQL 查询语句。对于这些正确结构已给定的数据集，可以直接利用监督学习算法进行匹配度训练。

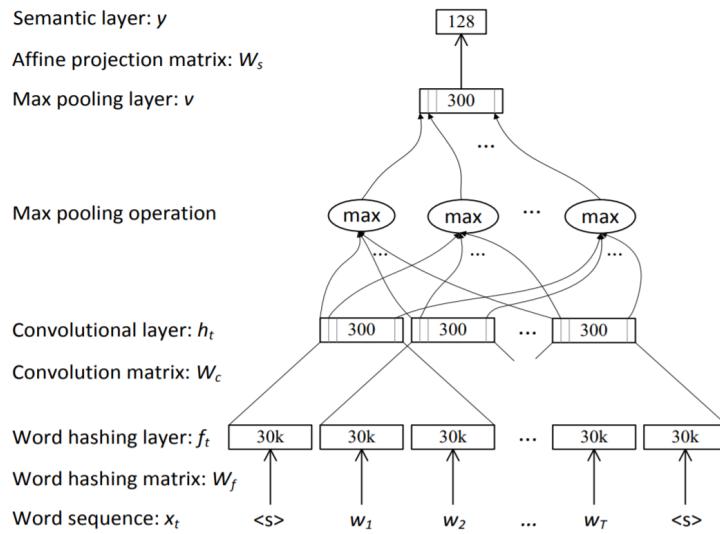
图 2-8 CDSSM 模型。<sup>[94]</sup>

Figure 2-8 The CDSSM model.

显然语义结构的标注需要知识库领域的专家，因此标注过程会消耗大量人力，更大规模的数据集例如 WebQuestions 和 ComplexQuestions 仅包含每个问题的正确答案，而没有语义结构信息。对于这些数据集，首先需要通过远距离监督方式构造可直接使用的训练数据，即对所有训练问题，自动生成语义结构的正负样本。已有的方法主要利用  $F_1$  分数衡量语义结构的好坏，兼顾其生成的查询结果的准确率与召回率，即  $F_1 = 2 \cdot P \cdot R / (P+R)$ ，其中  $P$  代表准确率， $R$  代表召回率。再通过设定阈值将不同的语义结构划分为正负样本，例如 Berant 等人<sup>[40]</sup> 仅将  $F_1$  分值为 1（即答案完全匹配）的语义结构作为正样本，而 Yih 等人<sup>[42]</sup> 则将阈值设为 0.5，容忍一定程度的答案不完全匹配。远距离监督方式避免了人工标注大量语义结构，但考虑到语义偏差的存在，即答案正确的语义结构未必正确，自动生成的训练数据也会引入一定量的错误。

### 2.3.2 基于信息抽取的问答模型

基于信息抽取的自动问答模型旨在直接从知识库中寻找正确答案，而不尝试对问题进行具体化的语义建模。模型主要包含三个步骤：1) 对问句进行实体链接，得到其中包含的相关实体；2) 在知识库中抽取出这些相关实体周围的其它实体，构成候选答案集合；3) 计算问句与每一个候选答案的匹配度，以此预测出其中的正确答案实体。显然模型的关键点在于第三步，即以怎样的特征描述候选答案与问句之间的关联。在知识库中，一个实体所具有的信息主要包含它的名称、类型、直接相连的谓词以及周围的其它实体。这些信息组成了知识库中以该实体为中心的局部图，不同的信息抽取模型都以

这样的局部图作为候选答案实体的输入。而这些模型的区别，在于特征的选取或学习方式。

Yao 等人<sup>[35]</sup>提出的模型利用特征工程方式，将问句特征与候选答案特征进行配对组合，得到大规模的关联特征。问句侧的特征来源于依存语法树，从中抽取出不同的依存路径，以及具有强烈语义的词汇（如动词，wh-疑问词）。答案侧的特征为答案的类型，以及与问句已知实体相连的谓词路径。通过训练，具有高相关性的配对特征（例如疑问词“where”与答案类型 *location* 配对）将具有更高的权重。

深度学习同样适用于基于信息抽取的问答模型。Bordes 等人<sup>[95]</sup>提出了 QASE 模型，同样基于“编码—比较”框架，如图2–9所示，问句和候选答案的局部图分别进行编码，问句的编码信息为每个词的出现次数，候选答案则通过二进制编码表示答案实体自身、所属类型、相邻的谓词等信息。模型学习映射矩阵  $W$ ，将各自编码转换为连续空间上的语义向量， $W$  的每一行对应一个元素（词、实体、类型、谓词）的向量表示，因此该模型实现了词向量和知识库向量的联合建模。

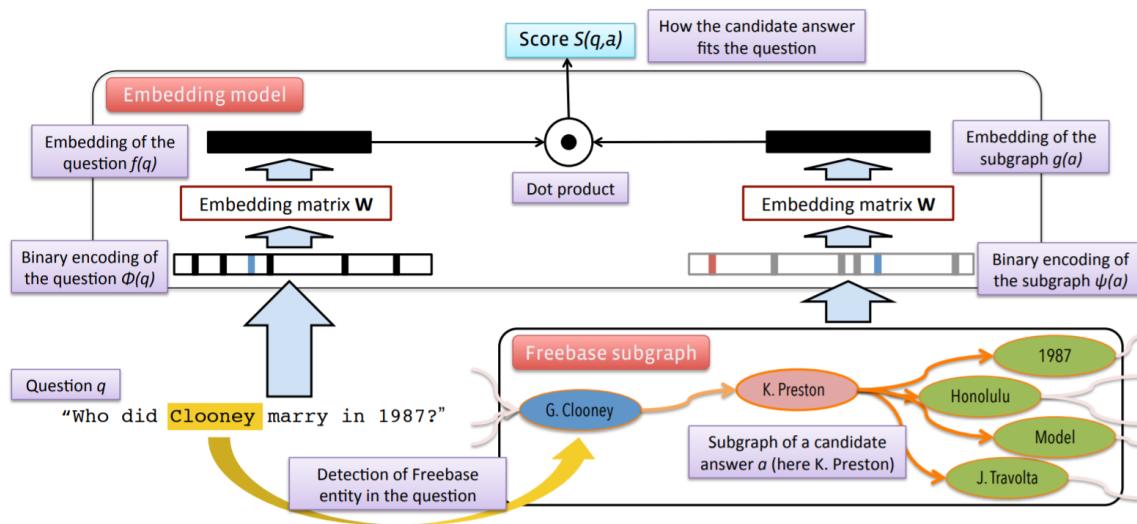


图 2–9 QASE 模型。<sup>[95]</sup>

Figure 2–9 The QASE model.

还有一些深度学习模型采用问句分别与答案相关的不同维度信息计算相似度，再将各个维度的相似度进行聚合，得到问句与候选答案的整体匹配度。Dong 等人<sup>[96]</sup>提出了 MCCNN 模型，如图2–10所示，模型使用多个不同的卷积神经网络层对问句进行编码，从而得到问句针对不同信息的向量表达。将它们分别与答案的类型、谓词、上下文向量表达计算相似度之后，最终的匹配度为这些相似度分值的总和，使得模型在寻找最佳答案时能兼顾来自不同方面的匹配特征。

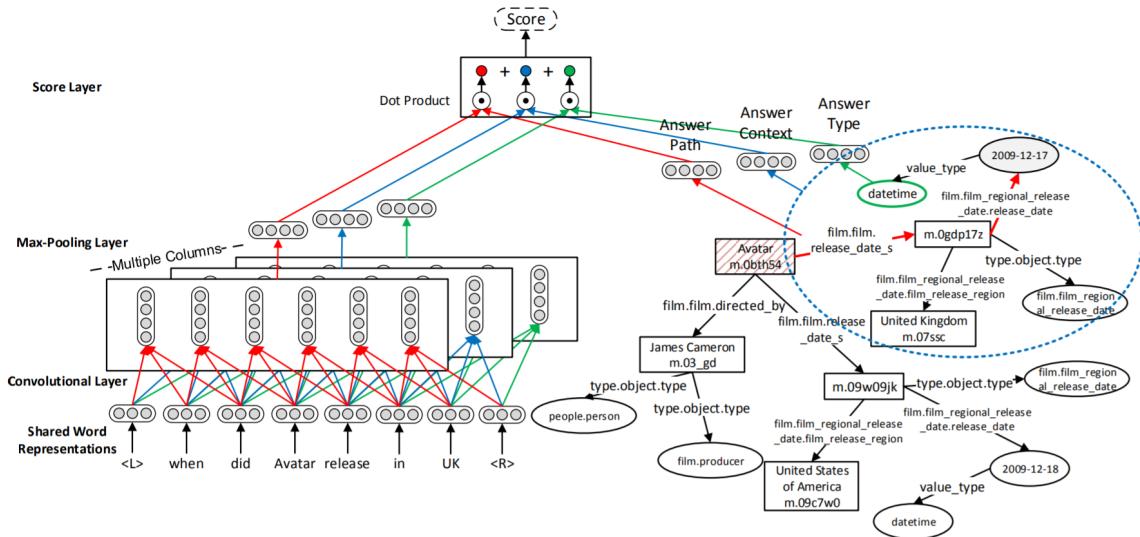
图 2-10 MCCNN 模型。<sup>[96]</sup>

Figure 2-10 The MCCNN model.

Hao 等人<sup>[37]</sup>的模型在 MCCNN 基础上进行了改良，除了将问句编码多个卷积层改为唯一一个双向 LSTM 层以外，主要的贡献在于模型中使用了问句和答案之间的双向注意力机制。一方面，针对答案在不同方面的表达，答案对问句的注意力能够动态调整问句中不同词的重要性，另一方面，问句对答案的注意力使得多个相似度分值互相之间也具有权重，模型训练效果要优于无差别的求和操作。

和语义解析模型比较，信息抽取模型实现了问答系统的端到端训练，直接以 < 问题，答案 > 作为训练数据，防止远距离监督引入错误。但同时也具有解释性较低的缺陷，无法直接输出模型所理解的问句语义结构，有时答案预测虽正确，但特征中可能存在语义偏差。对于复杂语义的自动问答研究，我们更在意语义结构的正确性，它能直接体现一个问答模型是否具有良好的语义理解能力。

## 2.4 本章小结

本章对实体、关系、问句理解这三个层面的研究进行了背景介绍和文献综述。实体理解方面，深度学习模型和跨语言词向量是我们较为关心的内容，将会在第三章的跨语言表格链接任务中使用。关系和问句理解方面，本章各介绍了两种路线不同的方法，分别是关系理解的规则推导、知识库向量表示，以及问句理解的语义解析、信息抽取。这四种方法之间存在着一些共性：规则推导和语义解析的共同点在于，语义理解需要显式的语义结构（一阶逻辑表达式，或与之等价的知识库子图）作为媒介；而另外两者的共

同点在于对实体、类型、谓词等知识库元素进行表示学习，以端到端的形式训练，模型更加面向具体任务。在第四章和第五章的研究中，我们更加在意机器是否能理解具有复杂语义的关系或问句，而不仅仅停留在特定任务的输出是否正确，因此规则推导和语义解析是本文关注的重点。



## 第三章 跨语言的表格实体链接研究

本章研究的实体链接任务中，待链接文本为以源语言编写的互联网表格，而知识库则以目标语言编写，因此我们将其称为跨语言的表格实体链接。为了捕捉不同于传统实体链接任务的特性，我们提出了基于神经网络和跨语言词向量的表格链接模型，旨在让不同语言的连续特征空间得以兼容，并捕捉表格具有的多种粒度的匹配特征。

### 3.1 概述

海量的互联网文本信息中，充斥着以 HTML 编写的表格，即互联网表格<sup>[97, 98]</sup>。和纯文本相比，互联网表格中的行列形式携带了非常有价值的结构化信息。为了能让机器理解，并且很好的处理表格中的信息，第一个步骤就是需要识别每个单元格中文本内容所对应的实体，并映射到一个标准词库，或是知识库上，例如维基百科或 Freebase。这样的一个在互联网表格上进行实体链接的任务，在本章节被称为表格链接<sup>[57, 99]</sup>。

对于表格链接任务，已有的研究工作<sup>[51, 57]</sup>主要针对英文表格，由于使用知识库也为英文，表格链接是在单一语言场景中进行的。然而，当需要链接的表格以其它语言编写的时候，对应语言的非英文知识库往往不够全面，无法涵盖目标表格中提及的所有实体。例如中文版维基百科，其中包含的实体（页面）数量仅为英文维基百科的 1/6 左右。基于不同语言知识库大小上的差异，本章探寻一种全新的方式将非英文表格与英文知识库相连，该任务也被称为**跨语言表格链接**。如图3-1所示，中文表格里的电影“邮差”在中文维基百科里没有对应的实体，但存在对应的英文维基实体“*Il Postino: The Postman*”，因此可以建立跨语言的链接。

帮助目标知识库补充事实三元组，是我们尝试跨语言表格链接的另一个动机。英文知识库比其它语言知识库更加庞大，也更加结构化，但仍然包含许多长尾实体。这些实体仅出现知识库的极少数事实三元组中，例如别国的电影、名人等，考虑到英文知识库的贡献者更多以英语为母语，这些实体的相关信息就很容易被忽略。另一方面，海量非英文的互联网表格成为了与长尾实体相关的丰富的语义信息来源。例如，图3-1描述了电影与它的原产国之间的关系。国产电影“线人”有对应的英文维基页面“*The\_Stool\_Piegon\_(2010\_film)*”，但与之相应的 Freebase 实体却缺少许多相关的知识。若我们准确将该电影链接至维基百科，并根据表格前两列的多个实体对推理出关系 *film\_country*，那么就可为知识库补充新的事实。

具体论述我们提出的跨语言表格链接方法之前，首先来讨论两种朴素的做法。第一

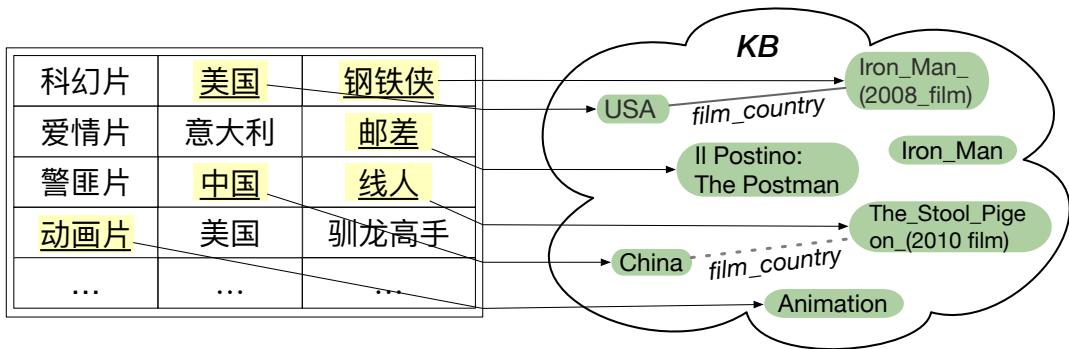


图 3-1 中文表格到英文知识库的跨语言链接示例。

Figure 3-1 Example of cross-lingual table linking from Chinese to English.

种方式主要基于已有单一语言的表格链接技术，将表格映射到语言一致的非英文知识库，然后再利用知识库之间存在的跨语言链接，将实体翻译至英文知识库。例如不同语言的维基百科之间就存在着人工编辑好的跨语言链接。这种方式的主要问题在于：1) 非英文知识库的信息量较低，可能无法覆盖每一个单元格的实体；2) 并不是每个非英文知识库都会存在和跨语言链接。

第二种做法中，整个非英文表格的内容首先直接被翻译成英文，然后整个问题便退化成英文上的表格链接，以往方法可以直接套用。它与远距离监督模型很相似，各单元格的(非英文名称，英文实体)对并不直接作为训练数据。此法的缺陷在于对已有翻译工具准确率的高度依赖：一方面，文本翻译过程仅生成单一结果，一旦错误则对后续链接步骤影响很大；另一方面，翻译工具如同黑盒，无法根据训练数据进行优化。

在本章中，为了使研究具有普适性，我们忽略不同语言知识库之间的跨语言链接，尝试在**不使用任何非英文知识库**进行过渡的情况下，解决跨语言的表格链接任务。据我们所知，本章节提出的解决方案，是对跨语言表格链接的第一次尝试。

对于实体链接任务而言，无论是否跨语言，第一个步骤总是为每个单元格生成一组候选实体，之后整个任务转换为排序问题，对每单元格寻找与其描述最接近的候选实体。主要的技术挑战在于表格描述和知识库来自不同的语言，无法依靠任何字面上的相似特征。此外，表格中缺少纯文本里的谓语、状语等相关上下文，给单个实体的消歧义带来了困难。

为了解决上述的两个挑战，我们提出了基于神经网络的联合模型来解决跨语言表格链接问题，它具有以下三个特点。首先，模型主体基于跨语言词向量，我们将单元格的描述短语、上下文、以及知识库的实体映射到不同语言对应的连续向量空间作为语义特征表示，并且使用线性变换的方式，实现不同语言的向量空间统一。其次，模型充分利用表格中同一行列的实体所具有的相关性，并通过神经网络学习不同粒度的相关性特

征。最后，模型基于联合训练思路，以优化整张表格的匹配程度作为目标函数，使用成对排序损失函数进行参数学习以及多轮迭代的预测方式，对新的表格完成链接。

本章的贡献可以总结为以下四个部分：

1. 我们首次尝试在跨语言场景上进行表格链接；
2. 我们提出了一个基于神经网络的联合训练模型，能有效捕捉原始表格与候选链接表格的语义相关性，并消除不同语言之间的语义间隔；
3. 联合模型除了捕捉单个单元格描述与候选实体间的语义关联特征，还提出了一种一致性特征，用于捕捉候选链接表格内部不同实体间的联系，有效提升模型的预测准确率；
4. 我们构建了从中文到英文的跨语言表格链接数据集用于实验，本章提出的模型效果显著优于其它基线模型，同时我们进行了一系列分析实验，以验证模型各部分的有效性。

## 3.2 相关工作

对互联网表格的研究最早开始于 Cafarella 等人的工作<sup>[97]</sup>，文中指出大约有 1.54 亿表格可以作为高质量的关系数据源。例如文献 [100, 101] 关注于从表格中寻找不同列之间的关系，从而实现向知识库中补充新的三元组。这些工作都假定实体链接已完成，而若要对更广范围的表格数据进行关系挖掘，表格链接始终是其前置步骤，链接准确度直接决定了后续步骤的质量。

和纯文本上的实体链接任务不同，表格文本上的链接聚焦于表格中的每一个单元格，并且对于任何一个待链接的单元格，其它同行或同列的单元格与其有着更加密切的语义联系。目前已有的表格链接研究主要基于特征工程。Limaye 等人<sup>[51]</sup> 以 YAGO 为知识库，解决更加宽泛的表格链接任务，包括将单元格链接至实体、列头链接至类型，以及两列之间的关系链接至谓词，同时创建了 WebManual 数据集。作者提出了一个概率图模型用于同时完成不同的链接子任务，并通过人为定义的多种势函数表示单元格、实体、类型、谓词语间的组合特征，整个表格链接的目标函数为多种势函数的连乘，不同子任务的决策互相影响，使得模型在捕捉单个单元格与实体相匹配的同时，也能兼顾实体与列头类型的一致性，以及不同列实体间与特定谓词的相关性。Bhagavatula 等人<sup>[57]</sup> 利用了表格上下文的词汇信息，对于待链接的单元格，将其行或列方向上的其它单元格文本合并形成上下文词袋，与候选实体所对应的词汇进行相似度计算，得到多个相似度特征用于模型训练，并采用迭代更新方式进行预测。Wu 等人<sup>[99]</sup> 首次尝试对中文表格进行链接，提出的模型首先构建由单元格和所有候选实体组成的连通图，然后在图中进行类似 PageRank 算法<sup>[102]</sup> 的随机游走，以选择最佳链接结果，因此是一种非监督学习方

式。候选实体是否同行列决定了图中是否存在直接相连的边，而单元格与实体、实体与实体之间所连边的权重则由预定义的相似度公式计算，使用了编辑距离、词袋相似度、实体于三元组中共现等特征。区别与以上研究，本文的工作基于深度学习，尝试不依赖常用的相似度计算公式，而是利用神经网络挖掘表格和目标实体在多个粒度上的特征。

跨语言的实体链接的主要目的是将文本中的实体短语链接至另一个语言构建的知识库上，近几年的 TAC-KBP 数据集<sup>[103-105]</sup> 中包含了跨语言的实体链接任务。为了解决此类问题，McNamee 等人<sup>[106]</sup> 提出了一种基线方法，利用已有的翻译工具将外文文本转换为英语，再使用传统的单语言链接模型完成任务。为了尽可能减少对翻译工具的高度依赖，模型需要能学习同一个实体或概念在不同语言下的抽象表达，并通过特定运算体现出不同抽象表达之间的联系，以完成语义的跨语言兼容。

基于跨语言词向量的链接模型是一种可行的解决方案，跨语言词向量的相关内容已在 2.1.3 节中介绍。Tsai 等人<sup>[107]</sup> 首先分别训练英文和外文的词向量，再用典型相关分析 (CCA) 学习各自语言的转移矩阵，使得不同语言词向量位于同一连续空间，之后依据该词向量计算短语和实体在不同粒度上下文中的余弦相似度，形成多个特征进行训练。Sil 等人<sup>[108]</sup> 提出了更加复杂的深度学习模型，以学习短语上下文和实体在句子级别和单词级别的相似特征，同时在实验中比较了 CCA、均方误差等多种生成跨语言词向量的方式。除了跨语言词向量以外，Zhang 等人提出的跨语言主题模型<sup>[109]</sup> 也可用于描述不同语言上的相同语义。传统的 LDA 主题模型<sup>[110]</sup> 旨在描述文档的语义表示，通过对“文档—主题”与“主题—单词”间的概率进行建模，将一个文档表示为抽象主题上的概率分布。考虑到同一个实体在不同语言中的维基页面，虽然单词不同，但其主题十分相似，因此双语 LDA 模型中，同一个抽象主题对应不同语言上的两个“主题-单词”概率分布，从而外语上下文和英语维基页面之间可以在主题层面上概率分布比较，实现链接过程。

本文的工作是表格链接和跨语言实体链接两者的综合体现，同时也是首次对此问题进行研究。

### 3.3 任务规范定义

输入的互联网表格  $X$  是一个具有  $R$  行和  $C$  列的矩阵，每一个单元格  $x_{ij}$  的内容是由语言  $L_1$ （例如中文）描述的词语序列。给定由另一种语言  $L_2$ （例如英文）编写，并包含大量实体  $e$  的知识库  $K$ ，跨语言表格链接的任务是寻找  $X$  对应的目标链接表格  $E$ ，使得链接表格中的每一个实体  $e_{ij} \in K$  对应单元格  $x_{ij}$  内容的消歧义表示。

在具体场景中，输入的表格包括一些无法被链接的单元格，例如数字、日期、时间以及一些知识库中尚不存在的新兴实体。一些已有工作<sup>[111]</sup> 主要负责在互联网表格中识别这些数字或时间实体，因此在本章中，我们不关注一个单元格是否能被链接的判断方

式。具体到任务定义中， $P$  为输入表格中所有可以被链接的单元格坐标  $(i, j)$  所构成的集合，并且我们假设在训练集和测试集中，每个输入表格  $X$  对应的可链接位置集合  $P$  都是已知的。

传统的实体链接方法通常在模型中定义一个评分函数  $S(x, e)$ ，用于衡量文本  $x$  与目标实体  $e$  之间的相关程度。在表格链接任务中，这样的做法等同于将不同的单元格分割开，单独计算相似度。然而缺陷在于，相邻或是同行列的目标实体之间的交互完全无法体现在链接模型中。为了将目标链接表格中不同实体间的耦合关系融入任务中，我们定义了在表格层面的评分函数，并以此预测最佳的链接表格  $\hat{E}$ ，如下所示：

$$\hat{E} =_{E \in GEN(X)} S(X, E), \quad (3-1)$$

其中  $GEN(X)$  表示由  $X$  生成的所有候选链接表格。该函数描述了输入表格与候选实体表格之间的整体相关性分数。

## 3.4 我们的方法

本节中，我们主要阐述使用联合训练模型解决跨语言表格链接的具体细节。图3-2为整个模型的示意图。之所以将整个模型成为“联合训练模型”，是因为神经网络的输入包含了整个互联网表格  $X$ ，以及对应的一个候选链接表格  $E$ ，而模型的输出代表两者的相关性分数  $S(X, E)$ 。

具体而言：1) 我们首先对表格中的每一个单元格内容生成一系列知识库中的候选实体；2) 模型对单元格词组和实体进行向量编码，并学习基于它们向量表示的**指示特征**以及**上下文特征**；3) 为了使不同语言下的语义向量互相兼容，模型利用双语翻译矩阵将向量表示从中文转为英文；4) 模型从候选表格  $E$  的内部学习第三类特征，即候选实体间的一致性特征。本节最后将介绍训练和测试的具体流程，以及整个模型中重要的一些实现细节。

### 3.4.1 候选实体生成

我们对中文表格  $X$  的每一个单元格内容生成一系列英文知识库中的候选实体。在本章的研究中，我们使用英文维基百科作为知识库。由于提出的方法不使用任何中文知识库进行过渡，为了实现语言转换，我们首先利用已有的翻译工具生成中文词组对应的多种翻译结果。接下来，对于每一个翻译结果，我们都使用预先定义的启发式规则，将英文词组转换为候选实体。这些实体的来源主要包括：1) 名称与翻译完全匹配的实体；2) 维基百科中，完全匹配的锚文本所指向的实体；3) 通过计算编辑距离（Edit Distance）进行模糊匹配，并且相似度足够高的实体。以中文词组“**疑犯追踪**”举例，不同的翻译

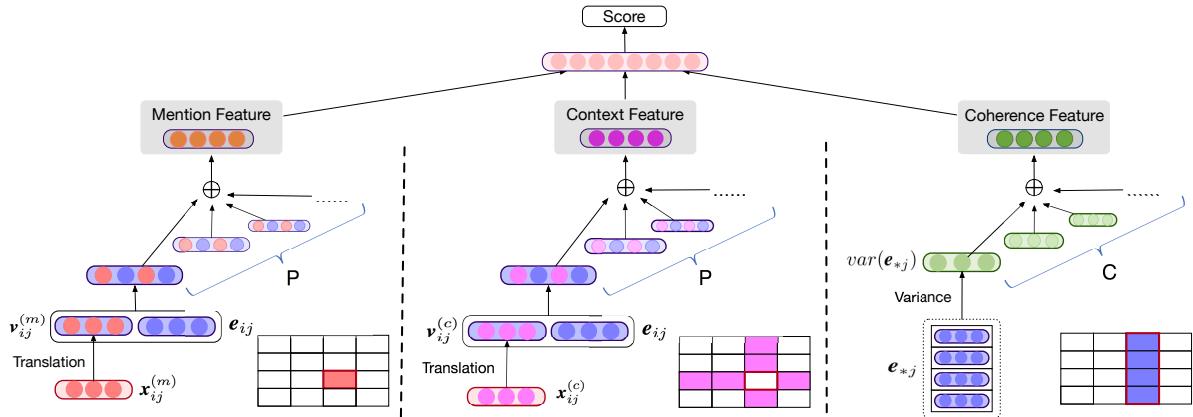


图 3-2 基于神经网络的联合训练模型示意图。

Figure 3-2 Overview of proposed neural network based joint model.

工具生成的结果不同，例如“person of interest”或者“suspect tracking”。整体候选实体来自于每一个翻译结果的映射，例如维基百科中的实体“person of interest”，“person of interest (tv series)”以及“suspect (1987 film)”。

### 3.4.2 向量表示及跨语言模块

给定一个单元格的字面描述短语  $x$ ，令  $\mathbf{x}^{(m)}$  代表其自身的语义向量，也称为**指示向量**。通常单元格字面描述较短（至多三个词语），因此模型计算字面描述包含的词向量的平均，作为  $\mathbf{x}^{(m)}$  的值。用  $\mathbf{e}$  表示候选实体  $e$  对应的实体向量，词向量和实体向量分别通过中文和英文的维基百科文本进行预训练。

考虑到语言的天生差异，且两者分别训练，因此词向量和实体向量所在维度空间并不兼容，这使得我们无法简单地对来自不同空间的向量进行比较和计算。为了应对这个问题，模型中引入了双语翻译层，将向量从一个语言的维度空间投影至另一个空间。 $\mathbf{x}^{(m)}$  为中文语义空间上对  $x$  的语义表示，该层通过线性变换将其映射为  $\mathbf{v}^{(m)}$ ，即英文维度空间上的语义向量： $\mathbf{v}^{(m)} = W_t \mathbf{x}^{(m)} + \mathbf{b}_t$ 。其中  $W_t$  为变换矩阵， $\mathbf{b}_t$  为偏置向量，两者均为模型参数，随着训练迭代而更新。

另外，我们通过少量的双语词对  $(w^{(ch)}, w^{(en)})$ ，对双语翻译层的参数  $W_t, \mathbf{b}_t$  进行预训练。预训练过程的损失函数定义如下：

$$L(W_t, \mathbf{b}_t) = \sum_i \|W_t \mathbf{w}_i^{(ch)} + \mathbf{b}_t - \mathbf{w}_i^{(en)}\|_2, \quad (3-2)$$

即最小化真实的英文词向量与线性变换后的词向量之间的欧氏距离。关于初始化，以及翻译预训练的更多细节，将在3.4.6节中进行叙述。

### 3.4.3 指示特征与上下文特征

如图3-2所示，最左边的部分对应指示特征模块，中间的部分对应上下文特征模块。两者的共同点在于，它们都关注互联网表格  $X$  与候选链接表格  $E$  之间的相似性或相关性，并且每个单元格各自计算的特征会聚合为一体。因此这两部分具有很相似的网络结构。

首先介绍指示特征，它捕捉一个单元格自身描述与目标实体的对应。给定字面描述  $x_{ij}$ ，我们将英文的指示向量  $v_{ij}^{(m)}$  与实体向量  $e_{ij}$  进行拼接，并送入全连接层，生成单元格在自身指示级别的隐含特征。收集所有需要被链接的单元格的指示特征，并对其进行平均，即可得到整张表格上的总体指示特征  $\mathbf{h}^{(m)}$ 。具体公式如下：

$$\begin{aligned} f_{ij}^{(m)} &= \mathbf{g}(W^{(m)}[v_{ij}^{(m)}; e_{ij}] + \mathbf{b}^{(m)}) \\ \mathbf{h}^{(m)} &= \frac{1}{|P|} \sum_{(i,j) \in P} f_{ij}^{(m)}, \end{aligned} \quad (3-3)$$

其中  $W^{(m)}$  以及  $\mathbf{b}^{(m)}$  为模型参数， $\mathbf{g}$  为非线性激活函数，实验中使用 ReLU 函数。

上下文特征的获取与指示特征类似。区别于指示特征的信息仅来自目标单元格，上下文特征还将考虑此单元格周围的有用信息。而在表格之中，位于同一行或同一列的其余单元格则具有直接的关联，因此成为上下文特征的信息来源。我们定义一个单元格的上下文向量  $\mathbf{x}_{ij}^{(c)}$  为这些相关单元格指示向量的平均：

$$\mathbf{x}_{ij}^{(c)} = \frac{1}{R+C-1} \left( \sum_{(i,k), k \neq j} \mathbf{x}_{ik}^{(m)} + \sum_{(k,j), k \neq i} \mathbf{x}_{kj}^{(m)} \right). \quad (3-4)$$

同样经过双语翻译层的转换，英文空间中每个单元格的上下文向量  $v_{ij}^{(c)}$  将用于生成整个表格的总体上下文特征，记做  $\mathbf{h}^{(c)}$ 。具体计算过程类似公式3-3，只需要把所有指示向量改为上下文向量作为输入即可。通过观察表格中的每个<字面描述，候选实体>对，并进行指示特征和上下文特征的学习，模型可以从两张表中捕捉大体上的语义相关程度。

### 3.4.4 一致性特征

前面叙述的两类特征都是对互联网表格与链接表格之间的契合度进行编码，另一方面，链接表格内部，不同实体之间的关系同样具有价值。之所以有这样的理解，是因为表格中同一列（有时同一行）的实体大多都属于同一种类型，也就是说，往往拥有更加相似的向量表达。例如概述部分的图3-1，表格中从左到右三列，对应的链接实体分别属于电影流派、国家、电影。我们提出的第三种特征，正是用来描述同一列候选实体之间的契合度。

关于同一类型的实体在表格中是按哪种方向进行排列，这涉及到另一个研究课题名为“表格类型分类”<sup>[112, 113]</sup>，主要用于区分表格的多种表现形式。本章中默认表格的形式为“垂直关系型”<sup>[113]</sup>，即和图3-1一样，相同类型实体按列方向排布。考虑到确定表格类型之后，大多数互联网表格都可以实现简单的格式转换，因此这个课题不在我们的讨论范围。

一致性特征的网络结构见图3-2的最右侧部分，为了衡量一列实体向量是否接近，我们对这些向量进行逐位的方差计算，方差越小，表明这些实体在对应位置的隐含语义上差别越小，反之亦然。同样对每一列的方差向量进行求平均的操作，我们便得到整个候选实体表格上的一致性特征  $\mathbf{h}^{(coh)}$ ：

$$\mathbf{h}^{(coh)} = \frac{1}{C} \sum_j \mathbf{var}(\{\mathbf{e}_{ij} | (i, j) \in P\}), \quad (3-5)$$

其中  $\mathbf{var}(\cdot)$  函数以向量集合作为输入，返回同样维度的逐位方差向量。一致性特征用于描述候选实体互相之间是否有良好的自我组织性，由于和字面描述表格  $X$  无关联，该特征可以看做对指示特征与上下文特征的补充。

### 3.4.5 训练及测试

我们首先定义输入表格  $X$  与候选链接表格  $E$  之间的整体相关性分数。前面提及的指示、上下文、一致性特征将被拼接，并送至一个两层的全连接网络得到总体特征  $\mathbf{h}_{out}$ ，第二层的输出维度为 1，即表示最终的表格相关度：

$$\begin{aligned} \mathbf{h}_{out} &= g(W_{out}[\mathbf{h}^{(m)}; \mathbf{h}^{(c)}; \mathbf{h}^{(coh)}] + \mathbf{b}_{out}) \\ S(X, E) &= \mathbf{u} \cdot \mathbf{h}_{out}, \end{aligned} \quad (3-6)$$

其中  $W_{out}$ ,  $\mathbf{b}_{out}$  以及  $\mathbf{u}$  均为模型参数。

训练集中的每一个互联网表格，都对应唯一一张正确的链接表格作为正样本。为了进行训练，我们需要准备若干张链接表格作为负样本。通过对正样本表格中的实体进行不同程度的篡改，我们可以自动生成一系列负样本表格，具体步骤如下：先随机指定要被篡改的单元格数量，再随机确定这些单元格在表格中的位置，最后将这些单元格的链接实体替换为对应候选集中的一个随机错误实体。这样可以使得篡改后的错误实体不至于太容易被发现。

训练过程中可能使用的更新方式有两种：基于最大间隔损失（Max Margin Loss，即 Hinge Loss），或者基于成对排序损失（Pairwise Ranking Loss）。对于前者，模型将最大化正样本表格与负样本表格间的分数差异。对于后者，单个正样本和多个负样本表格两两之间都会进行比较，具有更多正确链接实体的表格，要尽可能比另一张表格获得更高

的相关度分值。本章提出的模型采用了 RankNet 算法<sup>[114]</sup> 计算成对排序的损失函数，并使用 Adam 算法<sup>[115]</sup> 进行梯度下降。

测试过程涉及到更多的细节。理想状态下，对于互联网表格  $X$ ，我们需要枚举每一张链接表格  $E \in GEN(X)$ ，才能得到全局最优解。然而，候选表格集的数量与单元格的数量呈指数相关，同时每一个单元格又能对应大量候选实体，因此暴力枚举显然是不现实的。

---

### 算法 3-1 基于局部搜索下降的预测过程

---

**Input:** Mention table  $X$ , linking position  $P$ , initial entity table  $E_0$ ,

candidate generator  $Cand(\cdot)$ , scoring function  $S(\cdot, \cdot)$

**Output:** Entity table  $E$

```

1: procedure PREDICT( $X, P, E_0, Cand, S$ )
2:    $E \leftarrow E_0$ 
3:    $s_{max} \leftarrow S(X, E_0)$ 
4:   repeat
5:     Shuffle  $P$ 
6:     for  $(i, j)$  in  $P$  do
7:        $E' \leftarrow E$ 
8:       for  $ent$  in  $Cand(x_{ij})$  do
9:          $e'_{ij} \leftarrow ent$ 
10:         $s' \leftarrow S(X, E')$ 
11:        if  $s' > s_{max}$  then
12:           $e_{ij} \leftarrow ent$ 
13:           $s_{max} \leftarrow s'$ 
14:        end if
15:      end for
16:    end for
17:    until  $s_{max}$  converges
18:    return  $E$ 
19: end procedure

```

---

为此，我们使用局部搜索下降（Local-Search Descent）算法来逼近最优的链接表格。如算法3-1所示， $E_0$  为链接表格的迭代更新起点，每个单元格填充由生成器  $Cand(\cdot)$  产生的候选集中最可能的实体， $S$  为已学习的评分函数。预测步骤将以迭代形式进行。迭

代的每一轮中，所有需要链接的单元格按照乱序进行一一访问（第6行），对每一个被访问的单元格，预测算法固定其余单元格的链接结果不变，从该单元格的候选实体中，选择达到局部最优相关性分值的实体，并更新输出表格的对应位置（第12行）。迭代过程将持续进行，直到某一轮结束之后，输出表格  $E$  的相关性分数无法进一步提高。该算法可以类比为离散环境下的随机梯度下降，每个单元格的候选实体视为变量，输出表格的分值沿它们的离散梯度不断上升，打乱单元格的访问顺序则提供了随机扰动，防止预测过程陷入局部最优点。

### 3.4.6 模型实现细节

模型的主要实现细节包括了候选生成过程，双语翻译层的预训练，以及调参细节，下面将分别对这几个部分进行介绍。

**候选生成：**我们使用百度翻译<sup>1</sup>，谷歌翻译<sup>2</sup>以及腾讯翻译<sup>3</sup>的 API 用于候选生成。获取翻译结果之后，我们将英文字面描述与维基百科中的每一个实体进行比较，计算粗略的链接置信度。若某实体名称与字面描述完全匹配，或存在字面完全匹配的锚文本指向该实体，则将其置信度设为 1。对于非完全匹配的情况，我们去掉字面描述和锚文本中的所有停用词，并计算 Jaccard 相似度，作为字面描述与对应实体的链接置信度。综合各种可能的英文翻译，根据链接置信度对所有实体进行排序，排名前  $N_{cand}$  的实体将被保留，作为原字面描述的候选集。

**双语翻译层预训练：**我们利用必应翻译<sup>4</sup>的 API 收集了一个双语词典，其中包含 91,346 个单词级别的中英文翻译对，并且每对都关联了一个 0 到 1 范围的置信度。为了从中选取有价值的信息，我们保留那些置信度高于 0.5，且中英文词语均完全匹配某维基百科实体的翻译对。经过此法，我们总共收集了 3,655 个翻译词对用于转换矩阵的预训练。

#### 调参细节：

- 每个单元格对应的候选实体数量 ( $N_{cand}$ ) 的调参范围为 {1, 3, 5, 10, 20, 30, 40, 50}；
- 每个训练表格所生成的负样本表格数量 ( $N_{tab}$ ) 范围为 {9, 19, 49, 99}；
- 模型中，指示、上下文、总体特征对应向量的维度 ( $d_{cell}$ ,  $d_{cont}$ ,  $d_{out}$ ) 范围为 {20, 50, 100, 200}；
- 学习率  $\eta$  范围为 {0.0002, 0.0005, 0.001}；

<sup>1</sup><http://fanyi.baidu.com>

<sup>2</sup><http://translate.google.cn>

<sup>3</sup><http://fanyi.qq.com>

<sup>4</sup><http://www.bing.com/translator>

- 我们在每一个隐含特征计算上使用 dropout 层<sup>[116]</sup>，保留概率  $p$  范围为 {0.5, 0.6, 0.7, 0.8, 0.9}。

## 3.5 实验

本节中，我们首先介绍用于实验的跨语言表格链接数据集，以及已有的基线方法，这些方法主要是由单一语言上的实体链接方法转换而来。我们在跨语言以及单一语言场景下进行了端到端测试，并且通过横向对比实验分析方法中不同模块的重要性。

### 3.5.1 实验设置

**词向量、实体向量学习：**我们使用 2017 年 2 月版本的中文与英文维基百科<sup>1</sup>语料库，用于学习模型中的词向量与实体向量。语料库中包含 5,346,897 个英文实体以及 919,696 个中文实体。为了学习每个实体向量，我们将维基百科中的锚文本替代为一个特殊词语，与背后的实体一一对应。例如英文句子 “the Rockets All-Star player James Harden ...” 中，锚文本 “Rockets” 对应的实体为 “Houston Rockets”，因此我们使用与之对应的特殊词语 “[ [Houston\_Rockets] ]” 替代锚文本。这样处理的好处在于，实体和普通词语之间无差别，英文的词汇和实体用同一连续语义空间进行表达，这也使得模型经过翻译层后，更容易捕捉字面描述与实体间的相关性。预训练过程采用 Word2Vec[58] 分别学习中文和英文语料库上的词向量，特殊词语向量即为对应实体向量。预训练的词向量维度设为 100。

**表格链接数据集：**用于实验的跨语言表格数据集包含 150 个中文字面描述的互联网表格，以及对应的链接表格，标注的实体来自于英文维基百科。大部分表格来自 Wu 等人的研究 [99]，其公布的数据集包含 123 张中文表格，以及映射到中文维基百科上的实体。我们在互联网中收集了另外 40 张大小相似的中文表格，再利用维基百科的跨语言链接以及人工标注，生成所有的英文链接表格。大约 81% 的单元格可以找到对应的英文实体。我们过滤掉表格过小，或可被链接的单元格过少的表格。最终数据集包含 150 张表格，共有 3,818 个单元格，其中 2,883 个单元格标注了链接实体，平均每张表格包含 19.22 个链接实体。我们将数据集随机分为训练集、验证集和测试集，比例为 80: 20: 50。

### 3.5.2 基线模型

由于之前并没有直接针对于跨语言场景的表格链接工作，因此我们从两个角度出发，根据已有工作构建用于比较的模型。

<sup>1</sup><https://dumps.wikimedia.org/zhwiki/>，以及<https://dumps.wikimedia.org/enwiki/>。

第一个方向是单语言的表格链接系统，我们主要关注 Bhagavatula 等人<sup>[57]</sup>以及 Wu 等人<sup>[99]</sup>的工作。这两个系统分别在英文表格链接与中文表格链接上取得了不错的结果，分别简写为 *TabEL\_B* 以及 *TabEL\_W*。为了使这两个系统能在跨语言场景中进行测试，我们通过单一翻译工具将输入中文表格转换为英文，这样整个实验变成了单语言的场景，两个系统可以直接运行。

第二个方向是跨语言的实体链接系统，我们与 Zhang 等人<sup>[109]</sup>的工作进行比较，简写为 *TextEL*。该方法对 LDA 主题模型<sup>[110]</sup>进行改进，称为双语 LDA 模型。其核心在于同一个隐含主题具有两个不同语言上的词汇概率分布，通过比较字面描述上下文与候选实体在主题概率分布上的相似度，确定最佳的链接结果。为了将该模型用于表格上的实体链接，我们将表格按行遍历方向展开成普通文本，并标记文本中所有需要被链接的短语位置。经过此法，*TextEL* 可以在文本中捕捉更灵活的上下文信息，但有可能丢失列方向上实体相关的特性。

### 3.5.3 实验结果

#### 3.5.3.1 候选实体生成测试

本节中，我们关注将中文字面描述翻译为候选实体的精准度。根据3.4.6节中的介绍，我们使用了三种不同的翻译工具用来生成候选实体。我们通过 Hits@n 指标来衡量候选生成结果的好坏，以比较不同翻译工具带来的差别。Hits@n 的定义为正确的英文实体出现在前 n 个候选实体中的单元格比例。具体比较结果如表3-1所示，从中观察可知，百度翻译的结果稳定优于另外两者，而当所有翻译工具全部使用时，相比百度翻译结果，Hits@5 和 Hits@10 都能稳定增长约 4%。这说明了多个翻译工具之间互相补充，有助于发现更多正确的实体，同时有效的字面相似度的候选排序避免了过多错误的候选实体被引入。

表 3-1 候选生成步骤的 Hits@n 测评结果。

Table 3-1 Hits@n results on candidate entity generation.

Resources	n=1	n=5	n=10
Baidu	0.542	0.669	0.684
Google	0.463	0.585	0.596
Tencent	0.394	0.510	0.522
All Used	<b>0.558</b>	<b>0.708</b>	<b>0.726</b>

### 3.5.3.2 端到端测试

本节中，我们将与其它基线模型  $TabEL_B$ ,  $TabEL_W$  和  $TextEL$  在跨语言场景上进行端到端测试。与已有工作的实验保持一致，我们使用的评价指标为微观准确率（Micro Accuracy）和宏观准确率（Macro Accuracy）。微观准确率统计所有测试表格中，实体链接正确的单元格比例，而宏观准确率定义为每个表格各自链接准确率的平均值，避免了评价指标倾向于更大的表格。

由于  $TabEL_B$  和  $TabEL_W$  仅通过一种翻译工具生成输入表格的英文描述，出于公平考量，我与基线模型的比较实验均仅使用百度翻译。与此同时，我们也评估使用所有翻译工具，并且进行预训练的模型准确率。基于测试集上微观准确率的调参，我们使用的模型超参数为  $N_{cand} = 30$ ,  $N_{tab} = 49$ ,  $d_{cell} = d_{cont} = 100$ ,  $d_{out} = 200$ ,  $\eta = 0.0002$  以及  $p = 0.9$ 。

表3-2显示了端到端实验的比较结果。首先关注上面四行仅使用百度翻译的实验，我们模型的大幅度优于其余基线模型，准确率得到了约 12.1% 的相对提升。在此基础之上，使用多个翻译工具模型将微观准确率提升了 0.03，再次表明翻译工具之间的互补性给整个系统带来的帮助。双语翻译层的预训练步骤同样具有明显效果，进一步将微观准确率提升了 0.023。基于单语言表格链接模型的  $TabEL_B$  与  $TabEL_W$  受困于翻译过程带来的局限性：实体预测结果严重依赖唯一的英文翻译，一旦出现偏差便很难纠正，整个系统容错率较低。由于模型的后续训练切断了与原始中文描述之间的联系，这导致了翻译步骤无法收到训练数据提供的反馈，因此错误只能在模型中传播。作为对比，我们提出的模型利用多种英文翻译生成大量候选实体，并将原始中文描述作为输入学习特征表示，尽可能减轻了翻译过程的信息流失。

表 3-2 跨语言表格链接的测试结果，基线模型仅使用百度翻译工具。

Table 3-2 Cross-lingual table linking results. All baselines take Baidu as the only translating tool.

Approach	Micro Acc.	Macro Acc.
$TabEL_B$	0.512	0.507
$TabEL_W$	0.514	0.519
$TextEL$	0.472	0.458
Ours (Baidu Only)	0.576	0.573
Ours (Full, - pre-train)	0.606	0.591
Ours (Full, + pre-train)	<b>0.629</b>	<b>0.614</b>

接下来，我们进一步分析候选实体数量  $N_{cand}$  将对模型效果产生怎样的影响。显而

易见的是，一方面随着  $N_{cand}$  增大，候选实体中包含正确实体的概率也随之增大，意味着模型准确率的理论上限将会提高，而另一方面， $N_{cand}$  增大会引入更多干扰实体，整个系统也就更难达到理论上限。我们在不同的模型上改变  $N_{cand}$  值，进行了多组比较实验，微观准确率结果如图3-3所示，图中标出了微观准确率的理论上限。我们的方法在不同大小的候选数量上均有良好的适应性，随着  $N_{cand}$  增大，一直保持着稳定的效果提升。 $TabEL_B$  的效果比较稳定，但带有微小的准确率下降。而  $TextEL$  结果出现了急剧下降，拐点位置的候选数量甚至没有超过 10。我们认为主要原因在于双语 LDA 模型基于无监督学习方式，它没有获得任何直接的 < 中文描述，英文实体 > 信息用于训练，因此对干扰实体的数量非常敏感。

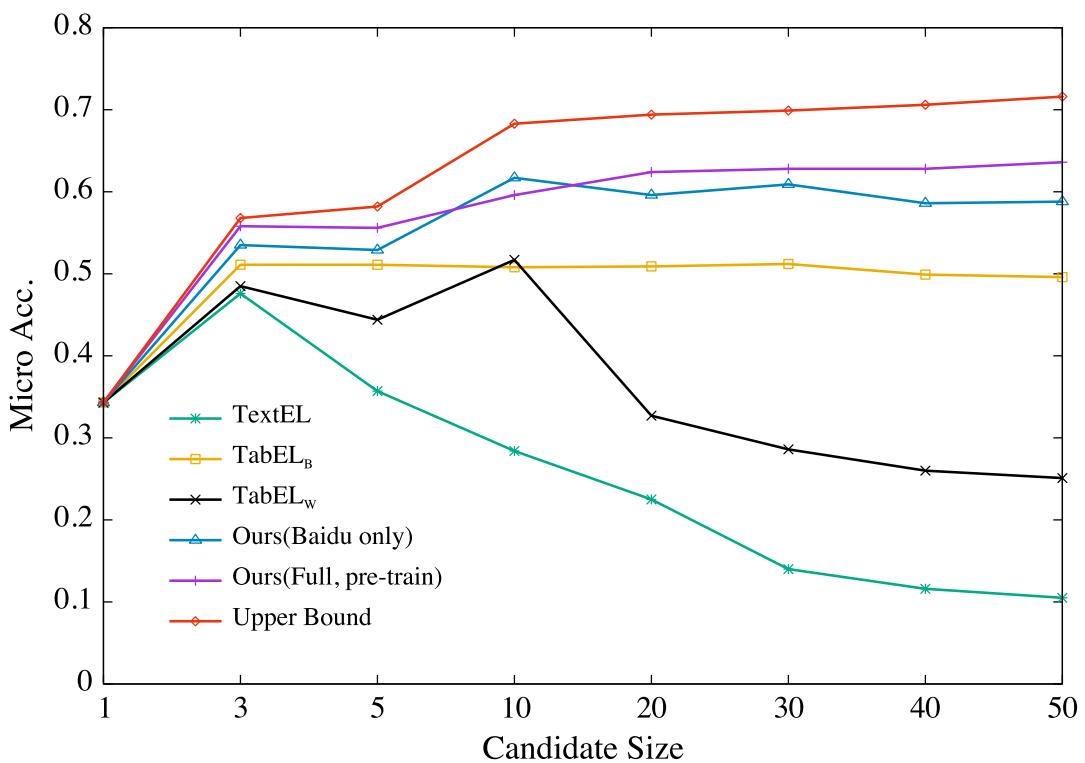


图 3-3 微观准确率随候选实体数量  $N_{cand}$  的变化情况。

Figure 3-3 Results of Micro Accuracy by different size of candidates.

为了更好地证明模型的有效性，我们对单语言场景的表格链接也进行了测试。由于跨语言的数据集利用中英文维基百科之间的链接构建，因此只需把标注实体替换为对应中文维基实体即可。相应地，我们从模型中移除双语翻译层，并保持其余设置不变。用于比较的系统依然为  $TabEL_B$  和  $TabEL_W$ ，两者均为表格链接的代表工作，其中后者为中文表格链接任务的最好结果。表3-3列出的实验结果显示，我们模型的单语言版本依

然优于两个基线系统，这在一定程度上说明了基于神经网络的联合训练模型的有效性，可以从表格的行列之中捕捉有意义的语义信息。

表 3-3 中文环境下的表格链接准确率。

Table 3-3 Accuracies on Chinese mono-lingual table linking.

Approach	Micro Acc.	Macro Acc.
$TabEL_B$	0.848	0.845
$TabEL_W$	0.852	0.848
Ours-mono	<b>0.886</b>	<b>0.868</b>

### 3.5.4 模型分析测试

本节中，我们对提出的神经网络模型进行了更加详细的实验分析，探索模型中指示、上下文、一致性特征各自的贡献程度，以及联合训练的模型架构所带来的优势。

#### 3.5.4.1 三类特征的作用

为了研究模型涉及到的三种不同特征的贡献程度，我们使用不同的特征组合，在跨语言场景中进行对比测试。比较结果如表3-4所示，模型中的每一类特征都对最终的准确率起到积极贡献。其中，指示特征是最重要的特征，因为它提供了字面描述与目标实体之间最为直接的信息。上下文特征的作用也十分明显，在维基百科中，实体对应的锚文本周围很可能出现与其同行或同列的其它描述，因此基于 CBOW 或 Skip-Gram 训练的实体向量包含这些上下文的语义。我们观察到，如果仅使用一致性特征进行训练，准确率的降低十分明显，约为 59.6%，这主要是因为模型难以获取实体与字面描述之间，最主导和直接的语义关联。但这并不影响一致性特征对指示及上下文特征的补充，若去除该特征，模型准确率将相对下降约 6%，依然是不小的差距。一致性特征旨在从全局角度发现实体之间的潜在关联，用来表征同一列实体之间是否具有一致性，例如隶属于相同的维基百科分类标签。即便模型没有直接利用每个实体的分类标签信息，一致性特征依然可以在向量表示中寻找依据。

我们用本章开头的图3-1举例讨论一致性特征的有效性。第三列中字面描述“钢铁侠”具有很高的歧义，在维基百科中，它可以对应超级英雄“Iron\_Man”，也可以对应电影“Iron\_Man\_(2008\_film)”。作为对比，“驯龙高手”（“How\_to\_Train\_Your\_Dragon\_(film)”）以及“线人”（“The\_Stool\_Pigeon\_(2010\_film)”）相对来说歧义较小。若只使用指示特征和上下文特征，模型预测的实体为超级英雄，考虑到钢铁侠在更多文本中确实代表超

表 3-4 不同特征组合在验证集上的跨语言链接准确率。

Table 3-4 Ablation test of feature combinations on validation set.

Feature Combination	Micro Acc.	Decrease in Acc. (%)
Mention Only	0.604	12.7
Context Only	0.576	16.7
Coherence Only	0.279	59.6
Mention + Context	0.652	5.78
Full	0.692	0.00

级英雄，因此这样的预测结果可以理解，但却是错误的。当一致性特征引入之后，联系其它两个歧义较低的实体，同一列实体之间强烈的相关性使得模型倾向于这一列都预测电影，因此模型能够实现正确的预测。

#### 3.5.4.2 联合模型的作用

这部分将验证整个联合模型框架的作用。相对于联合模型计算整个输入表格与链接表格的相关度，非联合模型中，单元格之间完全独立，各自计算字面描述与候选实体的匹配程度，最后求平均得到整张表格上的相关度。我们将联合模型进行退化，由于非联合模型仅考虑单个单元格，我们移除模型中的一致性特征模块，并无需对不同单元格的特征输出求平均。作为对比实验，我们同样从已有的联合模型中移除一致性特征，并尝试分别使用 RankNet 模型或最大间隔损失（Max Margin）进行训练。

表 3-5 不同模型训练方式在测试集上的跨语言链接准确率。

Table 3-5 Ablation test of train strategies on validation set.

Model	Optimizer	Coherence	Micro Acc.
Non-Joint	Max Margin	N	0.586
Joint	Max Margin	N	0.574
Joint	RankNet	N	0.598
Joint	RankNet	Y	0.629

表3-5列出了这一部分实验在测试集上的微观准确率结果。对比前两行结果，我们可以发现，若使用最大间隔损失，非联合模型的效果反而优于联合模型。主要原因有以下两点：1) 非联合模型中，每一个单元格的多个负样本实体都能在训练过程中被利用，而对于联合模型，由于负样本表格的生成依靠随机采样，并不是所有的负样本实体都会

被使用；2) 最大间隔损失侧重于正样本表格与不同负样本表格间的分值差距，而对于不同错误程度的负样本表格之间，它们的偏序关系并没有被有效利用。因此，相比于最大间隔损失，基于成对计算损失的 RankNet 更加适合于联合模型。此外，在算法运行速度方面，非联合模型无需迭代预测步骤，因此显然比联合模型更高效。而实验过程显示，联合模型平均只需要 6 轮迭代即可完成对每个测试表格的链接预测，是一个可以被接受的运行速度。

### 3.6 本章小结

据我们所知，本章的工作是首次提出了在跨语言场景中进行互联网表格的实体链接问题。为了使问题尽可能通用，本文研究在不利用任何非英文知识库作为过渡的情况下，完成非英文表格到英文知识库（维基百科）的链接。为此，本文提出了一个基于神经网络和跨语言词向量的链接模型，并利用模型学习三种不同粒度的链接特征，分别为单元格自身与目标实体的指示特征，单元格所在行列与目标实体的上下文特征，以及同一列目标实体之间的一致性特征。同时模型遵循联合训练框架，定义整张表格级别的链接匹配程度作为目标函数，并使用迭代更新方式完成所有单元格的链接。本文提出的模型在跨语言表格链接任务中取得了 63% 的准确率，考虑到此任务比单语言链接更具有挑战性，本文对后续的研究而言是一个良好的开端。在不同设定上的多组对比实验显示，三种粒度的特征对模型均起到明显效果，同时联合训练框架也具有实质性的帮助。

后续的研究主要包括对表格中的单元格判断是否需要被链接，本文的任务定义移除了这个问题带来的影响，但显然，不可链接的单元格在互联网表格中也会普遍存在，因此该研究具有其实际意义。

本章的研究成果已发表于 2018 年国际会议 Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)，论文题目为“Cross-Lingual Entity Linking for Web Tables”。



## 第四章 自然语言关系的语义理解研究

本章的研究中，我们关注从海量纯文本数据中挖掘出的关系三元组。二元关系是一个三元组的语义核心，它扮演谓语的成分，描述了主语和宾语实体间具有的特定联系。然而，由于关系具有多义性，以及知识库与自然语言间存在的语义间隔，我们很难直接像实体理解那样，建立关系和知识库谓词的一一对应。因此，我们尝试从多个角度出发，寻找关系与知识库之间存在的复杂匹配。

### 4.1 关系的主宾语类型搭配挖掘

这一节的研究中，我们旨在寻找不同关系连接的实体所具有的类型偏好，并利用知识库中的实体信息构建丰富的类型层次关系，从而挖掘具有代表性的（主语，宾语）类型搭配，在粗粒度上展现关系的不同含义。

#### 4.1.1 引言

开放式信息抽取（Open Information Extraction）任务的目标是从开放领域的文本语料库中挖掘命名实体或概念，并抽取出连接这些实体的各种不同的自然语言关系。之所以称为开放式抽取，是因为要挖掘的关系不局限于特定领域也不基于固定的匹配规则。学术界中，较为先进的开放式信息抽取系统<sup>[1-4]</sup> 可以从海量互联网语料库中，以很高的准确率提取百万甚至更高级别数量的关系实例， $(arg_1, rel, arg_2)$  三元组形式，我们将其称为关系三元组。其中， $rel$  为二元关系，一般表示为短语（词级别描述）或依存语法路径（语法级别描述）。 $arg_1$  和  $arg_2$  是关系的两个参数，即主语和宾语，同样表现为短语形式。

开放式信息抽取提供给我们海量关系实例的同时，我们有兴趣将这些实例进行归纳，寻找更加抽象的语义表示。我们关注的重点就是这些关系所具有的不同含义。以关系“play in”为例，开放式信息抽取系统可以提供一系列具有 $(X, play in, Y)$  形式的三元组。例如 ReVerb 系统<sup>[2]</sup> 可抽取出三元组 (Goel Grey, played in, Cabaret) 以及 (Tom Brady, play in, National Football League)。给定某关系已有的三元组实例，我们可以推理出一系列由类型三元组描述的关系模式，即主宾语类型搭配 $(t_1, play in, t_2)$ 。其中  $t_1$  以及  $t_2$  为标准化的实体类型，其来源为含有类型定义的知识库，例如 WordNet<sup>[5]</sup>，Yago<sup>[117]</sup>，Freebase<sup>[9]</sup> 以及 Probable<sup>[118]</sup>。每一个关系模式都可以用来表示一组特定的“play in”关系实例，其中主宾语分别属于对应的类型。对于上例“play in”，我们可以给出两个可能的

模式： $(film\_actor, play\ in, film)$ ，以及 $(pro\_athlete, play\ in, sports\_league)$ 。由此可见，二元关系“play in”具有明显歧义，不仅可以描述“运动员—体育联盟”联系，还可以描述“演员—电影”之间的联系。对于歧义较少的关系，我们依然可以推理出不同的主宾语类型搭配，例如关系“is the mayor of”可以推理出 $(person, is\ the\ mayor\ of, location)$ ，以及 $(politician, is\ the\ mayor\ of, city)$ 等不同模式，在类型上具有不同的粒度，后者显然更加具体。

对于自然语言理解任务，例如上下文相关的实体消歧，还有开放领域自动问答，关系模式是一个有用的信息。假设我们要对句子“*Granger* played in the NBA”进行实体识别。“*Granger*”对应一个人名，但由于只提供了姓氏，因此具有较高歧义。而“the NBA”几乎可以确定是人们熟知的体育联盟。再结合上面列举的“play in”所具有的关系模式，实体识别模型便可以获得额外特征，即“*Granger*”更有可能代表运动员，也就使得篮球运动员“*Danny Granger*”更容易被正确识别。考虑到这个实体并不非常著名，与之相关的关系实例数量可能较少，但类型特征依然可以提供很大的帮助。

为了生成关系模式，一种已有的方案是基于选择偏好（Selectional Preference）技术<sup>[119-121]</sup>，它可以帮助对关系中的主宾语实体计算各自具代表性的类型。选择偏好技术主要思路来自关系与类型之间的互信息计算<sup>[120]</sup>，这种方式倾向于选择当前关系所独有的类型，换句话说，如果一个类型普遍适用于不同关系中的实体描述，那么它便不容易被选为代表类型。然而在开放式信息抽取中，很多关系实际上是相关的，甚至非常相近，例如“play in”，“take part in”以及“is involved in”。这些关系实际上具有相同的语义，因此主宾语的类型搭配也应该相似，而选择偏好技术会因为关系的不同而对这些类型都进行弱化。

因此本章中，给定一个关系和一系列具体的三元组，我们的任务是寻找那些最具体的类型搭配，而同时包含尽可能多的关系实例。我们的方法首先将关系实例中的主宾语映射为知识库中的实体，即为每个三元组生成 $(e_1, e_2)$ 实体对。接着根据不同实体所属的类型，寻找可以覆盖尽可能多实体对的类型搭配 $(t_1, t_2)$ 。最后，当不同的类型搭配覆盖的实体对较为接近或一致时，我们利用知识库中已有的 *IsA* 关系，扩充知识库中类型之间的层次结构，以此寻找更加具体的类型搭配。

本章的贡献可以总结为以下三个部分：

1. 我们具体定义了基于开放式信息抽取的二元关系模式推理问题；
2. 我们设计了基于 Freebase 和实体链接任务的方法，对一类关系的主宾语所具有的类型分布进行联合建模；
3. 我们在 ReVerb 数据集上进行实验，根据人工标注的类型搭配结果，对不同二元关系生成的最佳模式进行测评。与传统选择偏好方法比较，我们的模型在 MRR

指标上得到了 10% 的相对提升。

#### 4.1.2 我们的方法

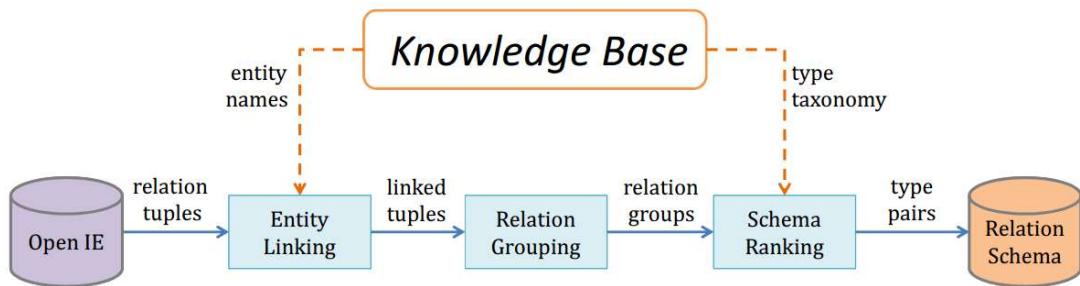


图 4-1 二元关系模式挖掘的流程框图。

Figure 4-1 System architecture of type inference of binary relations.

二元关系模式挖掘的系统架构如图4-1所示。整个系统的输入为开放式信息抽取系统中的所有关系三元组，经过实体链接、关系分组以及模式排序三个步骤之后，这些三元组将会转换为一系列排好序的主宾语类型搭配。每个步骤概括如下，本节将对它们进行具体描述。

**(1) 实体链接：**关系三元组中的参数实体均为字符串形式。我们通过模糊字符串匹配的方式，将主宾语分别映射到知识库中的不同实体。

**(2) 关系分组：**经过链接之后，关系表达形式相近的三元组将聚集在一起，形成一个大的分组。并且，每一个分组会从内部的不同关系中选择一个，作为整组的代表关系。

**(3) 关系模式排序：**对分组内的每一个具有链接的关系实例，其主宾语将转换为知识库中对应的类型。根据不同的类型搭配所覆盖的三元组数量，以及各个类型的宽泛或具体程度，对所有候选的关系模式进行排序并输出。

##### 4.1.2.1 实体链接

在实体链接步骤中，一个关系三元组的主宾语将分别映射到知识库中的实体，形成带链接的三元组  $(e_1, rel, e_2)$ ，并配有对应的链接分值。由于每一个三元组所具有的信息较少，并没有提供足够的上下文，因此实体链接过程主要基于主宾语名称以及实体在知识库中名称的模糊匹配。

实体在知识库中存在至多一个标准名称以及多个别名，例如 Freebase 中，实体的标准名称和别名分别对应 *type.object.name* 以及 *common.topic.alias* 属性。我们利用这些属性值构建了从单词指向不同名称的倒排索引，并进一步生成每个关系参数的候选实

体。我们用  $alias$  表示知识库中的一个名称（或别名），若将其看做单词的集合（bag-of-words），那么显然单词之间具有不同的重要性。直观上看，若  $alias$  中某单词  $w$  出现在极少数的名称中，那么它对整个名称而言更加重要；反之类似“of”，“the”等停止词会出现在大多数名称里，那么在模糊匹配的过程中，其权重就很低。因此我们利用文档频率倒数（Inverted Document Frequency）用于拟合单词  $w$  的权重：

$$idf(w) = 1 / \log(|\{alias : w \in alias\}|). \quad (4-1)$$

此外，我们直接从知识库的名称中过滤停止词，相当于它们的  $idf$  分值为 0。为了衡量关系三元组中的关系参数  $arg$  与知识库名称  $alias$  间的模糊匹配程度，我们计算两者之间的带权重叠分值：

$$overlap(arg, alias) = \frac{\sum_{w \in arg \cap alias} idf(w)}{\sum_{w \in arg \cup alias} idf(w)}. \quad (4-2)$$

对于候选实体  $e$ ，我们分别计算其不同名称与关系参数的模糊匹配分值，最终选取最高分代表实体  $e$  与关系参数  $arg$  的匹配度：

$$sim(e, arg) = \max_{alias \in Alist(e)} overlap(arg, alias). \quad (4-3)$$

为了控制候选实体的质量，对于由  $m$  个单词构成的关系参数（停止词忽略不计），我们仅考虑那些存在至少一个名称具有  $m - 1$  个单词重叠，同时模糊匹配度高于阈值  $\tau$  的候选实体。对于每个关系三元组中的主宾语，我们分别抽取匹配度排名前 10 的候选实体，用于后续的计算。

对单个关系参数进行匹配计算之后，我们将计算关系三元组  $(arg_1, rel, arg_2)$  与实体对  $(e_1, e_2)$  之间的联合匹配度。联合匹配度的定义方式有两种。第一种匹配方式较为朴素（Naive），仅考虑关系中的两个参数与各自实体的匹配程度，主宾语实体互相之间并无直接影响：

$$F(arg_1, e_1, arg_2, e_2, rel) = sim(e_1, arg_1) \cdot sim(e_2, arg_2). \quad (4-4)$$

第二种匹配方式除了考虑  $e_1$  和  $e_2$  各自的匹配分数，还考虑到了这两个实体之间存在的联系，在知识库上体现为连接它们的谓词或谓词序列。我们以  $\vec{w}$  表示  $rel$  的所有单词， $\vec{p}$  表示知识库中连接  $e_1$  和  $e_2$  的谓词路径，其长度至多为 2。若实体  $e_1$  与  $e_2$  可以通过长度为 1 的路径相连，则意味着知识库中存在通过某谓词  $p$  连接的事实三元组  $p(e_1, e_2)$ 。类似地，若  $e_1$  和  $e_2$  之间通过长度为 2 的路径相连，则意味着存在  $p_1, p_2$  以及中间实体

$e'$ , 使得事实  $p_1(e_1, e')$  以及  $p_2(e', e_2)$  存在于知识库中。我们利用朴素贝叶斯模型, 利用条件概率的形式定义谓词序列  $\vec{p}$  与关系  $\vec{w}$  之间的相关程度:

$$\begin{aligned} P(\vec{p} \mid \vec{w}) &\approx \prod_p P(p \mid \vec{w}) \\ &\propto \prod_p P(p) \prod_w P(w \mid p). \end{aligned} \quad (4-5)$$

Yao 等人<sup>[35]</sup> 将知识库谓词序列与关系的对应建模为机器翻译模型, 并根据对齐模型 IBM Model 1<sup>[122]</sup> 学习谓词的先验概率  $P(p)$  以及转移概率  $P(w|p)$ 。基于已有工作的概率模型, 给定关系后预测谓词序列的条件概率  $P(\vec{p} \mid \vec{w})$  便可计算得出。对于候选实体  $e_1$  和  $e_2$ , 它们之间的谓词序列与关系 *rel* 越接近, 则实体链接结果越有可能正确。因此, 我们通过枚举  $e_1$  和  $e_2$  之间所有满足长度条件的谓词序列, 计算关系实例与实体对之间的相似度:

$$F(arg_1, e_1, arg_2, e_2, rel) = sim(e_1, arg_1) \cdot sim(e_2, arg_2) \cdot \sum_{\vec{p}} P(\vec{p} \mid \vec{w}). \quad (4-6)$$

由于条件概率  $P(\vec{p} \mid \vec{w})$  的计算涉及到大量连乘, 其数值在不同实体对之间的差别较为明显, 这也使得其在公式4-6中具有较高的地位。而当所有候选实体间的谓词序列与当前关系都不相似的时候, 条件概率的随机波动反而会带来不小的干扰。因此, 我们采用了一种集成 (Ensemble) 方案: 首先定义条件概率阈值  $\rho$ , 对于当前关系实例的所有候选实体对, 若其中存在至少一条与关系足够相近的谓词序列, 即满足  $P(\vec{p} \mid \vec{w}) > \rho$  时, 模型使用公式4-6进行整体匹配度计算, 否则模型退回到公式4-4, 使用朴素的方式寻找最佳实体对。最后, 我们选择分数最高的实体对, 作为关系三元组的唯一链接结果。

#### 4.1.2.2 关系分组

这个步骤对所有已链接的关系三元组进行聚类, 拥有相似关系描述的三元组将归为同一分组。每个三元组仅存在于唯一一个分组中。

这个步骤的思路是通过语法转换, 将复杂的关系描述进行简化。如果两个不同的关系具有相同的简化形式, 那么视为其语义相同, 并归为同一分组。首先考虑到形容词、副词以及情态动词的存在与否, 基本上不会改变一个关系中主宾语实体所属的类型, 因此我们将这些词从关系描述中移除。此外, 大多数关系包含动词, 但时态并不一致, 因此我们将所有时态统一为现在时。此外, 关系中的被动语态将会被保留, 不做形式转变。例如经过语法转换之后, 下列关系实例将归为同一组: (X, *resign from*, Y), (X, *had resigned from*, Y) 以及 (X, *finally resignd from*, Y)。最后, 每一个分组的代表关系为组内关系的统一简化形式。如上例所示, 三个关系实例属于 “*resign from*” 组。

#### 4.1.2.3 类型搭配排序

给定一个关系分组  $r$ , 这一步骤将生成排好序的主宾语类型对, 即该关系的代表性模式。以二元关系“play in”举例, 理想情况下, 生成的结果里会包含模式  $\langle actor, film \rangle$  以及  $\langle pro\_athlete, sports\_league \rangle$ 。

对于带链接的三元组  $(e_1, rel, e_2)$ , 若在知识库中,  $e_1$  具有类型  $t_1$ , 而  $e_2$  具有类型  $t_2$ , 那么该三元组为类型搭配  $\langle t_1, t_2 \rangle$  的一个支持实例。一个实体有可能从属于多种类型, 无论类型宽泛或具体, 因此一个三元组可以支持多种类型搭配。对关系分组  $r$  中的所有实例进行处理, 我们可以得到每一种类型搭配所对应的支持集合:

$$sup_r(\langle t_1, t_2 \rangle) = \{(e_1, e_2) \mid (e_1, t_1) \in IsA, (e_2, t_2) \in IsA\}. \quad (4-7)$$

得到所有可能的类型搭配之后, 我们可以根据支持集合的大小进行排序。由于每个实体从属于多种类型, 因此显然更加宽泛的类型搭配通常会被排在前列。但是, 对于人类或是机器理解一个自然语言关系, 宽泛的关系模式所具有的信息量相对不足, 尤其是当两种类型对具有几乎一致的支持集合时, 往往更具体的类型对具有更好的代表性。例如对于关系“ $X die in Y$ ”, 在开放式信息抽取和实体链接均不产生错误的情况下, 类型对  $\langle person, location \rangle$  和  $\langle deceased\_person, location \rangle$  将对应完全一致的支持集合。后者对关系的描述更加具体, 在不丢失支持实例的同时, 尽可能缩小主语在知识库中的范围。

由此可见, 对候选类型对的排序需要考虑每个类型的相对粒度。接下来的目标就是提取知识库中类型之间的包含关系, 建立更加完整的层次结构。我们定义所有属于类型  $t$  的实体为  $cover(t) = \{e \mid (e, t) \in IsA\}$ 。理想情况中, 若  $t_1$  包含于  $t_2$ , 那么所有  $t_1$  中的实体都从属于  $t_2$ , 即  $cover(t_1) \subseteq cover(t_2)$ 。这样的包含规则称为“严格类型包含”。例如在 Freebase 中, 类型  $person$  所包含的其它类型包括  $actor$ ,  $politician$  以及  $deceased\_person$  等。

然而, 严格类型包含在知识库中并不多见, 主要原因是知识库的类型定义和人类对自然界的归纳存在一定差别, 以 Freebase 中的  $award\_winner$  为例, 类型中绝大多数实体都为自然人, 但依然包含少量的组织实体在内。基于严格类型包含的规则,  $award\_winner$  与  $person$  之间毫无包含关系, 但事实上, 考虑到非自然人实体仅存在极少数, 两个类别之间在很大程度上依然构成从属关系。另一方面, 由于实体的类型涉及到人工标记, 一旦出现类型标记错误, 就有可能导致类型之间无法满足严格包含条件。

为了能更好地建立类型层次关系, 我们使用一种更加松弛的类型包含定义方式。具体而言, 若  $t_1$  中足够数量的实体从属于  $t_2$ , 那么就认为包含关系成立。因此, 我们定义  $t_1$  包含于  $t_2$  的度, 即对应实体包含的比例:

$$deg(t_1 \subseteq t_2) = \frac{|cover(t_1) \cap cover(t_2)|}{|cover(t_1)|}. \quad (4-8)$$

若  $\deg(t_1 \subseteq t_2) > \epsilon$ , 则  $t_1$  包含于  $t_2$ 。阈值  $\epsilon$  表示松弛程度, 若  $\epsilon = 1$ , 则松弛包含退化为严格包含。若  $\epsilon$  太小, 那么类型之间将具有非常丰富的层次关系, 但其有效性则会下降。最后, 遍历知识库中所有的类型, 我们就可以得到特定松弛程度下的类型层次图。

随着类型层次关系建立完毕, 我们就可以定义不同类型搭配之间的包含关系。若类型对  $\langle t_1, t_2 \rangle$  被另一个类型对  $\langle t_3, t_4 \rangle$ , 则意味着以下条件之一成立: i)  $t_1 \subseteq t_3, t_2 \subseteq t_4$ ; ii)  $t_1 \subseteq t_3, t_2 = t_4$ ; iii)  $t_2 \subseteq t_4, t_1 = t_3$ 。最终的类型对排名体现为支持集合大小和类型对包含关系的共同作用。以支持集合降序排列为基础, 若类型对  $tp = \langle t_1, t_2 \rangle$  包含于另一个类型对  $tp'$ , 且各自的支持集合大小 ( $|sup_r(tp)|$ ) 几乎一致, 那么  $tp'$  将排在  $tp$  之前。我们同样可以根据重叠关系实例的覆盖程度, 来定义两个支持集合是否几乎一致:

$$\frac{|sup_r(tp)| - |sup_r(tp')|}{\max(|sup_r(tp)|, |sup_r(tp')|)} < \lambda, \quad (4-9)$$

其中  $\lambda$  为判断集合中的元素是否一致的阈值。

### 4.1.3 实验

#### 4.1.3.1 实验设置

我们在实验中使用的知识库为 Freebase<sup>[9]</sup> 在 2014 年 2 月 16 日的版本, 包含了大约 40,000,000 个不同实体, 以及 1,700 个主要类型。实验中使用的开放式信息抽取系统为 ReVerb<sup>[2]</sup>, ReVerb 数据集提供了多种版本, 我们使用的版本包含了置信度最高的 14,000,000 个关系三元组。

ReVerb 抽取的三元组中, 部分关系参数无法链接到 Freebase 中的某一个实体, 例如三元组 (*Metro Manila, consists of, 12 cities*), 其宾语显然不是一个实体, 而是用自然语言描述的类型。这部分三元组不是我们的研究对象, 需要进行过滤。考虑到在自然语言中, 概念通常对应非专有单词, 并且多为小写, 因此我们根据 WordNet 收集了常用的非专有单词。若一个三元组中包含纯小写, 或纯粹由非专有单词构成的主宾语, 那么该三元组将被过滤。除此之外, ReVerb 三元组中还具有时间或日期作为关系参数的情况, 例如 “Jan. 16th, 1981” 作为宾语, 但同样不对应 Freebase 的某个实体。为应对这种情况, 我们使用 SUTime<sup>[123]</sup> 工具识别时间或日期, 将它们替换为具有 *type.datetime* 类型的虚拟实体。经过清理之后, 系统共收集了 3,234,208 个三元组, 对应 171,168 个不同的关系分组。

实验中具体使用的参数值为:  $\tau = 0.667$ ,  $\rho = e^{-50}$ ,  $\epsilon = 0.6$  以及  $\lambda = 5\%$ 。关系分组步骤中, 我们使用 Stanford Parser<sup>[124]</sup> 对每个关系进行词性标注、语法分析以及时态转换。

### 4.1.3.2 结果分析

我们首先对实体链接进行评测。由于 ReVerb 没有提供主宾语的链接结果，我们从所有关系实例中随机挑选 200 个三元组，并人工标注这些主宾语所链接的实体。我们对比实体链接过程的朴素方法和集成方法，使用准确率（Precision），召回率（Recall）， $F_1$  分值，以及 MRR<sup>[125]</sup> 作为评价指标。MRR 为平均排名倒数（Mean Reciprocal Rank），即统计正确的链接结果在输出列表中的排名，再计算所有三元组上排名倒数值的平均。当一个三元组的主宾语均链接正确时，我们认为该三元组链接正确。实验结果比较如表4-1所示。不同于常规文本的实体链接，由于每个三元组的上下文极少，链接具有一定难度。基于集成的链接方法引入了关系与实体间语义的匹配模型，使主宾语的链接实体互相影响，链接过程的准确率和召回率均得到稳定提升。

表 4-1 ReVerb 三元组的实体链接实验结果。

Table 4-1 Entity linking result.

Linking Strategy	Precision	Recall	$F_1$	MRR
Naive	0.371	0.327	0.348	0.377
Ensemble	0.386	0.340	0.361	0.381

接下来我们衡量二元关系的主宾语搭配结果，主要关注具有较多实例的关系分组。我们首先从包含至少 500 个三元组的关系分组中，随机选择 50 个分组，对于每个分组，我们挑选出支持集合数量最大的 100 个类型对作为评测的对象。我们将这些类型对分配给 3 位对 Freebase 类型有了解的标注者，每个标注者根据自己的理解，判断类型对是否适合于描述对应关系，并标注 0 到 3 的分值。将三位标注者的打分进行平均，即可得到这 50 个关系分组的类型对排序。

我们使用点对点互信息（Pointwise Mutual Information）<sup>[126]</sup> 作为基线模型，该模型在选择偏好任务中被使用，例如文献 [119]。PMI 模型使用以下公式定义一个关系  $r$  与类型对  $tp$  的关联度：

$$PMI(r, tp) = p(r, tp) \log \frac{p(r, tp)}{p(r, *)p(*, tp)}, \quad (4-10)$$

其中  $p(r, tp)$  代表联合概率，即关系分组为  $r$ ，且支持  $tp$  的三元组占所有三元组的比重， $*$  代表任意关系或类型对。

我们使用 MRR 分数进行评测，衡量不同方法生成的最佳关系模式在标注列表中的位置。如表4-2所示，和基线模型进行比较，我们的方法在 MRR 指标上获得了 10.1% 的相对提升。

表 4-2 二元关系模式推理的评测结果。

Table 4-2 End-to-end schema inference results.

Approach	MRR Score
PMI Baseline	0.306
Our Approach	0.337

最后，表4-3列举了一些具体的关系分组，以及我们系统抽取的关系模式。我们可以看出，当构建了 Freebase 的类型层次结构之后，系统能够同时得到粗粒度和细粒度的类型信息，因此最终生成的类型对具有更加丰富的信息量。

表 4-3 生成的二元关系模式举例。

Table 4-3 Real examples of generated relation schemas.

Relation	Top-3 schemas
be found at	$\langle \text{location}, \text{location} \rangle$ $\langle \text{employer}, \text{location} \rangle$ $\langle \text{organization}, \text{location} \rangle$
appear on	$\langle \text{person}, \text{tv program} \rangle$ $\langle \text{person}, \text{nominated\_work} \rangle$ $\langle \text{person}, \text{winning work} \rangle$
be the writer of	$\langle \text{person}, \text{nominated\_work} \rangle$ $\langle \text{person}, \text{film} \rangle$ $\langle \text{person}, \text{book\_subject} \rangle$

## 4.2 关系的结构化语义挖掘

上一节的研究目标是挖掘一个关系所存在的主宾语类型搭配，用于区分不同的语义。本节的研究重点放在了深入理解关系本身，用结构化的符号代替字符形式的描述。我们提出了基于模式图的语义表示方法，与传统路径规则相比，图结构具有的分支可以更好地支持复杂语义，具有良好可解释性的同时，也可被用于知识库补全任务中。

### 4.2.1 概述

以 DBpedia、Freebase 等为代表的开放领域知识库包含了预先定义好的标准化的知识库谓词，用于连接知识库中的实体、类型和概念。知识库中的事实采用三元组形式表示，与关系三元组保持一致。本节中，我们假定每个关系三元组均已完成了实体链接步骤，用  $(e_{subj}, r, e_{obj})$  来表示。那么很显然，事实三元组和关系三元组的区别仅体现在谓语成分上。因此，利用知识库谓词来表示自然语言关系的语义，是一个很自然的想法，若能将开放式信息抽取中的每一个关系实例都映射为知识库中的三元组，那么机器将很容易理解海量非结构化文本中蕴含的结构化信息。这种基于直接对应的思路非常直观，但是对于现有的知识库，例如 Freebase<sup>[9]</sup>，即便其中包含十亿级别的事实三元组，仍然会面临两个主要的挑战。

首先，知识库和自然语言关系之间存在着语义鸿沟。以关系 “has grandfather” 为例，Freebase 中并不存在一个谓词能与之完全匹配，但存在一些和它相关的谓词，例如 *parents* 以及 *gender*。这是因为知识库的构建过程较为严谨，为了避免歧义，每一种谓词的语义都更加单一，同时为了避免信息冗余，能通过其它谓词进行描述的语义，通常不会对应一个单独的谓词。

其次，知识库的构建还远不够完整。即便拥有海量的事实三元组，但依然存在很多长尾的谓词，并没有多少事实与之相关。这个挑战也引入了另一个开放的研究课题，即知识库补全（Knowledge Base Completion）<sup>[25, 26, 127]</sup>。该课题的目标是，给定知识库中的目标谓词，根据其拥有的少量事实三元组进行学习，为其补充新的事实，这些新事实的主语和宾语均为知识库中已存在的实体。换言之，在已有的实体之间连接更多的谓词，使知识库更加稠密。

为了应对以上两个挑战，我们关注的重点在于能否利用知识库中已经存在的谓词，描述一个自然语言关系所具有的语义。已有的相关研究方法主要可以分为两大类。第一类方法为知识库的向量表示学习。这种方法类似于词向量技术，利用知识库中的三元组作为训练数据，学习每个实体以及谓词在连续空间中的特征表示，使得每个三元组的两个实体和谓词表示之间满足特定的代数关系。将开放式信息抽取的关系三元组与知识库已有的事实三元组合并，这类方法可以获取每一个目标关系的隐含语义。但考虑到知识库表示学习中涉及到的参数数量非常庞大，这种方法需要大量的训练数据以应对长尾实体，同时训练的时间开销也不可忽略。已有的研究工作主要集中在了较小的知识库上，例如 FB15K<sup>[29, 128]</sup>。

另一类方法为规则推导，每个目标谓词或关系的语义表达由明确的规则构建而成。这里的规则等价于知识库的子结构，用于连接自然语言关系中的主语和宾语实体。其中最基本的结构为路径的形式，即通过一个或多个谓词组成序列，连接主语和宾语。规则

推导方法的优势在于高度可解释性。一方面，知识库的子结构可以转换为知识库上的查询语言例如 SPARQL，因此可以通过在知识库上运行查询的方式，明确得知特定的两个实体之间是否可能存在某种关系。另一方面，相比知识库向量学习方式，基于规则推导的方法允许使用多条规则描述同一个关系，更好地适应自然语言中的多义性。此外，必要的情况下，人类可以对输出的规则进行微调。

根据以上论述，本节的研究建立在规则推导的基础之上。因此，我们将传统的基于路径的规则进行扩展，而是以树形结构的形式，不仅连接主语和宾语，同时还连接了其余相关实体，用于表示目标关系所具有的隐藏语义限制。这种树形结构是具有相同边结构的知识库中具体子图的抽象表示，我们将其称为模式图（Schema Graph）。图4-2是二元关系“has grandfather”的模式图，通过谓词路径 [ parents, parents ] 表示主宾语之间的祖孙关系，同时利用 *gender* 限制宾语的性别，以此精确描述关系语义。

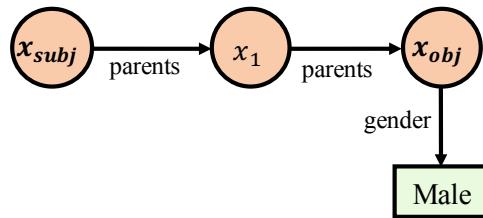


图 4-2 二元关系“has grandfather”的语义表示。

Figure 4-2 An example schema graph.

具体而言，给定自然语言中的关系  $r$  以及抽取出的三元组  $(e_{subj}, r, e_{obj})$ ，本章的研究任务是在知识库中挖掘出一系列与之相关的模式图，并且用概率分布的形式，描述用特定模式图代表该关系语义的可能性。在进行模式图推理的过程中，我们主要会面临以下三个技术性挑战：

首先，候选模式图的数量非常庞大。传统的规则推导中只考虑谓词路径，虽然候选路径的数量随长度呈指数增长，但在知识库中能够连接两个特定实体的路径仅有少数，因此简单遍历可以得到所有的候选路径。然而，具有树形结构的模式图中，不仅存在额外的谓词作为分支，而且包括用于语义限制的实体，任何一个实体的改变，都会产生一个新的模式图。若使用暴力枚举生成模式图，时间复杂度上无法承受，同时还会生成大量偏离语义的模式图。

其次，模式图推理需要做好粒度上的平衡。当一个模式图缺少足够的语义限制，它虽然能匹配已知的三元组，但也可能混淆了错误的三元组。反之，若一个模式图包含了不必要的语义限制，就很可能无法匹配已知的三元组。很显然，太具体或宽泛的模式图都无法精确表示一个关系的语义，但是如何兼顾这两点，并通过概率分布描述不同粒度

候选的语义匹配程度，这成为了模式图推理过程中的另一个难点。

最后，模式图推理模型仅有三元组作为训练数据，不存在标注好的模式图，同时没有明确给出不符合特定关系的错误三元组数据，这给学习过程增添了难度。一种规避方法是使用封闭世界假设（Closed World Assumption），即假定所有未见过的三元组都是错误的。但考虑到知识库本身远不够完整，封闭世界假设会带来大量的错误反例，这并不是一个最好的解决方案。

本章提出的基于模式图的规则推导模型旨在解决应对以上三个挑战，其主要贡献可以分为以下四个部分：

1. 我们定义了自然语言关系的模式图。和传统规则推导模型相比，模式图是谓词路径形式的规则扩展，通过挖掘隐藏的关联实体，在路径之上构建分支，准确描述关系的复杂语义；
2. 我们提出了一种基于局部搜索的启发式方法，通过高效的剪枝策略，快速生成关系所对应候选模式图；
3. 我们提出了一种基于数据驱动的方法，将模式推理问题转化为查询任务进行建模，并在不明确生成负面训练数据的情况下，学习候选模式图之间的概率分布，实现不同粒度模式图的统一比较；
4. 我们对自然语言关系以及知识库中已有的谓词进行了知识库补全任务的测评，包括主宾语预测和三元组分类两个子任务，我们的模型在这两个测评任务上均显著优于已有方法。具体生成的模式图结果表明，我们提出的模型能够挖掘出具体且精确的语义。

## 4.2.2 相关工作

随着大规模结构化知识库的提出与广泛使用，知识库补全任务成为了近年来的热门研究课题。该任务旨在对知识库中已有的谓词进行建模，通过预测潜在的  $(e_1, p, e_2)$  三元组，实现扩充知识库的最终目的。到目前位置，在该课题上的研究方法主要分为两类：基于知识库表示学习和基于规则推导。

知识库表示学习受到词向量技术<sup>[59, 60]</sup>的启发，将知识库中的实体类比为单词，每个实体具有一个向量表示，对应连续语义空间上的一个点。作为连接不同实体的桥梁，知识库中的每个谓词都对应着各自的向量或矩阵表示。通过定义不同的向量或矩阵之间的运算方式，这类方法可以计算每个三元组的置信度，以此实现对实体及谓词的表示学习。

RESCAL 模型<sup>[28]</sup>是一个基础的知识库向量模型，它基于实体向量和谓词矩阵表示的双线性运算。HOLE 模型<sup>[82]</sup>是 RESCAL 模型的改进，使用向量循环平移的技巧计算

实体间的组合语义向量，大幅度降低了谓词的表示维度。在众多知识库表示学习的方法中，有一组方法称为隐距离模型，它们对三元组置信度的计算方式主要基于连续空间中的距离度量：将主宾语向量经过某种方式的映射（翻译）之后，距离越小，置信度越高。最典型的研究工作为 TransE，其核心思路在于尽可能使每个三元组  $(h, r, t)$  对应的向量计算满足  $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ ，即利用谓词向量将连续空间中的主语进行平移，使其尽量与宾语重合。为了能更好地表示多对多的关系，相关文献 [30, 31] 对 TransE 模型进行了改良。Wang 等人提出了 TEKE 模型<sup>[129]</sup>，它对已有的翻译模型进行改良，充分利用结构化文本的知识，寻找三元组中单词级别的共现，并利用共现上下文微调实体和谓词的向量表示。

基于规则推导的方法旨在用逻辑规则的形式表达谓词的语义。例如  $\text{parent}(x, y) \wedge \text{parent}(y, z) \rightarrow \text{grandparent}(x, z)$  是一个常识性的规则，我们可以通过规则的左侧部分，在知识库中寻找出更多的祖孙间的关系。Jiang 等人的工作 [23] 基于马尔科夫逻辑，通过挖掘的规则对自动构建的知识库进行信息过滤。其它一些方法使用概率软逻辑或关联规则挖掘完成类似的任务<sup>[130, 131]</sup>。Galárraga 等人提出的 AMIE<sup>[22]</sup> 以及 AMIE+<sup>[132]</sup> 系统则直接根据知识库的三元组寻找置信度较高的一阶逻辑规则。最新的一些研究着眼于在知识库中寻找路径形式的规则，通过挖掘大量可能的路径，作为表示语义的特征。Lao 等人提出了 PRA 模型<sup>[25]</sup>，通过在谓词路径上的随机游走策略，衡量其连接一对实体的好坏程度，目标关系的语义等同于不同路径特征的带权组合。Gardner 等人对 PRA 模型进行改进，提出了 SFE 模型<sup>[26]</sup>，除了捕捉连接主宾语的路径以外，还从主宾语各自的知识库子图中挖掘独立的特征，同时谓词路径的定义更加宽泛，允许在其中使用通配符表示任意谓词。此外，Wang 等人提出了 CPRA 模型<sup>[79]</sup>，这是对 PRA 模型的另一种改进，通过挖掘目标关系中的相关性，使得相似关系之间的路径挖掘结果可以互相影响。然而，通过开放式信息抽取获得的三元组数量相对有限，不同的关系之间几乎不存在重叠的实体对，在这种场景下，CPRA 模型效果等价于原始的 PRA 模型。

一些相关的研究尝试在知识库向量学习的基础之上加入一定的逻辑规则。Guo 等人提出了 KALE 模型<sup>[133]</sup>，其主要思想是将规则转换为多个三元组之间的与或非逻辑操作，因此基于翻译模型计算的三元组置信度得以在逻辑规则级别产生交互。TRESCAL 模型<sup>[134]</sup> 在经典的 RESCAL 模型中加入了知识库的类型限制。而 Wang 等人的工作<sup>[135]</sup> 使用整数线性规划技术，将知识库向量表示和规则挖掘进行统一，

狭义的知识库补全任务只考虑知识库中的谓词，我们的工作将知识库补全的场景进行了扩展。考虑到为了降低知识库结构与自然语言描述的差距，知识库补全任务也可以针对自然语言中的二元关系。开放式信息抽取与这样的任务相契合，既提供了全新谓词，又有一定量的三元组用于补全学习。一些已有的工作也关注了自然语言关系到知识库的映射。Zou 等人的工作<sup>[136]</sup> 使用了非监督学习的方式，利用 TF-IDF 特征寻找关系

到谓词路径的匹配。Zhang 等人的工作<sup>[24]</sup> 利用马尔科夫逻辑网络<sup>[137]</sup>，学习自然语言关系对应于不同候选谓词路径的概率。这些方法对关系的表示局限于路径的形式，无法准确地描述一个形式简单但具有组合语义的关系。我们的工作旨在理解具有复杂语义的关系，挖掘其包含的隐含限制条件，并通过具有“路径 + 分支”结构的模式图进行语义建模。

### 4.2.3 任务定义

在本章中，我们定义知识库为  $KB = \{E, L, P\}$  三部分组成，具体如下： $E$  为知识库  $KB$  中所有实体集合； $L$  为  $KB$  中所有不同谓词的集合； $P$  为  $KB$  中所有事实三元组集合，每一个三元组表示为  $p(e_1, e_2)$ ，其中  $e_1, e_2 \in E$ ，并且  $p \in L$ 。此外，知识库中存在用于描述一个实体所拥有类型的谓词  $IsA$ ，为了简化描述，本章中我们将不同类型也看做实体，同属于集合  $E$  中。

一个模式图  $S$  同样由三部分构成， $S = \{E_S, X, P_S\}$ ，具体如下： $E_S \subseteq E$ ，为模式图中出现的具体的实体集合； $X$  为实体变量的集合，每一个变量  $x \in X$  在模式图中等同于占位符，为特定实体  $e \in E$  的抽象；模式图中包含两个特殊变量，即  $x_{subj}, x_{obj} \in X$ ，分别代表目标关系的主语和宾语实体； $P_S$  为模式图中的抽象三元组集合，每一个抽象三元组为  $p_s(v_1, v_2)$ ，其中  $v_1 \in X$ ， $v_2 \in E_S \cup X$  以及  $p_s \in L$ 。此外，模式图  $S$  具有以下性质：

- $S$  的表现形式为有向树形结构，且根节点一定为主语的实体变量  $x_{subj}$ ；
- 连接主语变量  $x_{subj}$  和宾语变量  $x_{obj}$  的谓词路径，称为模式图  $S$  的骨架；
- 骨架之外的所有抽象三元组称为模式图的限制（或分支）；
- 一个仅具有骨架而不包含任何限制的模式图，称为简单模式图，等价于谓词路径。

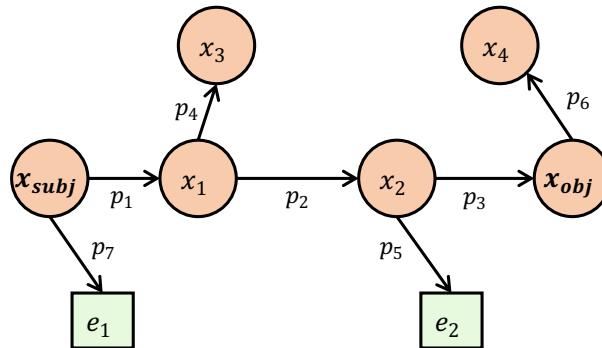


图 4-3 模式图的一般形式。

Figure 4-3 A general style of a schema graph.

图4-3显示了模式图的一般形式。可以发现，其中的每一条边都至少连接了一个实

体变量。模式图代表着知识库中，满足相同特定结构的一系列具体子图。这些具体子图称为实例图（Grounded Garph），作为模式图的实例化形式，所有的实体变量  $x_i$  被替换为特定的实体  $e_i \in E$ ，且每一个抽象三元组  $p_s(v_1, v_2)$  在实例化之后均对应存在于知识库中的事实  $p(e_1, e_2) \in P$ 。例如图4-2中的模式图，其不同的实例图囊括了知识库中所有已知的（个人，双亲，双亲父亲）知识。对于实例图中的主宾语对  $(e_{subj}, e_{obj})$ ，我们称其为模式图的一个支持实例。

根据以上符号定义，给定知识库  $KB$ ，自然语言关系  $r$  以及多个关系三元组  $\{(e_{subj}, r, e_{obj})\}$ ，我们对关系的深度语义挖掘任务为，推导出一系列描述其语义的候选模式图，并学习模式图上的概率分布，以此表示自然语言关系所具有的多义性。

#### 4.2.4 我们的方法

本节主要介绍将自然语言关系映射为模式图的具体方式。给定关系  $r$  以及其一系列关系实例作为训练数据，我们首先依据给定的主宾语对  $(e_{subj}, e_{obj})$ ，从它们支持的所有模式图中寻找可能性较高的候选模式图，然后对具有不同粒度的模式图进行重要性衡量。由于没有直接的  $<$  关系，模式图  $>$  对作为训练数据，我们提出了一种基于远距离监督学习的方式，学习所有候选图上的概率分布。

##### 4.2.4.1 候选模式图生成

根据已有的关系实例，我们提出了一种高效的搜索算法，在知识库上挖掘可能表示关系语义的候选模式图。其基本思路在于，首先通过主宾语对寻找仅由骨架（谓词路径）构成的简单模式图，带有限制的模式图生成则以简单模式图为起点，不断寻找与关系三元组契合的限制，并通过递归的形式将新的限制连接到已有的候选上，一步步生成具有复杂结构的模式图。

简单模式图的生成基于实体对在知识库中的直接连接。我们使用双向广度优先搜索，为每个实体对提取由主语连接到宾语的所有谓词路径。考虑到一个自然语言关系通常由短语构成，通常不会具有太多的语义跳跃，因此我们对谓词路径长度进行限制，避免生成大量无意义的路径。基于前人的工作 [24]，我们限制谓词路径最长不超过 3。此外，为了尽可能保证每一个候选图的质量，我们需要排除那些仅由偶然数据生成，实则偏离语义的候选图。一个有效的识别方式利用了候选图的支持率，即支持候选图的实体对占目标关系所有已知实体对的比例，记做  $sup(S)$ 。我们在生成过程中指定支持率阈值  $\gamma$ ，并移除那些支持率  $sup(S)$  小于  $\gamma$  的模式图。综上，对谓词路径和支持率的限制，可以使候选生成步骤过滤大量的干扰模式图。

在生成仅包含骨架的简单模式图之后，我们采用深度优先搜索的方式获取更多更加

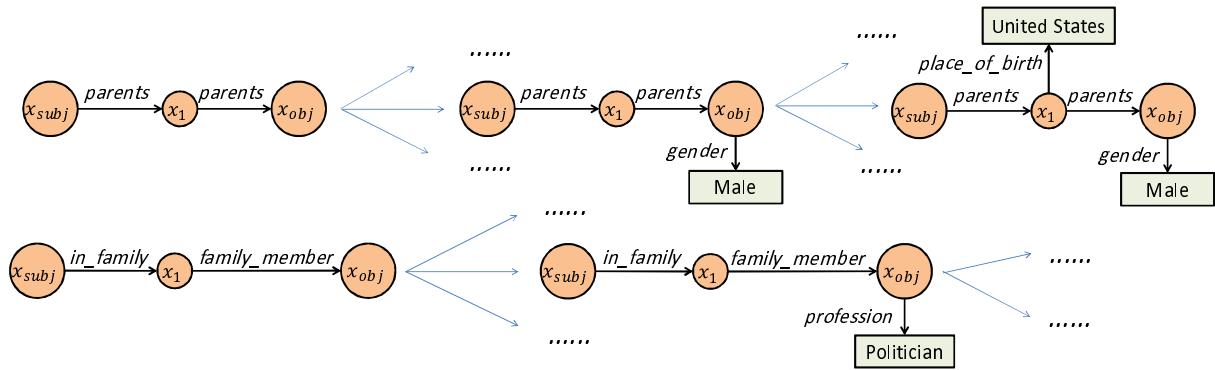


图 4-4 “has father” 模式图挖掘示例。

Figure 4-4 Candidate generation example of relation “has father”.

具体的模式图。如图4-4所示，“has grandfather”关系可以生成多种不同的简单模式图，在此基础上，我们逐步添加表示复杂语义的分支，让模式图更加具体。这个步骤的挑战在于，即便骨架长度得到限制，模式图扩展的搜索空间仍然异常庞大。为了提高效率，我们使用优先队列维护搜索过程中获取的高质量模式图，并进行剪枝操作，压缩候选图的搜索空间。具体步骤的伪代码流程如算法4-1所示。 $Q$  为存放模式图的优先队列，初始化为空，最大容量为  $B$ ，搜索过程中始终维护具有最大支持率的前  $B$  个候选图（第8行）。使用支持率作为剪枝依据的原因有二：一方面如同骨架生成中的论述，支持率高的模式图更不容易偏离语义，而支持率过低的候选图更有可能引入了不必要的限制，导致无法匹配大量已知三元组；另一方面，随着候选图上添加的限制越多，支持率一定呈非严格单调递减趋势，因此这种单调性特征可以直接用于剪枝。函数  $SchemaExpansion$  以模式图  $S$  为输入，返回值为一个模式图集合，其中每个模式图均为在  $S$  上加入一条新的限制所形成的更复杂的候选，例如图4-4中的  $(x_{obj}, gender, Male)$ ,  $(x_{obj}, profession, Politician)$  等。

为了使候选模式图之间具有多样性，我们期望最终保留的  $B$  个候选图中能包含多种不同的骨架，因为不同骨架的模式图通常代表更大的语义差别。因此在实际的搜索过程中，我们根据不同骨架的支持率，将整个大小为  $B$  的优先队列按比例分为多块，每个骨架上的深度搜索将使用各自独立的优先队列。这样的做法可以提高并行工作效率，同时保证候选集合不被某个高支持率的骨架主导。

#### 4.2.4.2 模式图概率推理

当关系  $r$  的候选图生成完成之后，下一步需要从中推理出最具有代表性的那些模式图。我们的目标是将关系的表示多义性表示为每个候选模式图  $S$  的条件概率  $P(S|r)$ ，这

**算法 4-1 复杂模式图搜索**

**Input:** Schema  $S$ , priority queue  $Q$ , budget  $B$ , minimum support ratio  $\gamma$

**Output:** Priority queue  $Q$  after expanding on  $S$

```

1: procedure SEARCH( $S, Q, B, \gamma$ )
2:   if  $sup(S) < \gamma$  then
3:     return  $Q$ 
4:   end if
5:   if  $Q.size < B$  or  $sup(S) > sup(Q.top)$  then
6:      $Q.push(S)$ 
7:     while  $Q.size > B$  do
8:        $Q.pop()$ 
9:     end while
10:     $NewList \leftarrow SchemaExpansion(S)$ 
11:    for  $S'$  in  $NewList$  do
12:       $Q \leftarrow Search(S', Q, B, \gamma)$ 
13:    end for
14:   end if
15:   return  $Q$ 
16: end procedure

```

样不同粒度的模式图之间可以直接比较。由于没有直接的  $<$  关系，模式图  $>$  训练数据，我们对概率分布的学习方式依靠三元组数据作为驱动，将学习过程建模为知识库查询场景上的一个最优化问题：给定  $r$  的一个关系实例中的主语（或宾语）实体，寻找最为合适的模式图概率分布，使得依照此分布在给定实体周围进行知识库查询时，能尽可能返回对应的宾语（或主语）实体。

为了能够在不同粒度的候选模式图之间得到平衡，我们使用最大化似然估计的方式定义目标函数，寻找最优的模式图概率分布，使得查询过程返回正确实体的概率最高。似然函数定义如下：

$$L(\vec{\theta}) = \prod_i P(obj_i|subj_i, \vec{\theta})P(subj_i|obj_i, \vec{\theta}), \quad (4-11)$$

其中，向量  $\vec{\theta}$  表示候选模式图的概率分布，即  $\theta_j$  对应条件概率  $P(S_j|r)$ ，且满足  $\sum_j \theta_j = 1$ 。 $subj_i, obj_i$  分别表示关系  $r$  的第  $i$  个实例中的主语和宾语。

接下来，我们通过两阶段的生成过程，对概率  $P(obj|subj, \vec{\theta})$  进行建模：首先根据模式图上的多项分布，随机挑选出一个模式图  $S \sim Multinomial(\vec{\theta})$ ，然后对模式图  $S$

进行查询（即在知识库上进行实例化），在所有主语为  $subj$  的实例图中，随机挑选其中的一个实例图，将其宾语实体返回。第一个阶段中，模式图的选取与主语  $subj$  条件独立，第二个阶段由于固定了模式图，因而与  $\vec{\theta}$  也条件独立。考虑这些条件独立之后， $P(obj|subj, \vec{\theta})$  的生成过程定义如下：

$$\begin{aligned} P(obj|subj, \vec{\theta}) &= \sum_j P(S_j|subj, \vec{\theta})P(obj|subj, S_j, \vec{\theta}) \\ &= \sum_j \theta_j P(obj|subj, S_j), \end{aligned} \quad (4-12)$$

概率  $P(obj|subj, S_j)$  的值对应模式图  $S_j$  在知识库上的查询结果：令  $q(subj, S_j)$  代表模式图  $S_j$  的实例图中，所有主语实体为  $subj$  的对应宾语集合，以均匀分布从中挑选一个实体  $obj$ ，公式展开如下：

$$P(obj|subj, S_j) = \begin{cases} 1 / |q(subj, S_j)| & obj \in q(subj, S_j) \\ \alpha & \text{otherwise} \end{cases} \quad (4-13)$$

公式中的  $\alpha$  为平滑参数，在目标宾语无法通过  $S_j$  得到时，我们将概率定位很小的数值，防止整个似然函数值变为 0。观察可知，对于过于宽泛的模式图  $S_j$ ， $q(subj, S_j)$  集合数量很大，从中随机选择到目标宾语的概率会因此降低；而对于过于具体的模式图，会使得较多的实体对无法被支持，因此同样会对似然带来降低。由此可见，基于两阶段生成的概率建模方式，可以实现宽泛与具体模式图之间的平衡，找到最适合的语义结构。此外， $P(subj|obj, \vec{\theta})$  的定义为公式4-12的对称版，代表着给定宾语实体，查询得到目标主语的概率。

综上，我们将模式图推理问题转化为了基于最大似然估计的最优化任务，并利用梯度下降算法对模型参数  $\vec{\theta}$  进行更新，使目标函数  $L(\vec{\theta})$  值最大。具体使用的梯度下降算法为 RMSProp<sup>[138]</sup>。

## 4.2.5 实验

本节中，我们首先对推理出的模式图进行直接的质量测评，然后使用主宾语预测和三元组分类这两个任务定量评估模式图的语义表达能力，最后我们分析一些错误例子，讨论当前模型的不足之处。

### 4.2.5.1 实验设置

**知识库：**为了和已有的知识库向量表示方法进行公平比较，我们在实验中使用了两个 Freebase 的子集：**FB3m** 以及 **FB15k**。FB15k 由 Bordes 等人提出<sup>[29]</sup>，它包含了 14,951 个实体，1345 种不同谓词，以及 483,142 个事实三元组。FB15k 的三元组被分为了训练

集、验证集、测试集三部分，我们仅选用训练集部分作为使用的知识库。与此同时，我们从 Freebase2015 年 6 月的版本抽取出最主要的 3,000,000 个不同的实体，并提取这些实体之间的联系，构成 FB3m 子集。FB3m 包含大约 50,000,000 个三元组，是 FB15k 的 100 倍。和完整的 Freebase 相比，FB3m 更加轻量化，但依然包含了大量有价值的信息。

**关系数据集：**我们使用了三个不同的关系数据集进行知识库补全的相关实验。在自然语言场景中，目标关系来源于开放式信息抽取系统 PATTY<sup>[4]</sup>，包含了大约 200,000 种不同的自然语言关系，以及百万级别以上的三元组。由于 PATTY 使用维基百科作为语料库，三元组中的所有实体均为维基百科页面，因此每个实体均自动链接至 Freebase。我们从 PATTY 中抽取子集 “PATTY-100” 以及 “PATTY<sup>+</sup>-100” 用于实验，PATTY-100 数据集与 FB15k 相匹配，其包含了 100 个具有较多数量三元组的关系，且三元组中所有实体均存在于 FB15k 中，平均每个关系包含 180 个关系实例。相对应地，PATTY<sup>+</sup>-100 与 FB3m 相匹配，同样包含 100 个自然语言关系，平均每个关系包含 388 个实例。两个数据集中，每一个关系的三元组均被分为训练集、验证集、测试集 (64% : 16% : 20%)。第三个关系数据集属于知识库场景，我们从 FB15k 的 “people”、“location” 以及 “sports” 三个领域内挑选出 37 个热门谓词，并将它们的所有三元组抽取出，组合为数据集 “FB15k-37”。每一个三元组出现在训练集、验证集、测试集的位置与 FB15k 保持一致。FB15k-37 是 FB122<sup>[133]</sup> 的一个子集，保证其中每一个关系在测试集中都具有至少 10 个三元组。

**用于比较的已有方法：**对于知识库向量表示的方法，我们与 TransE<sup>[29]</sup>，KALE<sup>[133]</sup>，TEKE<sup>[129]</sup> 以及 HOLE<sup>[82]</sup> 进行比较。对于规则推导的方法，我们与 SFE<sup>[26]</sup> 以及 AMIE+<sup>[132]</sup> 这两个系统进行比较。我们考虑使用 CPRA 模型<sup>[79]</sup> 作为另一个比较方法。但在 PATTY 相关的数据集中，不同关系之间几乎不存在相同的实体对，因此 CPRA 模型将会退化为传统的 PRA 模型<sup>[25]</sup>，被更优秀的 SFE 严格取代。这些模型在 2.2 节或 4.2.2 节中已有论述。

**模型实现细节：**我们评估了模型的两个变种，分别为生成带限制的模式图的 Ours-SC，以及仅生成简单模式图的 Ours-SK。以下是具体调参细节：

- 候选模式图的数量，即优先队列容量  $B$  设为 5000；
- 模式图骨架长度限制  $\tau$  设为 3，我们的方法可以支持更长的骨架，但具体测试中无明显的效果提升，同时候选生成时间显著增长，这里不展开讨论；
- 支持率阈值  $\gamma$  调参范围为 {5%, 10%, 15%, 20%}；
- 平滑参数  $\alpha$  调参范围为 {1e-6, 1e-5, 1e-4}；
- 学习率  $\eta$  调参范围为 {0.02, 0.05, 0.1}。

用于比较的系统中，具有开源代码的方法包括 AMIE+<sup>1</sup>，SFE<sup>2</sup> 以及 HOLE<sup>3</sup>。KALE 的代码由作者提供，TransE 基于 HOLE 的代码运行，并且我们在 TransE 的基础上自行实现了 TEKE 模型。以上基于知识库向量表示的模型均使用最大间隔损失进行训练，对于 KALE 模型，学习率调参范围为 {0.02, 0.05, 0.1}，最大间隔参数范围为 {0.1, 0.12, 0.15, 0.2}；对于 TransE，TEKE 以及 HOLE，学习率调参范围为 {0.05, 0.1, 0.2}，最大间隔参数范围为 {0.5, 1.0, 1.5, 2.0, 2.5}。

#### 4.2.5.2 模式图质量测评

这一部分的实验中，我们主要关注具有明确结构的模式图是否可以弥补 Freebase 和 PATTY<sup>+</sup>-100 之间的语义差距。我们首先通过具体的例子观察不同的规则推导方法，即 Ours-SC，Ours-SK，AMIE+ 以及 SFE 所生成的代表性结构。我们从 PATTY<sup>+</sup>-100 数据集中挑选出四个具有一定复杂性的关系，并在较大结构的 FB3m 上学习各自的规则。对于 Ours-SC 和 Ours-SK，我们使用选择概率最高的模式图作为代表性结构。SFE 模型中，每个规则（谓词路径）都对应一个特征，我们选择特征权重最高的规则作为代表性结构。AMIE+ 依靠准确率对规则进行排序，因此我们挑选准确率最高的规则，若多个规则准确率相同，我们则从中手动选择最合适的规则。

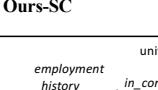
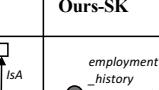
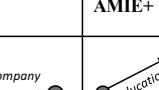
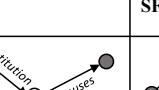
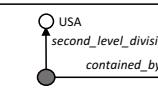
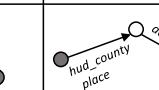
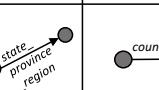
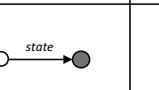
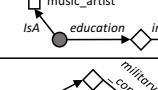
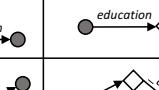
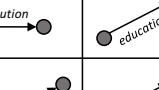
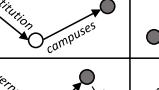
Natural Language Relation	Ours-SC	Ours-SK	AMIE+	SFE
(PER) taught at (LOC)				
(LOC) is a county located in (LOC)				
(PER) studied music at (LOC)				
(PER) 's invasion in (LOC)				

图 4-5 不同的规则推导系统对四个复杂关系生成的代表性结构。

Figure 4-5 Top structures produced by four systems on 4 complex relations.

图4-5列出了四个自然语言关系，以及不同系统生成的最佳结构。其中，圆点表示实体或变量，左右两个黑色圆点分别代表  $x_{subj}$  和  $x_{obj}$ 。方块代表知识库中的类型，菱形则代表用于维护多元关系的辅助节点。从这些例子中可以发现，Ours-SC 的模式图所具

<sup>1</sup><https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/amie/>

<sup>2</sup><https://github.com/matt-gardner/pr>

<sup>3</sup><https://github.com/mnick/scikit-kge>

有的分支结构，可以带来更加精确的语义。对比仅生成骨架的 Ours-SK，带有限制的查询图在每个例子上都表达了几乎完全正确的语义。另一方面，AMIE+ 和 SFE 输出的最佳结构不尽如人意。AMIE+ 按照准确率对规则排序，因此总是倾向于更具体的规则，但牺牲了召回率。同时随着规则长度提升至 4 甚至更高，AMIE+ 系统消耗了大量内存，无法返回任何结果。SFE 生成的规则中包含 [Any-Rel] 代表任意谓词，因此可以生成更多灵活的路径作为特征，但显然其中的大部分都不具有清晰的语义，人类难以直接理解。

作为补充实验，我们对 Ours-SC 和 Ours-SK 生成的模式图进行了人工测评。对每一个自然语言关系，我们从中抽取出至多前 5 个概率值至少为 0.05 的模式图，并由三位标注者进行人工打分，分值选择范围为 {0, 0.5, 1}，分别代表“不相关模式图”（骨架层次已出现语义偏离），“部分匹配”（骨架语义正确，但其余限制需要改善）以及“完全匹配”（骨架和限制的语义均无明显偏差）。我们将三位标注者的打分进行平均，得到每一个模式图的标注分值，并计算排名前  $n$  的所有模式图的平均分值，记做 AvgSc@ $n$ 。三位标注者之间的 Kappa 系数为 0.541，具有稳定的相关性。表4-4列出了不同的 AvgSc@ $n$  分值，Ours-SC 在骨架的基础上挖掘额外的语义限制，将结果提高了约 13%。

表 4-4 模式图列表的 AvgSc@ $n$  测评结果。

Table 4-4 AvgSc@ $n$  results on top-ranked schemas.

	n=1	n=3	n=5
Ours-SK	0.44	0.37	0.34
Ours-SC	<b>0.47</b>	<b>0.40</b>	<b>0.38</b>

#### 4.2.5.3 主宾语预测任务测评

主宾语预测任务的目标是预测三元组  $(e_{subj}, r, ?)$  或  $(?, r, e_{obj})$  所缺失的宾语或主语。测试集中的每一个三元组都对应两个这样的预测任务。公式4-12代表着给定一端实体，生成另一端未知实体的概率，因此对每一个带有未知实体的待预测三元组，我们根据该公式计算生成不同实体的概率，并衡量答案实体的概率排名高低。我们在实验中使用了两个评价指标，分别为 MRR 和 Hits@ $n$ ，前者衡量答案实体在所有预测任务中的平均排名，后者关注在多少比例的预测任务中，答案实体的概率排在前  $n$  位。不同的实验方法通过验证集的 MRR 分值进行独立调参。

以上对排名高低的衡量暗含着一个假设：除了答案实体之外，其余实体均为错误实体。然而考虑到关系可能具有的一对多性质，对于一个待预测的三元组，除了答案实体之外，还可能存在其它实体与给定的已知实体匹配，严格来讲，这些实体虽然不同于唯

一的答案，但也不应该算作错误。因此，我们使用和 TransE<sup>[29]</sup> 相同的设定，在测评中引入两种不同的模式，分别为原始模式和过滤模式：在过滤模式中，计算每个预测的答案实体排名时，均忽略不同于答案的其余正确实体，因此过滤模式下，排名值可能会提高；而原始模式则不做任何的过滤。

我们使用 FB15k 作为知识库进行实验，并与其它模型进行比较。在接下来的实验中，为了方便比较，我们的模型同一参数  $\gamma = 10\%$ ,  $\alpha = 1e-4$ , 以及  $\eta = 0.1$ , 对应着 PATTY-100 验证集上，在过滤模式下的最高 MRR 结果。表4-5和表4-6 分别展示了在 PATTY-100 和 FB15k-37 数据集上的实验结果。在两个数据集上，SFE 模型的代码均碰到了内存问题，因此表格中没有列出对应的结果。对于 PATTY-100 中的关系，我们基于模式图的语义表示方法，其效果优于其它用于比较的规则推导与知识库向量表示模型，以及仅生成简单模式图的变种。在 FB15k-37 数据集上，Ours-SC 与 Ours-SK 的结果十分接近，这主要是因为知识库上的一部分谓词具有等价形式，例如 *location.location.containedby* 和 *location.location.contains* 互为相反关系，对于这些关系，只需要依靠骨架结构就可以精确描述语义。对比两张表格可以发现，对于所有不同的模型和实验模式，自然语言关系上的主宾语预测结果都低于对应的知识库谓词上的结果。主要原因有两点：1) FB15k-37 上的每一个谓词平均包含接近千级别的训练三元组，而 PATTY-100 中的每个关系平均只有 115 个训练数据；2) 自然语言关系具有更多歧义，开放式信息抽取的结果会包含多种语义，而且还要考虑抽取错误的情况，相比之下，知识库上的谓词及三元组的制定经过了部分人工干预，因此歧义更少。

表 4-5 在 PATTY-100 上进行主宾语预测的测评结果。

Table 4-5 Link prediction results on PATTY-100 relations.

	Raw			Filtered		
	MRR	H@3	H@10	MRR	H@3	H@10
TransE	0.112	12.4	27.1	0.129	14.5	29.9
KALE	0.112	12.5	25.4	0.125	14.4	27.5
TEKE	0.101	10.9	24.1	0.114	12.6	26.3
HOLE	0.109	10.5	23.3	0.121	12.3	25.8
AMIE+	0.148	16.5	29.3	0.174	19.5	<b>31.9</b>
Ours-SK	0.169	18.2	29.3	0.179	19.1	30.4
Ours-SC	<b>0.172</b>	<b>18.5</b>	<b>29.8</b>	<b>0.185</b>	<b>19.9</b>	31.5

表 4-6 在 FB15k-37 上进行主宾语预测任务的测评结果。

Table 4-6 Link prediction results on FB15k-37 relations.

	Raw			Filtered		
	MRR	H@3	H@10	MRR	H@3	H@10
TransE	0.310	39.3	53.2	0.394	52.5	65.0
KALE	0.342	40.6	53.0	0.410	48.7	60.6
TEKE	0.288	35.7	49.2	0.339	43.0	56.5
HOLE	0.234	26.7	39.5	0.323	36.5	50.5
AMIE+	0.395	46.1	53.7	0.562	60.0	68.9
Ours-SK	0.425	47.8	55.6	0.664	68.8	73.0
Ours-SC	<b>0.427</b>	<b>48.1</b>	<b>55.7</b>	<b>0.671</b>	<b>69.3</b>	<b>73.3</b>

#### 4.2.5.4 三元组分类任务测评

三元组分类任务的目标是预测一个未知三元组  $(e_1, r, e_2)$  是否描述了一个正确的客观事实。考虑到这是个二分类任务，测试数据中需要包含负样本三元组，因此我们使用和 KALE<sup>[133]</sup> 相同的生成策略，对测试集和验证集中的每个三元组生成 10 个不同的负样本，其中 5 个三元组替换了主语，另外 5 个替换了宾语。为了保证负样本不至于显得过于错误，我们保证用于替换的主语（或宾语）都曾出现在目标关系的某个已知三元组的同样位置上。

对于每一个目标关系，我们通过公式4-11计算各个未知三元组的似然值，以此作为置信度对所有测试集的所有正负样本进行排序。我们使用 FB15k 作为知识库进行了实验，并使用 MAP (Mean Average Precision) 作为测评指标，衡量不同的模型在三元组分类任务上的效果。表4-7列出了 PATTY-100 和 FB15k-37 数据集上的效果，我们的模型在两个数据集上均大幅度优于其它方法。此外我们发现，仅生成简单模式图的方法效果要优于生成完整模式图的做法。我们对实验数据进行了分析，造成这个现象的原因源于负样本生成方式的天然缺陷。例如对于“father of”关系，我们期望负样本中能包含表示母子关系的实例，识别这种负样本需要较高难度，必须依靠额外限制才能和正样本进行区分。然而，负样本的生成方式决定了主语只能替换为某个随机小孩的父亲，判断三元组正确与否主要依靠骨架的正确性，因而很难体现模式图的额外限制为给语义理解带来的优势，减少候选模式图的数量和复杂度反而能得到更好的效果。

表 4-7 三元组分类任务的 MAP 测评结果。

Table 4-7 MAP results on triple classification task.

	PATTY-100	FB15k-37
TransE	0.304	0.666
KALE	0.309	0.654
TEKE	0.282	0.631
HOLE	0.308	0.680
SFE	0.329	0.621
AMIE+	0.226	0.730
Ours-SK	<b>0.408</b>	<b>0.804</b>
Ours-SC	0.403	0.803

#### 4.2.5.5 错误分析

对于一些自然语言关系，我们的模型可能难以寻找出较为正确的模式图。我们对结果进行了分析，并总结出以下几类主要错误。

1. 开放式信息抽取提供的关系三元组存在错误。考虑到 PATTY 主要利用依存语法分析对句子进行关系识别，语法分析本身的偏差将导致生成错误的三元组。例如对于关系 “*served as*”，给定句子 “*Dennison served as the 24th Governor of Ohio and as U.S. Postmaster General ...*”，PATTY 提取的实体对 (William Dennison Jr., Ohio) 有误，正确的宾语应为 “*Governor of Ohio*”。

2. PATTY 数据集中，每个关系实际代表着一个关系同义集，即由多个具有相似结构的关系组成的组合，这导致部分关系同义集混入了语法相似但语义不同的关系，产生本不存在的歧义。以 PATTY 中的关系同义集 “*'s wife*” 为例，其中混入了少部分可能由 “*the wife of*” 产生的三元组，其中主语为妻子，宾语反而为丈夫。在混入的三元组干扰下，模型会误以为该关系的准确语义为不带有性别限制的配偶关系，因此正确的模式图很难获得较高的概率。

3. 对于部分关系，知识库本身缺乏用于描述其语义的谓词。对于一些琐碎的自然语言关系例如 “*talk to*”，知识库显然不包含这类事实。但即便对于一些不那么琐碎的关系，知识库依然可能缺乏必要的谓词。例如关系 “*(singer) performed in (LOC)*” 描述的是歌手和演唱会举办地的联系，但 Freebase 中并不包含类似于 *place\_visited* 或 *hold\_concerts\_in* 的谓词，因此难以通过已有知识表示目标关系的语义。

4. 由于搜索空间的限制，部分有意义的模式图无法在候选生成步骤被过滤。例如

关系“(actor) starring with (actor)”，由于 Freebase 通过辅助节点(Mediator)维护多元关系，这使得最合适的骨架<sup>1</sup>长度为 4，并不满足候选生成的骨架长度限制，因此模型无法得到这样的模式图。

### 4.3 本章小结

本章的研究着眼于自然语言中的二元关系，根据关系已有的三元组实例，推理出其所具有的语义。第一部分的工作将关系模式定义为知识库中的主宾语类型搭配，并利用知识库的类型层次结构实现模式推理。我们提出的方法基于一个直观的思路，即尽可能使用具体的模式匹配更多的已知实例。在 ReVerb 上进行的人工测评实验表明，此方法推理出的最具有代表性的模式具有较高的准确度，效果优于传统的选择偏好模型。

第二部分的工作直接挖掘关系语义和结构化知识之间的匹配。为了使语义理解具有良好的可解释性，我们提出了基于模式图的规则推导模型，模式图是对传统路径规则的泛化，以“路径 + 分支”的结构描述具有更多限制的复杂语义。该模型将关系语义表示为多个模式图的概率分布，以适应关系的多义性。我们对 PATTY 中的热门关系进行模式图推理，多个具体例子表明，基于模式图的结构表示有能力描述更加细化的关系语义，而且质量优于其它已有的规则推导模型。此外，基于模式图的语义表示还可用于知识库补全任务中，在主宾语预测和三元组分类两个子任务上，效果优于其它规则推导及知识库向量模型。

后续的研究主要包括两部分：数据预处理方面，关系三元组的实体链接需要优化，主语和宾语都可能存在不可链接实体，需要进行识别从而过滤杂乱三元组；语义理解模型方面，本章的两个工作均基于数据驱动，对于已知三元组较少的长尾关系，模型效果会明显降低，如何利用关系本身的短语信息作为额外特征进行推理，是值得研究的方向。

本章中，关系主宾语类型搭配挖掘的研究成果已发表于 2015 年国际会议 Empirical Methods in Natural Language Processing (EMNLP 2015)，论文题目为 “Inferring Binary Relation Schemas for Open Information Extraction”；关系结构化语义挖掘的研究成果已发表于 2017 年国际会议 International Joint Conference on Artificial Intelligence (IJCAI 2017)，论文题目为 “A Data-Driven Approach to Infer Knowledge Base Representation for Natural Language Relations”。

---

<sup>1</sup>骨架具有  $actor \rightarrow med. \rightarrow film \rightarrow med. \rightarrow actor$  形式，其中  $med.$  为辅助节点。



## 第五章 面向复杂语义的知识库自动问答研究

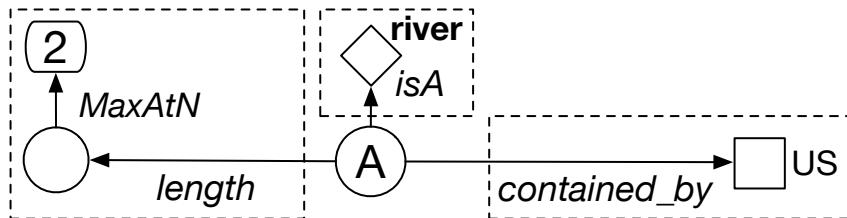
本章的研究为基于知识库的自动问答任务。用户提出的问句可能具有复杂语义，其中包含了未知答案与相关实体的多种关系，因此复杂问句的回答过程充满了挑战。我们提出了面向复杂语义的知识库问答模型，主要特点在于，我们利用神经网络学习复杂语义结构的整体连续特征表示，从而捕捉不同语义成分之间的信息交互。

### 5.1 概述

基于知识库的自动问答（KBQA）是自然语言处理中的经典应用场景。该任务以自然语言问句作为输入，并根据已有结构化知识库提供的信息，寻找到问句的一个或多个答案。以 Freebase, YAGO, DBpedia 为代表的结构化知识库主要以维基百科为骨架构建而成，它们包含真实世界的广域知识，因此常用于自动问答任务中。

在自动问答任务中，我们关注的问题称为“事实类问题”，其特点在于它们询问的是与句子中实体相关的客观事实，因此答案为知识库中存在的实体、数值或时间。以一个较简单的问题为例，“What's the capital of the United States?”，为了准确回答这个问题，一个较为直接的方式是，首先识别句子中的相关实体并链接到知识库，再将该实体与目标答案之间的自然语言关系映射为知识库中的一个谓词（或为词序列），那么原问题即可转换为具有（实体，谓词，目标答案）三元组形式的查询语句，例如 (*united\_states*, *capital*, ?)，通过在知识库上运行查询语句，生成最终的结果。将已有的<问题，答案>对作为训练数据，我们可以通过远距离监督（Distant Supervision）的形式学习问句和查询语句之间的映射关系。

对于只包含简单语义的问题，我们可以通过上述方法将其转为知识库上的一个基本三元组查询，但这样的方法并不适用于其它具有更复杂语义的问题。例如图5-1所示，为了准确回答问题“What is the second longest river in United States?”，我们实际上需要对其进行推理，得出以下三条语义线索：1) 答案实体位于美国内部；2) 答案实体的类型是河流；3) 在满足前两个条件的所有实体中，根据长度属性进行降序排列，目标答案排在第二位。具体分析，第一条语义类似于简单问题，描述相关实体和答案间的关联，第二条语义则描述了知识库中的特定类型与答案的包含关系，第三条语义和序数相关，它甚至不能简单地对应到知识库中已有的事实三元组。由此可见，我们需要挖掘出多条不同的关系，才能准确地定位目标答案。对于这类无法通过单个三元组查询来精确描述语义的问题，我们将它称为“复杂问题”，也是这个章节研究的重点。



What is the second longest river in the United States?

图 5-1 一个具有复杂语义的问句示例。

Figure 5-1 Running example of complex question.

回答复杂问题的核心，在于问答系统是否能准确理解问句中多部分语义之间的组合关系，而不仅仅是通过搜索的方式得到答案。这条思路对应了解决自动问答的语义解析技术（Semantic Parsing）<sup>[39, 40]</sup>。对于一个问句，基于语义解析的模型会将其转换成一棵语义解析树，这样的解析树等价于知识库中的查询图（Query Graph），与关系理解中的模式图类似，是包含未知实体知识库子结构。本章中，“语义解析树”，“查询结构”和“查询图”表示同一概念。图5-1为问题“What is the second longest river in United States ?”的查询图，具有树形结构。代表未知答案的节点 A 为解析树的根节点，三个叶节点 US, river, 2 则由问句的字面描述中抽取出来，并已链接到知识库中的实体、类型、时间或是数值上。这些叶节点通过知识库中的谓词（序列）与答案节点连接，从而对未知答案进行限制，因此本节中也称叶节点为问句的“相关节点”。此外，近年来神经网络模型在提高自动问答系统的性能方面显示出了巨大的前景，在多个不同的自动问答数据集上，通过神经网络改善语义解析的方法成为了目前最先进的技术<sup>[42, 43, 139]</sup>。基于以上论述，本章所讨论的工作围绕语义解析技术结合神经网络模型的思路，并将其扩展至复杂问题场景。

语义解析模型可以分为两个部分：生成候选查询图，以及预测最佳查询图。候选查询图的生成可以采用自底向上的方式构建<sup>[40, 41]</sup>，或是分阶段形式，由简到繁逐步生成所有候选<sup>[42, 43]</sup>。预测最佳查询图，主要是基于计算问题和查询图之间的语义相似度，挑选出最佳查询图。对于回答简单问题，目前已有的神经网络模型主要遵循“编码-比较”框架，即首先利用卷积神经网络（CNN）或循环神经网络（RNN），将原始问题以及候选的谓词序列分别进行编码，形成在同一个向量空间中的两个不同的语义向量，两者之间的语义相似度则可以定义为向量空间中的距离度量。

当输入的问题具有复杂语义时，候选的查询图无法简化为线性的谓词序列，如何对复杂的查询图进行编码，成为了语义相似度模型的关键问题。一个较为直观的做法，是将整个查询图看做由答案节点到不同叶节点的路径集合，例如图5-1中的虚线框将查询

图分成三个语义成分，分别对应指向不同相关实体的谓词序列。这使得针对简单问题的神经网络模型可以被直接应用，即分别计算问句与不同语义成分的相似度分值，并将其聚合（平均或相加），用来代表问句与查询图整体的语义相似度。

这种基于查询图拆分的方式具有其合理性，每个语义成分仅对应一个相关实体，类似人类对问句推理得到的平行语义线索。然而，基于此法套用简单问题的神经网络模型，依然存在两个缺陷。第一个缺陷是，将独立的语义成分与问句直接比较会带来风险。对于简单问题，唯一的谓词路径代表了整个问句的语义，问句和查询对应的语义向量越相近，代表它们匹配度也越高。然而复杂问题的查询图中，每一个独立的路径仅包含问句部分语义，即便是正确的谓词路径，与问句整体依然存在语义差距。若整体相似度由各部分相似度相加产生，则可能导致训练陷入局部极值，即问句经编码后的语义向量倾向于查询图中的某条特定谓词路径，而难以和其余正确的语义成分产生匹配。第二个缺陷是，分别计算相似度再简单相加的形式会丢失信息。将查询图的多个谓词序列分别进行编码，计算相似度再合并，这样的做法视作互相独立的多个部分。因此这样的模型无法理解不同语义成分之间存在的重叠、互补等语义交互。模型没有学习整个查询图的语义向量，因此无法从一个全局的角度描绘复杂查询图所包含的语义组合。

已有的文献 [42, 139] 尝试规避上述两个缺陷，它们的共同点在于从查询结构中仅挑选一条主路径，与问句计算语义相似度，对于查询结构中的其它限制，则依赖于人工定义的规则特征，或引入外部非结构化文本进行额外过滤。问答模型效果得以提升，但并没有直接应对这样的不足。

在本章中，我们着手于利用神经网络模型改善问句与复杂查询图之间语义相似度计算的效果，并尝试解决之前论述的两个缺陷。该模型整体基于对问句和谓词序列的编码，将其表示为同一个语义空间下的语义向量。我们的模型和之前方法主要区别，在于模型对各个语义成分编码后的向量进行结合，形成对于查询图整体的语义向量表示。同时，为了弥补问句和语义成分之间的信息不对等，在对问句进行编码的过程中，我们利用依存语法分析 (Dependency Parsing) 寻找问句中和特定谓词序列相关的局部信号，以此作为对问句字面信息的补充，使模型能更好地将问句和不同的语义成分对齐。

本章的贡献可以总结为以下四个部分：

1. 提出了一个轻量化和有效的神经网络模型来解决具有复杂语义的自动问答任务。  
据我们所知，这是第一次尝试在模型中对复杂查询图的完整语义进行明确编码；
2. 通过融入依存语法分析信息来丰富模型中问句的语义表示，并进行模型分析以验证其有效性；
3. 通过一种集成的方法，对已有的实体链接工具进行改良，丰富从问句中获得的候选实体，并进一步提升任务的整体效果；

4. 在多个自动问答数据集上进行实验，在由复杂问题组成的 ComplexQuestions 数据集中，模型的效果超过了已有的方法，在主要有简单问题组成的 WebQuestions 和 SimpleQuestions 数据集中，模型依然具有很强的竞争力。

## 5.2 相关工作

基于知识库的自动问答是最近几年的热门研究。最主要的用于解决自动问答的方法可以分为两类：基于信息抽取（Information Extraction）和基于语义解析（Semantic Parsing）。

基于信息抽取的问答模型首先通过实体链接寻找句子中的相关实体，将它们在知识库上邻近的实体抽取出作为候选答案。对于候选答案的排序，则依赖以候选答案为中心的知识库子图与问句之间的关联特征。早期的文献 [35] 利用特征工程进行训练，而一系列深度学习模型<sup>[37, 95, 96]</sup>则通过神经网络学习答案在类型、谓词、上下文等多个不同维度与问句的语义关联程度，并取得了明显的效果提升。基于语义解析的系统则会先生成带有复杂结构的候选查询图，将查询图翻译为能在运行在知识库上的结构化查询语句，得到最终的答案。早期的语义解析系统<sup>[38, 39]</sup>根据 PCCG 文法生成和具体知识库无关的中间表达形式，通常以  $\lambda$  算子的形式呈现，再将  $\lambda$  算子中的谓词和常量，映射到知识库中的具体谓词和实体。Liang 提出的  $\lambda$ -DCS<sup>[89]</sup> 是对 PCCG 的简化，语义解析树依然为自底向上的方式，但  $\lambda$  表达式由简单的相交、合并等规则生成，大大降低了解析树生成的复杂程度。最近的研究中，分阶段候选差选图的生成<sup>[42, 43]</sup>已证明了其有效性，它利用深度搜索，通过由简到繁逐步扩展查询图，不需要定义操作，也摆脱了自底向上生成过程中，组合顺序与单词顺序相关的限制。

随着深度学习的发展，神经网络模型被广泛使用于知识库上的自动问答任务，并且展示出了优秀的结果。这些方式的基本思路是利用神经网络的对特征表示的学习能力，将问句转换为连续空间上的向量表示，同时再将查询结构（或答案实体）映射到同一语义空间，并定义问句和答案的语义相似度，根据 <问题, 答案> 对进行学习，预测正确的查询。处理简单语义的神经网络问答模型具有较多的变种，例如文献 [48, 92] 使用了字符级别的循环神经网络以及注意力机制，对谓词序列和相关实体均进行相似度计算，对于未在训练数据中观察到的单词，模型依然具有鲁棒性；Bordes 等人<sup>[46]</sup>利用知识库向量学习，关注候选答案在知识库中的类型、相连谓词、相邻实体等信息，学习它们在知识库上的向量表示，并以此对候选答案进行编码；Yu 等人<sup>[49]</sup>引入了多层循环神经网络，并通过残差连接的方式，同时捕捉问句在词级别和整体级别与特定谓词序列的语义匹配；Qu 等人<sup>[93]</sup>提出了 AR-SMCNN 模型，除了利用循环神经网络捕捉问句和谓词序列在语义上的相关性，还利用了类似与卷积神经网络处理二维图像的方式，在词级别

相似度矩阵中寻找纹理，学习问句和谓词序列的另一种相似度量。

对于利用神经网络回答复杂语义的问题，已有的工作进行了不少尝试，但并没有尝试学习查询图整体的语义表示。例如文献 [42, 139] 倾重于用神经网络计算问句和查询图中主路径的匹配关系，相当于退化至简单语义场景。对于查询图中，除去主路径的其余语义成分，Yih 等人<sup>[42]</sup> 利用人工定义特征捕捉少数特殊语义，但基于特征工程的方法不具有较好的扩展性；Xu 等人<sup>[139]</sup> 则挖掘非结构化文本中的上下文信息，对满足主路径的候选答案进行过滤，这种方式被视为模型计算之后的处理，而并没有从本质上解决问题。Bao 等人<sup>[43]</sup> 利用每个相关实体在问句中的上下文窗口表示局部语义，并和查询图中的对应的谓词路径进行相似度匹配计算，但谓词路径之间仍缺少关联。

此外，依存语法分析可以描述一个句子中，词汇间的远距离依赖关系，考虑到它与查询图的结构较为相似，因此候选查询结构的生成可以基于依存分析树进行转换，语义匹配过程也更多利用了结构上的相似关系，例如文献 [90, 91]。我们的模型同样使用了依存语法分析，但将其视为语义特征的信息来源，而并非直接决定候选查询图的形状，因此我们可以生成更灵活的查询图。

### 5.3 我们的方法

本节将具体阐述复杂语义下的自动问答模型。主要包括四个部分：1. 基于分阶段的方式生成所有候选查询图；2. 通过神经网络定义问句和查询图整体之间的语义相似度；3. 基于集成的方式对已有的实体链接结果进行扩充；4. 具体的训练以及测试流程。

#### 5.3.1 分阶段查询图生成

本节中主要阐述分阶段候选查询图的生成过程。与已有的工作比较，例如文献 [43]，我们对候选生成的策略进行了优化，主要利用了查询图中对答案类型的隐含限制，以及知识库中用来维护和时间段事实相关的特殊设计。本文中，我们主要考虑四种不同的语义限制，分别是实体、类型、时间、顺序限制。例如在问句中，实体限制描述了答案与某已知实体的联系，顺序限制描述了答案按某种方式排序所具有的序号。以图5-2为例，我们通过问句“who is the youngest president of the united states after 2002?” 阐述候选图的具体生成过程，该问句同时包含了上述四种语义限制。为了方便描述，本节假设 Freebase 为问答系统所使用的知识库。

**阶段一：相关节点链接。**该步骤寻找问句中代表相关实体、类型、时间、顺序的词汇或短语，并链接到知识库上。相关节点作为候选查询图的叶节点，是不同类别语义限制的起点。图5-2(a) 列出了可能的 < 短语，叶节点 > 对，同一个短语可以对应到多个候选叶节点。不同语义限制类别（实体、类型、时间、顺序）的叶节点有着各自的链接

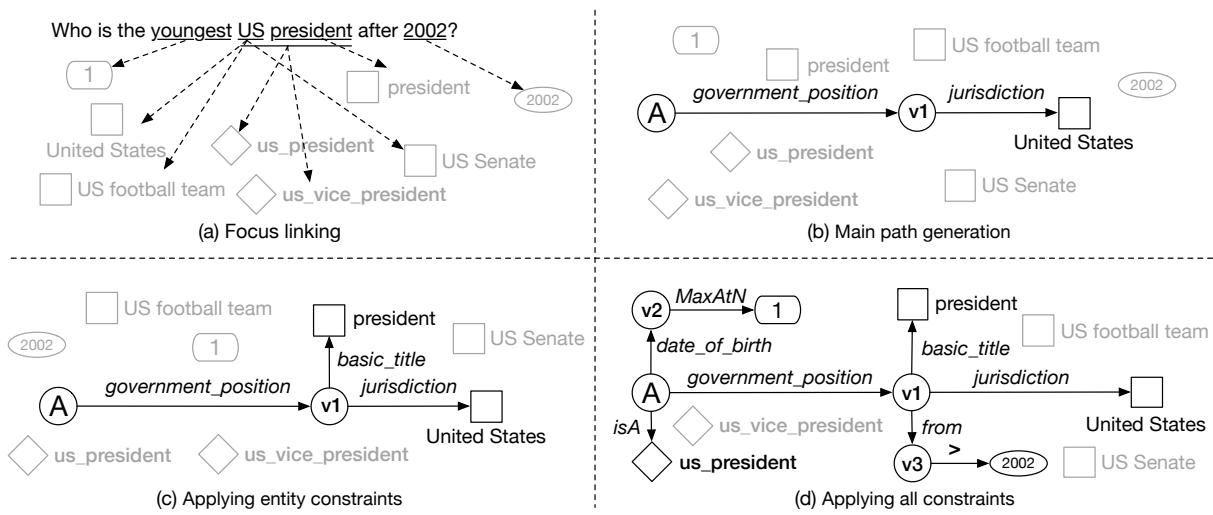


图 5-2 分阶段候选图生成的具体例子。

Figure 5-2 Running example of candidate generation.

方式。对于实体链接，我们使用了已有的链接工具 S-MART<sup>[54]</sup>，在多个已有的自动问答研究均被使用。S-MART 对所有可能的 < 短语，实体 > 进行打分，并保留了至多前十组结果。对于类型链接，考虑到知识库中不同的类型数量有限，我们枚举问句中所有长度不超过 3 的短语，并根据预训练的词向量，计算不同短语和类型之间的余弦相似度，同样保留至多前十组结果。对于时间链接，我们通过正则表达式识别句中出现的所有年份。对于顺序链接，我们利用预先定义的形容词最高级词汇列表（例如 largest, highest, latest 等描述客观事实的最高级词汇），并在问句中匹配最高级词汇，或“序数词 + 最高级”的词组，如“second longest”。对应的叶节点表示顺序值，若匹配到序数词，则顺序值为序数词对应的数字，否则为 1。如图 5-2(a) 所示，<“youngest”，1> 为生成的唯一顺序链接。

**阶段二：生成主路径。** 主路径是一个查询图的基础，代表着问句最主要的语义。考虑到几乎所有的事实类问题都和问句中至少一个实体相关，因此它被定义为从答案出发，通过谓词序列连接至某个实体节点的路径，等同于一个简单问题的查询图。我们枚举所有被链接的实体，以及它们在知识库中相连的合法谓词序列，即可生成一系列候选主路径。谓词序列的长度为 1 或 2，后者实质是描述了多元关系中某两个实体的关联。图 5-2(b) 显示出了某一个主路径，其中答案节点 A 以及中间节点 v1 都是变量节点。对于后续更复杂的语义限制，在图中均表示为由主路径上某变量节点出发，指向特定的叶节点的谓词序列。

**阶段三：添加额外实体语义限制。** 这个步骤的目的是在主路径之上扩充与实体相关的语义限制。受到 4.2.4.1 节中复杂模式图生成的启发，我们同样采用深度优先搜索的方

式，由简到繁进行查询图生成。对搜索空间中的每一个查询图，我们尝试单个谓词连接不同的变量节点与实体节点，构建出具有不同复杂程度的查询图。如图5–2(c)所示，在主路径上添加的实体语义限制为( $v_1, basic\_title, president$ )。基于深度优先搜索的优势在于查询图中的实体数量不受限，和基于模板的候选生成方法相比，具有更高的覆盖率，同时搜索过程中可以通过剪枝策略排除无法生成答案的查询图，提高候选生成速度。

**阶段四：添加类型限制。**类型限制只能和答案节点关联，利用知识库中的*IsA*谓词连接某个具体的相关类型节点。在该步骤中，我们对已有方法进行了改进：通过答案节点直接连接的谓词，推测出其具有的隐含类型，以此对类型限制进行过滤。如图5–2(c)所示，与答案直接相连的谓词为*government\_position*，根据知识库对谓词的定义，其主语类型为*politician*，因此成为答案的隐含类型。因此，我们可以过滤与隐含类型无关联的相关类型节点，从而防止语义偏离，并提升候选差选图的生成速度。具体而言，为了定义两个类型是否相关，我们采用了4.1.2.3节中通过松弛类型包含构建的Freebase类型层次关系。若某相关类型不包含任意一个隐含类型，或不被任意一个隐含类型包含，我们则将其视为无关类型，不用于候选生成。

**阶段五：生成时间、顺序限制。**完成类型限制的添加后，主路径上所有变量节点的类型（显式类型限制以及隐含类型）都已确定，因此我们可以枚举隶属于这些类型的特定谓词，完成时间和顺序限制的添加。如图5–2(d)所示，时间限制通过长度为2的谓词序列表示，例如序列[*from*, >]，其中前一个谓词在知识库中指向时间，后一个谓词为虚拟谓词，指明了和特定时间比较的方向，由问句中位于时间前的介词进行确定，例如“before”，“after”以及“in”。类似地，顺序限制同样由长度为2的谓词序列表示，例如序列[*date\_of\_birth*, *MaxAtN*]，前者在知识库中指向整数、浮点数或时间，后一个谓词表示降序排列。我们并不能从问句中获取直接的信号确定排序方向<sup>1</sup>，因此生成具体的排序限制时，两种方向都进行枚举。值得注意的是，对于时间限制，我们的方法进行了针对性优化。已有的文献[42, 43]仅考虑使用一条谓词与时间相连，我们的改进在于使用了知识库中存在的成对时间谓词，来描述更加准确的时间限制。Freebase中，成对时间谓词用来描述和时间段相关的事，例如图5–2(d)中的*from*谓词，存在谓词*to*与之对应<sup>2</sup>，两者分别为起始时间谓词和终止时间谓词。我们通过简单的名称匹配方式，收集了知识库中356组成对谓词，对于时间比较为“in”的形式，例如句中出现“in 2002”，我们在图中使用起始时间谓词进行连接，但生成SPARQL查询语句时，起始和终止谓词均会被使用，从而确保问句中的相关时间能够限制在一个时间段内，而不是仅仅等同

<sup>1</sup>部分形容词最高级较为明显，例如largest, longest等词几乎一定对应降序排列，但为了减少人工指定的规则，我们不对这些形容词事先指定方向，而是通过模型训练进行学习。

<sup>2</sup>*from*和*to*分别为*government.government\_positions\_held.from*和*government.government\_positions\_held.to*的简写。

于起始或终止时间点。

所有阶段结束后，我们将所有生成查询图转换为 SPARQL 查询语句，并在 Freebase 中查询最终答案。图5–2(d) 中的查询图对应的完整 SPARQL 查询语句<sup>1</sup>对应如下：

代码 5–1 SPARQL 查询语句示例

```

1 PREFIX fb: <http://rdf.freebase.com/ns/>
2 SELECT ?ans ?name WHERE {
3   ?ans fb:government.politician.government_positions_held ?v1 .
4   ?v1 fb:government.government_position_held.jurisdiction_of_office fb:m.09c7w0 .
5   ?v1 fb:government.government_position_held.basic_title fb:m.060c4 .
6   ?v1 fb:government.government_position_held.from ?v3 .
7   ?ans fb:type.object.type fb:government.us_president .
8   ?ans fb:people.person.date_of_birth ?v2 .
9   ?ans fb:type.object.name ?name .
10  FILTER (?v3 >= "2002-01-01"^^xsd:dateTime) .
11 } ORDER BY DESC(?v2) LIMIT 1

```

最后，我们舍弃掉没有结果的查询图，以及使用的相关实体对应词组出现重叠的查询图。和已有系统相比，本节的候选图生成使用了更少的人工规则，并在类型限制和时间限制上进行了改进，加快生成速度的同时，描述更加准确的语义限制。

### 5.3.2 基于神经网络的语义匹配模型

本节介绍的语义匹配模型如图5–3所示。作为预处理部分，查询图中使用的实体（或时间）节点对应于问句中的短语被替换为单词  $\langle E \rangle$ （或  $\langle Tm \rangle$ ），这样问句的语义将不会被具体的实体或年份所干扰。为了对查询图整体进行编码，我们首先将其分拆为从答案节点出发，指向不同叶节点的谓词路径，也称为语义成分。同样为了去除具体的实体、时间、顺序值对语义的干扰，谓词序列不包括叶节点的信息，类型限制是一个特例，作为模型输入的谓词序列为 [ *IsA*, *river* ]，类型节点的信息被包含在内。接下来将逐个介绍对问句和谓词序列的编码，基于查询图整体语义表示计算相似度的方式。

#### 5.3.2.1 语义成分编码

为了对语义成分  $p$  进行编码，模型对主要利用谓词序列的名字信息，以及每个谓词在知识库中的编号信息。以图5–3为例，查询图的第一个语义成分仅由一个谓词构成，对应的编号序列为 [ *contained\_by* ]。将序列中的每个谓词在知识库中显示的名字相连，即可得到谓词名字序列，即 [ “*contained*”, “*by*” ]。

<sup>1</sup> *m.09c7w0* 为实体 “United States”， *m.060c4* 为实体 “President”。

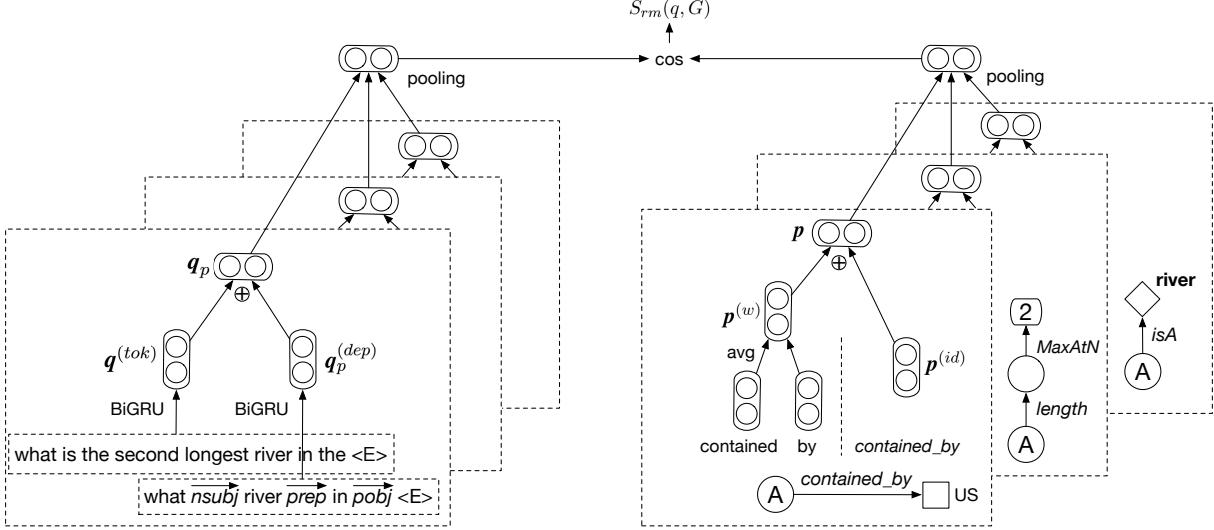


图 5-3 语义匹配模型的整体结构

Figure 5-3 Overview of proposed semantic matching model.

对于语义成分的谓词名字序列  $\{p_1^{(w)}, \dots, p_n^{(w)}\}$ , 我们首先通过词向量矩阵  $E_w \in \mathbb{R}^{|V_w| \times d}$  将原始序列变为词向量  $\{\mathbf{p}_1^{(w)}, \dots, \mathbf{p}_n^{(w)}\}$ , 其中  $|V_w|$  表示自然语言词汇数量,  $d$  表示词向量维度。接着我们采用词平均的方式计算整个名字序列的语义向量, 即  $\mathbf{p}^{(w)} = \frac{1}{n} \sum_i \mathbf{p}_i^{(w)}$ . 对于谓词编号序列  $\{p_1^{(id)}, \dots, p_m^{(id)}\}$ , 我们将整个序列视为整体, 并根据序列级别的向量矩阵  $E_p \in \mathbb{R}^{|V_p| \times d}$ , 直接转换为语义向量表示, 其中  $|V_p|$  代表训练数据中不同的编号序列数量。之所以将编号序列看做整体, 而不使用编号的向量平均或循环神经层表示语义, 主要原因有以下三点: 1) 根据候选图生成方式, 每个语义成分的谓词编号序列长度不超过 3; 2) 通常情况下, 对单个谓词序列进行打乱重排操作, 新的序列是非法的, 不会出现在其它查询图中; 3) 不同的谓词序列数量约等于知识库中不同的谓词数量, 不带来成倍增长。将名字序列和编号序列的向量进行按位置相加, 我们得到了单个谓词序列的向量表示,  $\mathbf{p} = \mathbf{p}^{(w)} + \mathbf{p}^{(id)}$ .

### 5.3.2.2 问句编码

对问句的编码需要考虑全局和局部两个层次, 其目的是捕捉问句中与某特定语义成分  $p$  相关的语义信息。对问句全局语义的编码, 输入信息为问句词序列。我们利用同一个词向量矩阵  $E_w$  将词序列向量化, 得到  $\{\mathbf{q}_1^{(w)}, \dots, \mathbf{q}_n^{(w)}\}$ 。将该输入通过双向 GRU 层<sup>[140]</sup>, 并将前向序列和后向序列的最后一个隐藏状态进行拼接, 作为整个词序列的语义向量:  $\mathbf{q}^{(tok)} = [\overleftarrow{\mathbf{h}}_1^{(w)}; \overrightarrow{\mathbf{h}}_n^{(w)}]$ .

为了对表示问句的局部语义, 核心在于提取与特定语义成分对应的信息。我们在模

型中利用依存语法分析，寻找答案与语义成分中的实体之间的依赖关系。由于在问句中，wh-词用于指示答案，因此我们抽取依存语法树中，连接 wh-词和实体所对应短语的路径，该路径有且仅有一条。与<sup>[139]</sup>类似，在依存语法树上的一条路径包含了词，以及词之间带有方向的依存弧。例如图5-3中的句子，答案“what”与实体“United States”之间的依存路径为 [ what,  $\overrightarrow{nsubj}$ , is,  $\overrightarrow{prep}$ , in,  $\overrightarrow{pobj}$ , <E> ]。我们使用另一个具有不同参数的双向 GRU 层，对依存路径进行编码，生成向量表示  $\mathbf{q}_p^{(dep)}$ ，其中包含了语法层面的以及与语义成分  $p$  直接相关的特征。最后，我们同样将句子在两种粒度上的向量进行按位置相加，得到整个问句对应特定语义成分的向量表示， $\mathbf{q}_p = \mathbf{q}^{(tok)} + \mathbf{q}_p^{(dep)}$ 。

### 5.3.2.3 语义合并

给定具有  $N$  个语义成分的查询图  $G = \{p^{(1)}, \dots, p^{(N)}\}$ ，每个语义成分已经被投影至同一个连续语义空间上的不同向量，体现了不同方面的隐藏特征。受卷积神经网络应用于二维图像处理所启发，图像整体的特征表示取决于是否存在某些局部区域，其样式与对应隐藏特征相吻合，而忽略这些局部区域的相对位置。考虑到复杂查询图内部的多个语义成分是并列的，互相之间并无次序之分，因此，模型对语义成分的向量表示进行最大池化（Max Pooling），获得整个查询图的组合语义表示。相应地，针对每个语义成分所对应的问句语义表示，我们同样进行最大池化操作，将多个语义向量合并为问句的整体表示。最后，我们利用余弦相似度计算问句和整个查询图之间的语义相似程度：

$$S_{rm}(q, G) = \cos(\max_i \mathbf{p}^{(i)}, \max_i \mathbf{q}_p^{(i)}). \quad (5-1)$$

基于以上框架，本节提出的语义相似度模型能尽可能使问句与单个语义成分具有可比性，同时捕获查询图不同部分之间的互补语义特征。

### 5.3.3 实体链接扩充

S-MART 实体链接器<sup>[54]</sup>在本模型中类似于一个黑箱，不具有操控性，并且生成的结果倾向于高准确率，而牺牲了一定召回率。为了在实体链接步骤寻找一个更好的准确率与召回率间的平衡，我们提出了一个基于集成的方式对实体链接结果进行扩充。首先，我们通过维基百科建立一个大的<词组，实体>对应表，每个实体和如下词组相对应：1) 实体页面的标题；2) 实体所在的重定向、消歧义页面标题；3) 实体在其它实体页面提及的链接文字，即锚文本（Anchor Text）。之后，每一对<词组，实体>都关联上一组统计特征，包括实体的链接概率、词级别的 Jaccard 相似度、三连字符级别的 Jaccard 相似度、实体在维基百科中的热门度、实体在知识库中的热门度。最终，我们使用一个双层全连接的线性回归模型，将所有出现在 S-MART 链接结果中的词组实体对

作为模型训练数据，用来拟合每一对的 S-MART 链接分值。模型训练完毕后，词组实体对应表中的每一对条目都将计算出一个虚拟的链接分值。对于每个问题，我们挑选出不在 S-MART 已有结果中，且分数排在前  $K$  位的条目，作为实体链接结果的扩充，阈值  $K$  为模型超参数。

### 5.3.4 问答系统整体训练及预测

为了从一系列候选中预测最佳查询图，我们用  $S(q, G)$  表示问句  $q$  和查询图  $G$  之间的整体关联分值。前一小节的语义匹配模型关注谓词路径层面的相似性，而整体关联分值还涉及到更多维度的特征，例如实体链接的置信度，以及查询图本身的结构特征。所以  $S(q, G)$  为一系列实体链接、语义匹配、查询结构层面上的特征进行加权求和而得。表5-1为完整的特征列表，实体链接特征为链接分数之和，以及每个链接的来源（S-MART 或链接扩展）；语义匹配特征即神经网络的输出  $S_{rm}(q, G)$ ；查询图结构特征为不同类别限制的数量、主路径长度以及输出的最终答案个数。我们利用最大间隔损失函数进行模型训练，尽可能较好查询图  $G^+$  和较差查询图  $G^-$  之间的分数差距：

$$loss = \max\{0, \lambda - S(q, G^+) + S(q, G^-)\}, \quad (5-2)$$

由于问答数据集通常只包含正确答案，而不标注查询图，我们依据查询图生成的答案对应的  $F_1$  分数区分正负样本。对于每一个  $F_1$  分数高于一定阈值（设定为 0.1）的查询图，我们将其视为正样本  $G^+$ ，并从候选集中随机选择最多 20 个具有更低  $F_1$  的查询图作为  $G^-$ ，组成不同的样本对。

表 5-1 预测最佳查询图所使用的特征。

Table 5-1 Full set of features for predicting query graphs.

Category	Description
Entity	Sum of linking scores of all entities; Number of entities from S-MART; Number of entities from enriched lexicon;
Semantics	Semantic similarity score $S_{rm}(q, G)$ ;
Structural	Number of each kind of constraints in $G$ ; Whether a kind of constraints is used in $G$ ; Whether the main path is one-hop; Number of output answers, discretized by {1, 2, 3, 5, 10, 50}.

## 5.4 实验

本节主要介绍我们所使用的自动问答数据集，以及用于比较的已有问答模型。具体实验包括在多个数据集上的端到端测试，以及一系列切除测试，用来分析方法中不同模块的重要性。

### 5.4.1 实验设置

**自动问答数据集：**我们在实验中使用了三个开放领域的数据集，分别为 ComplexQuestions<sup>[43]</sup>，WebQuestions<sup>[40]</sup> 以及 SimpleQuestions<sup>[47]</sup>，对应缩写为 CompQ，WebQ 和 SimpQ。CompQ 数据集来源于 Bing 搜索引擎日志，一共包含 2,100 个具有复杂语义的问题，以及人工标注的答案，前 1,300 个问句为训练集，后 800 为测试集。WebQ 数据集收集了 5,810 个通过 Google Suggest API 抓取的问题，以及对应的人工标注答案，约有 15% 的问句为复杂语义，同样数据集被分为 3,778 句训练集，以及 2,032 句测试集。SimpQ 一共包含 108,442 个具有简单语义的问句以及标注的答案，答案形式为 < 相关实体，谓词 > 对，我们主要利用该数据集进行补充实验，验证回答复杂问题的模型在简单语义场景中的性能。对于其它自动问答的数据集，例如 QALD，由于测试集数量过小，我们没有在这之上进行实验。

**知识库：**对于在 CompQ 和 WebQ 上进行的实验，我们跟随文献 [40, 139] 的实验设置，使用完整版本的 Freebase<sup>1</sup> 作为知识库，共包含约 46,000,000 个不同实体，以及 5,323 种不同谓词。同时通过开源图数据库 Virtuoso<sup>2</sup> 对 Freebase 进行访问与查询。对于 SimpQ 上进行的实验，我们使用数据集中提供的 FB2M 知识库，它是 Freebase 的一个子集，包含大约 2,000,000 个实体和 10,000,000 个事实三元组。

**模型实现及调参细节：**对本节中的所有实验，我们使用基于 GloVe<sup>[59]</sup> 预训练的词向量作为模型词向量矩阵的初始化。词向量维度  $d$ ，以及双向 GRU 层的隐藏状态维度均设为 300。损失函数中的  $\lambda$  的调参范围为 {0.1, 0.2, 0.5}，实体链接优化的集成阈值  $K$  范围为 {1, 2, 3, 5, 10, +INF}，训练批量大小  $B$  范围为 {16, 32, 64}.

### 5.4.2 端对端实验比较

我们首先对 WebQ 和 CompQ 数据集进行端到端测试。实验所使用的评价指标为所有测试问题的平均  $F_1$  分数。Berant 等人<sup>[40]</sup> 提供的官方评测代码<sup>3</sup>通过预测答案和标准答案的完全字面匹配计算每个问题的  $F_1$  分数，对于 CompQ 数据集，其中标注的实体

<sup>1</sup>该版本具体数据可从<https://github.com/syxu828/QuestionAnsweringOverFB>下载。

<sup>2</sup><http://virtuoso.openlinksw.com>.

<sup>3</sup><http://www-nlp.stanford.edu/software/sempre>.

名称和 Freebase 内实体名称存在大小写不一致的情况，因此我们参照 Bao 等人<sup>[43]</sup>的做法，计算  $F_1$  分数时忽略大小写。通过对验证集进行调参，WebQ 数据集的实验参数为  $\lambda = 0.5$ ,  $B = 32$ ,  $K = 3$ , CompQ 数据集的参数为  $\lambda = 0.5$ ,  $B = 32$ ,  $K = 5$ 。

表5-2列出了在两个数据集上的具体实验结果。Yih 等人<sup>[42]</sup>在 CompQ 上的实验结果基于 Bao 等人<sup>[43]</sup>对其模型的实现。在 CompQ 数据集上，我们提出的神经网络模型超过了其它已有方法，将平均  $F_1$  分数提升了 1.9，而在 WebQ 数据集上，与大量已有工作进行对比，我们的模型排在第二位，文献 [141] 基于记忆网络模型，成为分数最高的系统，其方法并不基于语义解析，无法直观解释一个答案是基于怎样的语义而生成，并且问答过程涉及的隐含语义与单一谓词路径相似，难以应对类型、时间、顺序等语义限制。需要指出的是，Xu 等人<sup>[139]</sup>利用维基百科的非结构化文本进行候选答案的验证，过滤掉满足主路径语义，但不匹配剩余语义的答案。由于此方法引入了大量由人工社区提供的额外知识，它达到了一个略高于我们方法的分数 (53.3)，但将此步骤去掉之后，模型分数跌落至 47.0。此外，文献 [42, 43] 额外使用了 ClueWeb 数据集<sup>[142]</sup>学习谓词与自然语言词组之间的语义匹配关系。根据 Yih 等人公布的比较结果，把这一部分信息移除之后，WebQ 数据集上的  $F_1$  分数将下降了约 0.9。此外，结果显示，扩充实体链接可以进一步提升问答系统的整体性能，在两个数据集上都获得了大约 0.8 的提升，是对语义匹配模型的一个良好补充。我们认为，和其它使用了 S-MART 链接工具的问答系统相比，我们的结果可以与之直接比较，这是因为 S-MART 的算法同样基于维基百科的半结构化信息进行学习，例如重定向链接、消歧义页面、锚文本等信息，实体链接扩充的步骤没有并没有引入额外的知识，因此可以直接比较。

针对语义匹配本身，我们在 SimpQ 数据集上进行了测试。由于 SimpQ 提供了标注的相关实体，我们可以消除实体链接步骤带来的差错，单独衡量语义匹配的性能。我们根据相关实体的名字，倒推出它在问句中对应的短语，将其替换为 <E> 之后，预测问句所表达的知识库谓词，使用准确率作为评价指标。表5-3列出了具体的实验结果。相关文献主要针对简单问题，尝试了许多模型变种，例如文献 [93] 的准确率最高，该模型利用循环神经网络对问句语义进行建模，同时利用卷积神经网络，从问句和谓词名称的词级别二维相似度矩阵中学习隐藏匹配样式。文献 [49] 使用了双层双向 LSTM 网络对问句进行编码，并在两层中使用残差连接方式捕捉不同粒度的语义。我们的语义匹配准确率略低一些，考虑到重点在于多个语义成分的组合，而不是回答简单问题，我们的模型更加轻量，同时 93.1% 的准确率也确保了模型的有效性。

### 5.4.3 模型分析

本节主要对模型的各个主要进行分析测试，并讨论模型回答错误的一些例子。

表 5–2 CompQ 和 WebQ 数据集上的实验结果，评价指标为平均  $F_1$  分数Table 5–2 Average  $F_1$  scores on CompQ and WebQ datasets.

Method	CompQ	WebQ
Dong et al. (2015) [96]	-	40.8
Yao et al. (2015) [36]	-	44.3
Bast et al. (2015) [45]	-	49.4
Berant et al. (2015) [143]	-	49.7
Yih et al. (2015) [42]	36.9	52.5
Reddy et al. (2016) [90]	-	50.3
Xu et al. (2016) [139] (w/o text)	-	47.0
Bao et al. (2016) [43]	40.9	52.4
Jain (2017) [141]	-	<b>55.6</b>
Abujabal et al. (2017)[144]	-	51.0
Cui et al. (2017) [44]	-	34.0
Hu et al. (2018) [91]	-	49.6
Talmor et al. (2018) [145]	39.7	-
Ours (w/o linking enrich)	42.0	52.0
Ours (w/ linking enrich)	<b>42.8</b>	52.7

#### 5.4.3.1 谓词路径表示

我们改变模型对谓词路径的编码方式，并在 CompQ 和 WebQ 上进行分析测试。首先对于谓词名字序列，我们尝试使用双向 GRU 层（和问句编码部分结构一致，但不共享参数）拼接隐藏状态的方式替代词向量平均。对于谓词编号序列，我们将对路径整体编码方式改为谓词向量的平均。

实验结果如表5–4所示。观察发现，前三行的基线方法移除了名字序列或编号序列，在两个数据集上的  $F_1$  分数明显低于后三行的方法。这说明了谓词的名字序列和编号序列所提供的语义可以互相补充。另一方面，对比最后两行实验，在 CompQ 数据集上，对名字序列使用词向量平均要优于使用双向 GRU，而在 WebQ 上，这个差距变得更小，我们认为原因主要来自于训练数据量的区别，WebQ 的训练集大小约为 CompQ 的三倍，因此可以支持更复杂的模型。

表 5–3 SimpQ 数据集上的语义匹配测试结果

Table 5–3 Accuracy on the SimpQ dataset.

Method	Relation Inputs	Accuracy
BiLSTM w/ words	words	91.2
BiLSTM w/ rel_name	rel_name	88.9
Yih et al. (2015) [42]	char-3-gram	90.0
Yin et al. (2016) [48]	words	91.3
Yu et al. (2017) [49]	words+rel_name	93.3
Qu et al. (2018) [93]	words+rel_separated	<b>93.7</b>
Ours	words+path	93.1

表 5–4 对谓词表示的分析结果。

Table 5–4 Ablation results on path representation.

Word repr.	Id repr.	CompQ $F_1$	WebQ $F_1$
None	PathEmb	41.11	51.86
Average	None	42.18	51.74
BiGRU	None	41.80	51.87
Average	Average	42.16	52.00
BiGRU	PathEmb	41.52	52.33
Average	PathEmb	<b>42.84</b>	<b>52.66</b>

#### 5.4.3.2 问句表示及语义组合

为了说明语义组合的有效性，我们建立一个基线模型：不使用公式5–2对应的最大池化操作，替代方式是分别计算每个问句表示和每个语义成分之间的相似度，并将各部分相似度分值相加，作为查询图与问句的整体相似度： $S_{rm}(q, G) = \sum_i \cos(\mathbf{p}^{(i)}, \mathbf{q}_p^{(i)})$ 。对于问句的编码方式，我们进行一系列比对实验，观察不使用字面序列或依存语法路径对整体性能带来的影响。

表5–5显示了在 CompQ 和 WebQ 上的具体比较结果。相比仅使用问句字面信息的模型，当依存语法分析提供的路径信息被使用后，问答系统整体性能平均提升了 0.42。在隐藏语义的角度，答案和相关实体之间的依存语法路径主要包含了词之间的语法依赖，以及每个词的功能化特征，是对整个问句序列信息的良好补充。然而，如果对问句编码只使用依存语法信息， $F_1$  分数会大幅度下降约 2.17。对于具有特殊语法结构的问

题，如果仅关注疑问词和实体短语间的路径，会使得模型丢失句中表达语义的关键词，例如以下两例：“*who did draco malloy end up marrying*”以及“*who did the philippines gain independence from*”，其中相关实体用斜体标出，代表语义的关键词为粗体。经过观察发现，WebQ 中大约有 5% 的问句具有类似的结构，在丢失关键语义信息后很难预测出正确的查询图。

语义组合的比较结果显示，模型中使用的最大池化操作要一致优于对应的基线方法。在 WebQ 上的提升要低于 CompQ，主要原因是 WebQ 中约 85% 的问句依然是简单语义形式，无法体现语义组合的区别。移除依存语法信息和池化操作的模型可以视为一个基础的利用深度学习改善语义解析的问答模型。在复杂语义场景中，局部信息和语义组合的引入，两者结合使得 CompQ 数据集上效果提升 1.28。

我们通过以下例子，进一步阐述模型中语义组合带来的优势。给定问句“*who is gimli's father in the hobbit*”，由于“gimli”的实体链接结果中既存在自然人，也存在名字一样的虚拟角色，我们主要关注下面两个可能代表真实语义的查询图：

1. (? , *children*, *gimli\_person*);
2. (? , *fictional\_children*, *gimli\_character*)  $\wedge$  (? , *appear\_in*, *hobbit*)。

两个查询图涉及到三个不同的语义成分，如果独立观察其中每一个语义成分，谓词 *children* 与问句整体的匹配程度最高，因为“father”一词包含了很强的语义信息，训练数据中也包含较多“'s father”和 *children* 的关联，因此它们的关联特征容易被学习。相比之下，*fictional\_children* 过于生僻，而 *appear\_in* 与“father”无关联，这两个语义成分的相似度远不如 *children*，因此基线模型认为第一个查询图更加正确。而我们的模型中，不同语义成分的隐藏特征通过池化方式汇集起来，分别将各自突出的隐藏语义传递出去，构成查询图整体的语义向量。与单独的 *children* 语义向量相比，查询图整体语义能兼顾与“'s father”以及“in the hobbit”匹配，因此模型能正确预测第二个查询图为答案。

表 5-5 问句表示和语义组合的分析测试。

Table 5-5 Ablation results on question representation and compositional strategy.

Composition	Q_repr	CompQ $F_1$	WebQ $F_1$
Baseline	sentential	41.56	52.14
Baseline	both	42.35	52.39
Ours	dependency	41.48	49.69
Ours	sentential	42.59	52.28
Ours	both	<b>42.84</b>	<b>52.66</b>

#### 5.4.3.3 错误分析

我们从 CompQ 数据集中完全回答错误的问题中随机挑选 100 个例子进行分析，并归纳出下列几类错误原因。

**主路径错误 (10%)**: 模型完全没有理解问句语义，哪怕最主要的语义也没有预测出来。这类错误对应的问题通常较难回答，例如“What native american sports heroes earning two gold medals in the 1912 Olympics”。

**语义限制错误 (42%)**: 模型预测的查询图中包含正确的主路径，但其余语义限制存在偏差。比较典型的一类限制是隐含时间限制，例如问句“Who was US president when Traicho Kostov was teenager”无法准确回答，因为“when Traicho Kostov was teenager”暗示了时间限制，受限于候选生成方法，这类限制无法被识别。

**实体链接错误 (16%)**: 这类错误的主要原因是问句中的一些实体词组具有高度歧义。例如问句“What character did Robert Pattinson play in Harry Potter”，而“Harry Potter”可以对应 7 部不同的电影，因此很难猜测问句中指的是哪一部。

**杂项 (32%)**: 包含了一些较明显的答案标注错误，以及问题本身语义不明确或不合逻辑。例如问句“Where is Byron Nelson 2012”，根据标注答案可以帮助确定问句中“Byron Nelson”的具体所指，然而此人已于 2006 年去世，因此该问题的真实意图难以捉摸，或许提问者想问的是他的逝世地点，或葬于何处。

### 5.5 小结

本章讨论了面向复杂语义的知识库自动问答任务，其难点在于复杂问句中包含多个关系，并不能转换为知识库上的简单三元组查询。我们沿用关系理解中的模式图思路，提出了基于复杂查询图的语义解析模型，以解决复杂问句的语义结构表示和语义匹配计算。据我们所知，我们的工作是首次通过神经网络模型学习查询图整体的连续语义表示，相对于已有工作，整体语义表示通过池化操作，聚合查询图中不同语义成分的特征，以捕捉其中的语义相近、互补等交互。与此同时，我们研究了提升问答效果的多种不同的方法，主要包括候选查询图生成的时间、类型限制优化，引入依存语法信息捕捉与特定语义成分的局部匹配，以及利用集成方法扩充实体链接结果，提高候选查询图的召回率。我们在三个广泛使用的问答数据集上进行了测试，在全部由复杂问题组成的 ComplexQuestions 中，我们提出的模型取得了目前最好的效果，并且显著优于已有模型；在主要由简单问题构成的 WebQuestions，以及全部为简单问题的 SimpleQuestions 中，基于复杂查询图的模型依然拥有竞争力，领先于绝大部分已有模型，同时语义匹配模型具有轻量级、参数少等优势，证明了其有效性。

后续的研究主要包括了对更多种语义限制的挖掘，例如隐含时间限制，即问句中不出现具体的时间，而是以从句形式描述与该时间相关的事件。一些研究工作对问句进行从句提取的方式，先回答从句部分，再将时间答案代回主句进行第二次回答。为了减少对问句进行特殊处理的步骤，我们会研究如何将隐含时间限制的挖掘纳入现有的查询图框架中，进一步提升问答模型效果和适用性。

本章的研究成果已发表于 2018 年国际会议 Empirical Methods in Natural Language Processing (EMNLP 2018)，论文题目为“Knowledge Base Question Answering via Encoding of Complex Query Graphs”。

## 第六章 总结与展望

### 6.1 论文工作总结与主要贡献

自然语言理解是人工智能的重要分支。如何让机器理解人类语言的含义，是一系列任务的研究重点，尤其是对于问答系统、阅读理解、多轮对话等下游任务，它们都依赖于机器对语义的充分认知。伴随着互联网中海量结构化信息积累，知识库的诞生和相关技术的发展给自然语言理解提供了一种有效的解决方案，即以知识库中的实体、类型和谓词为载体，描述自然语言中的实体、实体间的关系，甚至蕴含多个关系的复杂句子。在此背景下，本文对基于知识库的自然语言理解分为三个递进的层面，即实体理解、关系理解和问句理解。针对这三个层面理解问题，本文展开了一系列研究，并提出了具有针对性的语义匹配模型。

实体理解的目标，是将自然语言文本中表示实体的短语映射至知识库的对应实体，是一种直接匹配的过程。本文进行了中文到英文的跨语言场景中，对表格文本进行链接的研究。表格链接过程中，同行列的实体具有明显的相关性，这是传统实体链接任务所不具备的特性，也是链接模型的关注重点。而知识库和链接文本不在同一个语言中，使得模型无法利用任何字面上的相似信息，这给链接任务带来了更多挑战。本文是学术界首次研究跨语言的表格链接任务，本文提出了基于跨语言词向量和深度神经网络的链接模型，目标在于克服翻译步骤带来的错误传播，以及自动学习不同粒度的语义匹配特征。具体而言，本文提出的方法贡献如下：

1. 候选实体生成中，利用多种翻译工具进行过渡，并保留足量候选，将黑盒翻译工具出错的影响尽可能降低；
2. 训练跨语言词向量，使得中英文单词、实体的特征表示在连续语义空间中互通，保证在不依赖字面相似特征和共现统计特征的情况下，实现高质量的链接；
3. 定义了三种语义匹配特征，即单个单元格到实体的指示特征，单元格行列信息到实体的上下文特征，及同列实体之间的一致性特征，通过神经网络对三类特征进行表示学习，并提出了逐位方差进行一致性特征计算的方式；
4. 模型遵循联合训练框架，以整张表格级别的匹配程度作为目标函数，并利用基于成对排序损失的 RankNet 进行训练，充分利用负样本表格生成产生的偏序关系；
5. 实验表明，本文提出的模型在跨语言表格链接任务中明显优于其它基线模型，同时模型对一致性特征的建模以及联合训练框架均带来实质性的帮助。

关系理解的目标，是将自然语言中的二元关系通过知识库中的谓词进行表示。相对

于实体理解的直接匹配过程，关系理解较难做到二元关系和谓词的一一对应，一方面在于关系的多义性，更重要原因在于知识库和自然语言之间存在语义间隔，使得一些语法简单的关系，在知识库中却对应复杂的语义。基于这两个不同的挑战，本文对二元关系进行了两种不同粒度的研究。

粗粒度的关系语义研究中，本文旨在分析关系在大跨度上的多义性，挖掘关系的主语和宾语所具有的不同类型搭配。本文提出了挖掘关系具有代表性类型搭配的方法，其思路在于尽可能使用具体的类型匹配更多的已知关系三元组，主要贡献列举如下：

1. 提出了一种主宾语联合进行实体链接的方式，利用关系名称和主宾语间谓词路径存在的关联特征，提升整体链接准确率；
2. 去除关系名称中不影响类型搭配的成分，并利用语法变换将相似语义关系归为一组，使长尾关系能够被有效利用；
3. 利用松弛类型包含构建更丰富的知识库类型层次关系，并可用于其它任务中；
4. 人工测评实验表明，本文提出的方法可以改善互信息模型对热门类型搭配的惩罚情况，同时推理出的代表性的类型搭配也具有不错的质量。

细粒度的关系语义研究中，本文旨在深入挖掘关系语义的精确表达，定义了具有树形结构的模式图，它是知识库中满足特定语义的子图的抽象表达，同时具有良好的可解释性。本文提出了基于复杂模式图的规则推导模型，由已知关系三元组出发，挖掘语义相近的候选模式图，并学习它们的概率分布，从而以结构匹配的形式描述关系语义，并运用于知识库补全任务中。本文提出的方法贡献如下：

1. 定义了具有“路径 + 分支”结构的模式图，它是对传统规则推导模型中，基于谓词路径形式的规则扩展，对复杂语义关系具有更强的表示能力；
2. 利用深度优先搜索采集不同的模式图，并通过优先队列实现搜索过程的高效剪枝，在获取和关系语义较为接近的模式图同时，维持不同模式图间的多样性；
3. 将二元关系语义表示为候选模式图上的概率分布，可以更好应对关系的多义性，同时任何一个查询图自身都具有独立的描述能力，使人类易于理解；
4. 模式图概率通过生成模型学习，实现了宽泛和具体模式图之间的平衡；
5. 多个自然语言关系的模式图实例表明，基于模式图的结构有能力准确描述复杂关系语义，并且质量显著好于其它基于路径的规则推导模型；
6. 本文提出的模型能有效运用于知识库补全任务中，在主宾语预测和三元组分类两个子任务上，效果优于其它规则推导模型，以及新兴的知识库向量模型。

问句理解的目标，是学习问句和答案之间的推理匹配。本文关注于通过知识库回答客观事实类问题，由于单个问句可能包含未知答案和其它实体的多个关系，和语义仅对应单个谓词的问句相比，复杂问句的回答更具有挑战性，体现在如何对复杂问句进行语

义描述，以及如何度量和问句的语义匹配程度。针对以上挑战，本文提出了面向复杂语义问句的问答模型。对于问句的语义表示，本文沿用关系理解中的模式图思路，由问句出发生成可解释性高的查询图，以表示答案实体与问句中多个相关实体、类型、时间等信息的关联。同时，模型通过神经网络训练问句与查询图的匹配程度，为复杂查询图整体学习连续空间中的特征表示，捕捉不同成分间的语义交互。具体贡献如下：

1. 沿用模式图思路，利用多阶段生成方式构建问句的候选查询图，并在前人基础上对类型语义限制和时间语义限制进行改进；
2. 提出了一个轻量级的神经网络模型，以计算问句和查询图的语义匹配程度，据我们所知，这是知识库问答研究中首次尝试学习复杂查询图整体的连续语义表示；
3. 对问句的表示学习引入依存语法路径，作为问句字面序列信息的补充，以体现问句与特定语义成分的关联；
4. 通过集成方法，对已有实体链接工具的结果进行扩充，在链接准确率不受较大影响的前提下，提升候选查询图的召回；
5. 本文提出的模型在复杂问题数据集上取得了最优的效果，在简单问题数据集上依然保持竞争力，更多对比实验显示，学习查询图整体的连续特征表示有助于提升问答系统的效果。

## 6.2 未来工作展望

由于时间关系，本文的工作中还存在一些没有得到解决的问题，列举如下：

1. 表格链接，以及关系三元组的实体链接中，都存在着无法链接到具体实体的短语。除了较容易识别的数字、时间以外，考虑到知识库并不完整，部分实体（尤其是人名）不存在于知识库中，此时模型需要识别出这样的短语，而不是强行链接。我们对表格链接的任务定义绕开了此问题，而对三元组的实体链接则忽略了这种情况，这是一个需要改进的方向。
2. 关系三元组的链接方式较为粗糙，采用了主谓宾各自匹配度连乘的方式，并没有使用模型训练各部分权重。4.1.2节提到的集成链接方案并不是最优的解决办法，未来将利用神经网络表示三元组各自成分的链接特征，从而提升这一步骤的准确率。
3. 知识库问答研究中，我们尝试使用注意力层<sup>[63]</sup>取代依存语法序列，让语义匹配模型自动学习和特定谓词最相关的问句短语，但实验显示注意力层对问答指标几乎没有改进。一个可能的解释是，输入的问句长度大多在 10 左右，而不是类似一段话的形式，因此注意力模型效果不明显。在今后的研究中，会在这个问题上继续调研。

此外，在未来的研究工作中，我们以问句理解为核心，关注以下两个主要研究问题。关系理解和问句理解具有很高的相关性。给定问句中的二元关系，若已知其主宾语类型搭配，那么对于候选查询图而言，答案类型与类型搭配的查询图更有可能表示了正确的语义。类似地，二元关系所对应的模式图也可指引问句查询图的排序，提供额外的匹配特征。我们在过去的工作中，对主宾语类型搭配与自动问答的结合进行了一定的尝试，但效果提升有限，除了类型搭配本身出现偏差，将问句与特定二元关系的对应是另一个瓶颈。基于语法转换的方式进行映射过于确定，由于用户提问可能不具有严谨的语法，可能需要使用更加灵活的方式实现这一对应。在未来的研究中，我们将尝试由陈述句出发生成疑问句，并引入一定的非严谨语法形式，以此构建训练数据，学习更加准确的问句到二元关系的映射。真实问答系统的问答对数据（例如 Yahoo Answers，以及大量 FAQ 资源）可以帮助对不规范的语法进行理解和近似，使模型学习启发性的知识。

在现有的问答模型中，候选结构的生成过程是一次性的，对于测试问句，必须先生成所有查询图，再从中挑选最匹配的结构。为了保证候选生成速度，搜索规模需要受限，例如主路径长度限制为 2，对于某些特殊问句，则无法生成出正确的查询图。因此，一种可能的改进方式，是将查询结构的生成看做序列，通过使用序列到序列模型，以问句为输入，输出查询图的生成序列。Golub 等人<sup>[92]</sup> 使用了这样的模型用于回答简单问句，而 Jain<sup>[141]</sup> 使用记忆网络模型在 WebQuestions 上取得了最佳的效果，其模型的多层设计暗含了谓词的多步跳转。对于复杂问句，虽结构复杂，但多阶段生成过程很容易转换成序列形式，如何将复杂语义结构与序列到序列模型结合，是未来的一个研究方向。

## 参考文献

- [1] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an Architecture for Never-Ending Language Learning.[C]// AAAI. Vol. 5: 2010: 3.
- [2] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]// EMNLP: 2011: 1535-1545.
- [3] SCHMITZ M, BART R, SODERLAND S, et al. Open language learning for information extraction[C]// EMNLP: 2012: 523-534.
- [4] NAKASHOLE N, WEIKUM G, SUCHANEK F. PATTY: a taxonomy of relational patterns with semantic types[C]// EMNLP-CoNLL: 2012: 1135-1145.
- [5] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
- [6] KINGSBURY P, PALMER M. From TreeBank to PropBank.[C]// LREC. Citeseer: 2002: 1989-1993.
- [7] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[M].: Springer, 2007.
- [8] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]// WWW. ACM: 2007: 697-706.
- [9] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// SIGMOD: 2008: 1247-1250.
- [10] LESKOVEC J, RAJARAMAN A, ULLMAN J D. Mining of massive datasets[M].: Cambridge university press, 2014.
- [11] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American society for information science, 1990, 41(6): 391-407.
- [12] HOFGMANN T. Probabilistic latent semantic analysis[C]// Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.: 1999: 289-296.

- [13] MIHALCEA R, CSOMAI A. Wikify!: linking documents to encyclopedic knowledge[C]// Proceedings of the sixteenth ACM conference on Conference on information and knowledge management: 2007: 233-242.
- [14] GUO S, CHANG M W, KICIMAN E. To link or not to link? a study on end-to-end tweet entity linking[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 2013: 1020-1030.
- [15] RATINOV L, ROTH D, DOWNEY D, et al. Local and global algorithms for disambiguation to wikipedia[C]// Proceedings of ACL:HLT: 2011: 1375-1384.
- [16] SHEN W, WANG J, LUO P, et al. Linden: linking named entities with knowledge base via semantic knowledge[C]// Proceedings of the 21st international conference on World Wide Web. ACM: 2012: 449-458.
- [17] FRANCIS-LANDAU M, DURRETT G, KLEIN D. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks[C]// Proceedings of NAACL-HLT: 2016: 1256-1261.
- [18] SUN Y, LIN L, TANG D, et al. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation.[C]// IJCAI: 2015: 1333-1339.
- [19] GUPTA N, SINGH S, ROTH D. Entity linking via joint encoding of types, descriptions, and context[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: 2017: 2681-2690.
- [20] NGUYEN T H, FAUCEGLIA N, MURO M R, et al. Joint learning of local and global features for entity linking via neural networks[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers: 2016: 2310-2320.
- [21] RUDER S, VULI I, SØGAARD A. A survey of cross-lingual word embedding models[J]. ArXiv preprint arXiv:1706.04902, 2017.
- [22] GALÁRRAGA L A, TEFLIOUDI C, HOSE K, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C]// Proceedings of the 22nd international conference on World Wide Web. ACM: 2013: 413-422.
- [23] JIANG S, LOWD D, DOU D. Learning to refine an automatically extracted knowledge base using markov logic[C]// ICDM: 2012: 912-917.

- [24] ZHANG C, HOFFMANN R, WELD D S. Ontological Smoothing for Relation Extraction with Minimal Supervision.[C]// AAAI: 2012.
- [25] LAO N, MITCHELL T, COHEN W W. Random walk inference and learning in a large scale knowledge base[C]// EMNLP: 2011: 529-539.
- [26] GARDNER M, MITCHELL T. Efficient and expressive knowledge base completion using subgraph feature extraction[C]// EMNLP: 2015: 1488-1498.
- [27] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: 2014: 601-610.
- [28] NICKEL M, TRESP V, KRIEGEL H P. Factorizing yago: scalable machine learning for linked data[C]// WWW: 2012: 271-280.
- [29] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// NIPS: 2013: 2787-2795.
- [30] WANG Z, ZHANG J, FENG J, et al. Knowledge Graph Embedding by Translating on Hyperplanes.[C]// AAAI: 2014: 1112-1119.
- [31] LIN Y, LIU Z, SUN M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion.[C]// AAAI: 2015: 2181-2187.
- [32] XIAO H, HUANG M, ZHU X. TransG: A generative model for knowledge graph embedding[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1: 2016: 2316-2325.
- [33] CHEN D, FISCH A, WESTON J, et al. Reading Wikipedia to Answer Open-Domain Questions[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1: 2017: 1870-1879.
- [34] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[J]. ArXiv preprint arXiv:1611.01603, 2016.
- [35] YAO X, VAN DURME B. Information extraction over structured data: Question answering with freebase[C]// ACL: 2014: 956-966.
- [36] YAO X. Lean question answering over freebase from scratch[C]// NAACL: 2015: 66-70.

- [37] HAO Y, ZHANG Y, LIU K, et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge[C]// ACL: 2017: 221-231.
- [38] CAI Q, YATES A. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension.[C]// ACL: 2013: 423-433.
- [39] KWIATKOWSKI T, CHOI E, ARTZI Y, et al. Scaling semantic parsers with on-the-fly ontology matching[C]// ACL: 2013.
- [40] BERANT J, CHOU A, FROSTIG R, et al. Semantic Parsing on Freebase from Question-Answer Pairs.[C]// EMNLP: 2013: 1533-1544.
- [41] BERANT J, LIANG P. Semantic parsing via paraphrasing[C]// ACL: 2014: 1415-1425.
- [42] YIH W T, CHANG M W, HE X, et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base[C]// ACL-IJCNLP: 2015: 1321-1331.
- [43] BAO J W, DUAN N, YAN Z, et al. Constraint-Based Question Answering with Knowledge Graph.[C]// COLING: 2016: 2503-2514.
- [44] CUI W, XIAO Y, WANG H, et al. KBQA: learning question answering over QA corpora and knowledge bases[J]. Proceedings of the VLDB Endowment, 2017, 10(5): 565-576.
- [45] BAST H, HAUSSMANN E. More accurate question answering on freebase[C]// CIKM: 2015: 1431-1440.
- [46] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer: 2014: 165-180.
- [47] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks[J]. ArXiv preprint arXiv:1506.02075, 2015.
- [48] YIN W, YU M, XIANG B, et al. Simple Question Answering by Attentive Convolutional Neural Network[C]// COLING: 2016: 1746-1756.
- [49] YU M, YIN W, HASAN K S, et al. Improved Neural Relation Detection for Knowledge Base Question Answering[C]// ACL: 2017: 571-581.

- [50] LUKOVNIKOV D, FISCHER A, LEHMANN J, et al. Neural network-based question answering over knowledge graphs on word and character level[C]// Proceedings of the 26th international conference on World Wide Web. International World Wide Web Conferences Steering Committee: 2017: 1211-1220.
- [51] LIMAYE G, SARAWAGI S, CHAKRABARTI S. Annotating and searching web tables using entities, types and relationships[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1338-1347.
- [52] LIN T, ETZIONI O, et al. Entity linking at web scale[C]// Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction: 2012: 84-88.
- [53] HOFFART J, YOSEF M A, BORDINO I, et al. Robust disambiguation of named entities in text[C]// Proceedings of EMNLP: 2011: 782-792.
- [54] YANG Y, CHANG M W. S-MART: Novel Tree-based Structured Learning Algorithms Applied to Tweet Entity Linking[C]// ACL-IJCNLP: 2015: 504-513.
- [55] LUO G, HUANG X, LIN C Y, et al. Joint entity recognition and disambiguation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: 2015: 879-888.
- [56] JOACHIMS T. Optimizing search engines using clickthrough data[C]// Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: 2002: 133-142.
- [57] BHAGAVATULA C S, NORASET T, DOWNEY D. TabEL: entity linking in web tables[C]// International Semantic Web Conference. Springer: 2015: 425-441.
- [58] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in neural information processing systems: 2013: 3111-3119.
- [59] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]// EMNLP: 2014: 1532-1543.
- [60] MIKOLOV T, LE Q V, SUTSKEVER I. Exploiting similarities among languages for machine translation[J]. ArXiv preprint arXiv:1309.4168, 2013.

- [61] XU K, FENG Y, HUANG S, et al. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: 2015: 536-540.
- [62] XU Y, MOU L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]// Proceedings of the 2015 conference on empirical methods in natural language processing: 2015: 1785-1794.
- [63] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// ICLR: 2015.
- [64] FANG W, ZHANG J, WANG D, et al. Entity disambiguation by knowledge and text jointly embedding[C]// Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning: 2016: 260-269.
- [65] KLEMENTIEV A, TITOV I, BHATTARAI B. Inducing crosslingual distributed representations of words[J]. Proceedings of COLING 2012, 2012: 1459-1474.
- [66] LAZARIDOU A, DINU G, BARONI M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Vol. 1: 2015: 270-280.
- [67] HERMANN K M, BLUNSMON P. Multilingual distributed representations without word alignment[J]. ArXiv preprint arXiv:1312.6173, 2013.
- [68] GOUWS S, BENGIO Y, CORRADO G. Bilbowa: Fast bilingual distributed representations without word alignments[C]// International Conference on Machine Learning: 2015: 748-756.
- [69] VULI I, MOENS M F. Bilingual distributed word representations from document-aligned comparable data[J]. Journal of Artificial Intelligence Research, 2016, 55: 953-994.
- [70] XING C, WANG D, LIU C, et al. Normalized word embedding and orthogonal transform for bilingual word translation[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 2015: 1006-1011.

- [71] ZHANG Y, GADDY D, BARZILAY R, et al. Ten Pairs to Tag–Multilingual POS Tagging via Coarse Mapping between Embeddings[C]// Proceedings of NAACL-HLT: 2016: 1307-1317.
- [72] ARTETXE M, LABAKA G, AGIRRE E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: 2016: 2289-2294.
- [73] FARUQUI M, DYER C. Improving vector space word representations using multilingual correlation[C]// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics: 2014: 462-471.
- [74] HOTELLING H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321-377.
- [75] MURTHY V, KHAPRA M, BHATTACHARYYA P, et al. Sharing network parameters for crosslingual named entity recognition[J]. ArXiv preprint arXiv:1607.00198, 2016.
- [76] ZOU W Y, SOCHER R, CER D, et al. Bilingual word embeddings for phrase-based machine translation[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing: 2013: 1393-1398.
- [77] BORDES A, GLOROT X, WESTON J, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(2): 233-259.
- [78] LAO N, COHEN W W. Fast query execution for retrieval models based on path-constrained random walks[C]// Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: 2010: 881-888.
- [79] WANG Q, LIU J, LUO Y, et al. Knowledge base completion via coupled path ranking[C]// ACL: 2016: 1308-1318.
- [80] NICKEL M, MURPHY K, TRESP V, et al. A review of relational machine learning for knowledge graphs[J]. Proceedings of the IEEE, 2016, 104(1): 11-33.
- [81] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]// Advances in neural information processing systems: 2013: 926-934.
- [82] NICKEL M, ROSASCO L, POGGIO T. Holographic embeddings of knowledge graphs[C]// AAAI: 2016.

- [83] BORDES A, WESTON J, COLLOBERT R, et al. Learning Structured Embeddings of Knowledge Bases.[C]// AAAI. Vol. 6. 1: 2011: 6.
- [84] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 2013: 746-751.
- [85] XIAO H, HUANG M, HAO Y, et al. TransA: An adaptive approach for knowledge graph embedding[J]. ArXiv preprint arXiv:1509.05490, 2015.
- [86] CIMIANO P, LOPEZ V, UNGER C, et al. Multilingual question answering over linked data (qald-3): Lab overview[C]// International Conference of the Cross-Language Evaluation Forum for European Languages. Springer: 2013: 321-332.
- [87] ZETTLEMOYER L S, COLLINS M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[J]. ArXiv preprint arXiv:1207.1420, 2012.
- [88] KWIATKOWSKI T, ZETTLEMOYER L, GOLDWATER S, et al. Inducing probabilistic CCG grammars from logical form with higher-order unification[C]// EMNLP: 2010: 1223-1233.
- [89] LIANG P. Lambda dependency-based compositional semantics[J]. ArXiv preprint arXiv:1309.4408, 2013.
- [90] REDDY S, TÄCKSTRÖM O, COLLINS M, et al. Transforming dependency structures to logical forms for semantic parsing[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 127-140.
- [91] HU S, ZOU L, YU J X, et al. Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(5): 824-837.
- [92] GOLUB D, HE X. Character-level question answering with attention[J]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.
- [93] QU Y, LIU J, KANG L, et al. Question Answering over Freebase via Attentive RNN with Similarity Matrix based CNN[J]. ArXiv preprint arXiv:1804.03317, 2018.

- [94] SHEN Y, HE X, GAO J, et al. Learning semantic representations using convolutional neural networks for web search[C]// Proceedings of the 23rd International Conference on World Wide Web. ACM: 2014: 373-374.
- [95] BORDES A, CHOPRA S, WESTON J. Question answering with subgraph embeddings[C]// EMNLP: 2014.
- [96] DONG L, WEI F, ZHOU M, et al. Question Answering over Freebase with Multi-Column Convolutional Neural Networks.[C]// ACL (1): 2015: 260-269.
- [97] CAFARELLA M J, HALEVY A, WANG D Z, et al. Webtables: exploring the power of tables on the web[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 538-549.
- [98] WANG J, WANG H, WANG Z, et al. Understanding tables on the web[C]// International Conference on Conceptual Modeling. Springer: 2012: 141-155.
- [99] WU T, YAN S, PIAO Z, et al. Entity Linking in Web Tables with Multiple Linked Knowledge Bases[C]// Joint International Semantic Technology Conference. Springer: 2016: 239-253.
- [100] MUÑOZ E, HOGAN A, MILEO A. Using linked data to mine RDF from wikipedia's tables[C]// Proceedings of the 7th ACM international conference on Web search and data mining. ACM: 2014: 533-542.
- [101] SEKHAVAT Y A, DI PAOLO F, BARBOSA D, et al. Knowledge Base Augmentation using Tabular Data.[C]// LDOW: 2014.
- [102] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: Bringing order to the web.[R]. Stanford InfoLab, 1999.
- [103] JI H, GRISHMAN R, DANG H T, et al. Overview of the TAC 2010 knowledge base population track[C]// Third Text Analysis Conference (TAC 2010). Vol. 3. 2: 2010: 3-3.
- [104] CANO A E, RIZZO G, VARGA A, et al. Microposts2014 neel challenge: Measuring the performance of entity linking systems in social streams[J]. Proc. of the Microposts2014 NEEL Challenge, 2014.
- [105] CARMEL D, CHANG M W, GABRILOVICH E, et al. ERD'14: entity recognition and disambiguation challenge[C]// ACM SIGIR Forum. Vol. 48. 2. ACM: 2014: 63-77.
- [106] MCNAMEE P, MAYFIELD J, LAWRIE D, et al. Cross-Language Entity Linking.[C]// IJCNLP: 2011: 255-263.

- [107] TSAI C T, ROTH D. Cross-lingual wikification using multilingual embeddings[C]// Proceedings of NAACL-HLT: 2016: 589-598.
- [108] SIL A, KUNDU G, FLORIAN R, et al. Neural cross-lingual entity linking[J]. ArXiv preprint arXiv:1712.01813, 2017.
- [109] ZHANG T, LIU K, ZHAO J, et al. Cross Lingual Entity Linking with Bilingual Topic Model.[C]// IJCAI: 2013.
- [110] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
- [111] IBRAHIM Y, RIEDEWALD M, WEIKUM G. Making Sense of Entities and Quantities in Web Tables[C]// Proceedings of the 25th ACM International Conference on Conference on Information and Knowledge Management. ACM: 2016: 1703-1712.
- [112] EBERIUS J, BRAUNSCHWEIG K, HENTSCH M, et al. Building the dresden web table corpus: A classification approach[C]// Big Data Computing (BDC), 2015 IEEE/ACM 2nd International Symposium on. IEEE: 2015: 41-50.
- [113] NISHIDA K, SADAMITSU K, HIGASHINAKA R, et al. Understanding the Semantic Structures of Tables with a Hybrid Deep Neural Network Architecture.[C]// AAAI: 2017: 168-174.
- [114] BURGES C J. From ranknet to lambdarank to lambdamart: An overview[J]. Learning, 2010, 11(23-581): 81.
- [115] KINGMA D, BA J. Adam: A method for stochastic optimization[J]. ArXiv preprint arXiv:1412.6980, 2014.
- [116] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting.[J]. Journal of machine learning research, 2014, 15(1): 1929-1958.
- [117] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: A Core of Semantic Knowledge[C]// 16th international World Wide Web conference (WWW 2007). Banff, Canada: ACM Press, 2007.
- [118] WU W, LI H, WANG H, et al. Probase: a probabilistic taxonomy for text understanding[C]// SIGMOD Conference: 2012: 481-492.
- [119] RESNIK P. Selectional constraints: An information-theoretic model and its computational realization[J]. Cognition, 1996, 61(1): 127-159.

- [120] ERK K. A simple, similarity-based model for selectional preferences[C]// ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: 2007: 216.
- [121] RITTER A, ETZIONI O, et al. A latent dirichlet allocation method for selectional preferences[C]// ACL: 2010: 424-434.
- [122] BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.
- [123] CHANG A X, MANNING C D. SUTime: A library for recognizing and normalizing time expressions.[C]// LREC: 2012: 3735-3740.
- [124] KLEIN D, MANNING C D. Accurate unlexicalized parsing[C]// ACL: 2003: 423-430.
- [125] LIU T Y. Learning to rank for information retrieval[J]. Foundations and Trends in Information Retrieval, 2009, 3(3): 225-331.
- [126] CHURCH K W, HANKS P. Word association norms, mutual information, and lexicography[J]. Computational linguistics, 1990, 16(1): 22-29.
- [127] LAO N, COHEN W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine Learning, 2010, 81(1): 53-67.
- [128] TOUTANOVA K, CHEN D, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: 2015: 1499-1509.
- [129] WANG Z, LI J. Text-enhanced representation learning for knowledge graph[C]// AAAI: 2016: 1293-1299.
- [130] PUJARA J, MIAO H, GETOOR L, et al. Large-scale knowledge graph identification using psl[C]// AAAI Fall Symposium on Semantics for Big Data: 2013.
- [131] VÖLKER J, NIEPERT M. Statistical schema induction[C]// Extended Semantic Web Conference. Springer: 2011: 124-138.
- [132] GALÁRRAGA L, TEFLIOUDI C, HOSE K, et al. Fast rule mining in ontological knowledge bases with AMIE+[J]. The VLDB Journal, 2015, 24(6): 707-730.
- [133] GUO S, WANG Q, WANG L, et al. Jointly embedding knowledge graphs and logical rules[C]// EMNLP: 2016: 1488-1498.

- [134] CHANG K W, YIH S W T, YANG B, et al. Typed tensor decomposition of knowledge bases for relation extraction[C]// EMNLP: 2014: 1568-1579.
- [135] WANG Q, WANG B, GUO L. Knowledge Base Completion Using Embeddings and Rules.[C]// IJCAI: 2015: 1859-1866.
- [136] ZOU L, HUANG R, WANG H, et al. Natural language question answering over RDF: a graph data driven approach[C]// SIGMOD. ACM: 2014: 313-324.
- [137] RICHARDSON M, DOMINGOS P. Markov logic networks[J]. Machine learning, 2006, 62(1-2): 107-136.
- [138] TIELEMANS T, HINTON G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2).
- [139] XU K, REDDY S, FENG Y, et al. Question Answering on Freebase via Relation Extraction and Textual Evidence[C]// ACL: 2016: 2326-2336.
- [140] CHO K, VAN MERRIËNBOER B, BAHdanaud D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. Proceedings of SSST-8, 2014: 103-111.
- [141] JAIN S. Question answering over knowledge base using factual memory networks[C]// Proceedings of the NAACL Student Research Workshop: 2016: 109-115.
- [142] GABRILOVICH E, RINGGAARD M, SUBRAMANYA A. FACC1: Freebase annotation of ClueWeb corpora, Version 1. <http://lemurproject.org/clueweb12/>. 2013.
- [143] BERANT J, LIANG P. Imitation learning of agenda-based semantic parsers[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 545-558.
- [144] ABUJABAL A, YAHYA M, RIEDEWALD M, et al. Automated template generation for question answering over knowledge graphs[C]// WWW: 2017: 1191-1200.
- [145] TALMOR A, BERANT J. The Web as a Knowledge-base for Answering Complex Questions[C]// NAACL-HLT: 2018.

## 致 谢

时光荏苒，在交大已经度过了六年的时光。博士生的求学之路，应该是人生第一段称得上艰苦的旅行，一路走来经历了很多，要感谢很多人，让我能一步步走到今天。

衷心感谢我的导师朱其立教授。您对科研的不懈追求以及对学术的严谨态度，一点点感染着我，使我不满足于追求表面的结果，而是去多问自己为什么，多去探寻问题的本质。在和您的无数次讨论交流中，经常有学术上的争论，但我总能无压力地讨论自己的想法。您对问题的敏锐嗅觉和对每一位学生的充分尊重，让我深深敬佩。您经常工作到午夜，耐心地为我们的学术论文进行修改，还有一次次的 Noodle time，大家一起愉快而又充满干劲地向论文提交发起冲刺。六年来自己的科研进步离不开您的悉心指导，尤其在基础薄弱的前几年始终对我充满信心。在科研之余，您经常和我们一起运动，组织一年一度甚至两度的实验室集体出游，邀请无法回家的同学来您家中，共度中秋、元旦和除夕。为师者常有，亦师亦友却可遇不可求，衷心感谢您对我在科研上的指导和生活中的关心，这是我此生莫大的荣幸。

特别感谢同一个课题小组的三位学弟，骆徐圣、陈显扬和林封利。很庆幸几年来能和你们一起努力，每一次通宵赶论文有你们的陪伴，每一篇发表成果都离不开你们的倾力付出。骆徐圣和我合作三年，不仅科研一起奋战，生活中也是非常好的朋友，在各方面给了我非常多的鼓励，也与我分享对人生、情感的感悟，相信这些都会对我一生受用。

感谢实验室的三位师兄，赵凯祺、蔡智源、王拯，不仅带领我走进科研的殿堂，而且让我更快融入实验室大家庭，至今依旧怀念当年 ABG 的美好时光。感谢其他几位博士生，罗志一、刘乙竹、黄姗姗，在实验室朝夕相处，怀揣共同的理想，面临相似的困难，平日互相支持与鼓励，和彼此都有着深厚的友谊。感谢姜凯、孙伟、赵天宛、方文静、龚禹、徐栋、沙雨辰、梁玉鼎、唐洋洋、许方正、林禹臣、章梦雪、黄圣蕾、张海军、贾琪以及更多的 ADAPTERs，和大家相处非常愉快，从大家身上学到了许多。

感谢我的室友李冉，不仅有着共同的爱好与话题，平日也经常畅聊对未来生活的规划与思考，让我对科研之外的生活有了更多的理解。感谢王珏、孔奎权、王小乐等光彪楼演唱厅的小伙伴们，很欣慰能被你们亲切称呼一声“大师兄”，你们的存在让我的博士生活不再单调。

最后，特别感谢我的父母和所有家人，在博士的求学之路中，是你们给予我无条件的包容和支持，让我无需担心生活上的种种压力。在我经历科研挫折、一度失去自信的时候，总能在家中给我鼓励，让我能放下包袱，继续奋力前行。



## 攻读学位期间发表的学术论文

- [1] LUO K, LUO X, ZHU K. Inferring Binary Relation Schemas for Open Information Extraction[C]// EMNLP: 2015: 555-560.
- [2] LUO K, LUO X, CHEN X, et al. A Data-driven Approach to Infer Knowledge Base Representation for Natural Language Relations[C]// IJCAI: 2017:1174-1180.
- [3] LUO X, LUO K, CHEN X, et al. Cross-lingual Entity Linking for Web Tables[C]// AAAI: 2018: 362-369.
- [4] LUO K, LIN F, LUO X, et al. Knowledge Base Question Answering via Encoding of Complex Query Graphs[C]// EMNLP: 2018: 2185-2194.
- [5] LUO K, LU J, ZHU K, et al. Layout-aware Information Extraction from Semi-structured Medical Images[J]. Computers in Biology and Medicine, 2019. (已录用)



## 攻读学位期间参与的项目

- [1] 参与自然科学基金中韩合作交流基金项目(2014年–2015年)“多语言、跨文化语义联想概念网络的研究”
- [2] 参与自然科学基金面上项目(2014年–2017年)“基于动作概念的本体知识库及在文本处理上的应用”