软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2024,35(9):4324−4345 [doi: 10.13328/j.cnki.jos.006951]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于胶囊异构图注意力网络的中文表格型数据事实验证*

杨 鹏[1,2], 查显宇[1,2], 赵广振[1,2], 林 茜[3]

[1](东南大学 计算机科学与工程学院, 江苏 南京 211189)

[2](计算机网络和信息集成教育部重点实验室 (东南大学), 江苏 南京 211189)

[3](福州大学 计算机与大数据学院, 福建 福州 350108)

通信作者: 杨鹏, E-mail: pengyang@seu.edu.cn

**摘 要**: 事实验证旨在检查一个文本陈述是否被给定的证据所支持. 由于表格结构上具有依赖性、内容上具有隐含性, 以表格作为证据的事实验证任务仍面临很多挑战. 现有工作或者利用逻辑表达式来解析基于表格证据的陈述, 或者设计表格感知神经网络来编码陈述-表格对, 以此实现基于表格的事实验证任务. 但是, 这些方法没有充分利用陈述背后隐含的表格信息, 从而导致模型的推理性能下降, 并且基于表格证据的中文陈述具有更加复杂的语法和语义, 也给模型推理带来更大的困难. 为此, 提出基于胶囊异构图注意力网络 (CapsHAN) 的中文表格型数据事实验证方法, 所提方法能充分理解陈述的结构和语义, 进而挖掘和利用陈述所隐含的表格信息, 有效提升基于表格的事实验证任务准确性. 具体而言, 首先通过对陈述进行依存句法分析和命名实体识别来构建异构图, 接着对该图采用异构图注意力网络和胶囊图神经网络进行学习和理解, 然后将得到的陈述文本表示与经过编码的表格文本表示进行拼接, 最后完成结果的预测. 更进一步, 针对现有中文表格型事实验证数据集匮乏而难以支持基于表格的事实验证方法性能评价的难题, 首先对主流 TABFACT 和 INFOTABS 表格事实验证英文数据集进行中文转化, 并且专门针对中文表格型数据的特点构建了基于 UCL 国家标准的数据集 UCLDS, 该数据集将维基百科信息框作为人工注释的自然语言陈述的证据, 并被标记为蕴含、反驳或中立 3 类. UCLDS 在同时支持单表和多表推理方面比传统 TABFACT 和 INFOTABS 数据集更胜一筹. 在上述 3 个中文基准数据集上的实验结果表明, 所提模型的表现均优于基线模型, 证明该模型在基于中文表格的事实验证任务上的优越性.

**关键词**: 基于表格的事实验证; 异构图注意力网络; 胶囊图神经网络; 依存句法分析; 命名实体识别

**中图法分类号**: TP18

中文引用格式: 杨鹏, 查显宇, 赵广振, 林茜. 基于胶囊异构图注意力网络的中文表格型数据事实验证. 软件学报, 2024, 35(9): 4324–4345. http://www.jos.org.cn/1000-9825/6951.htm

英文引用格式: Yang P, Zha XY, Zhao GZ, Lin X. Fact Verification with Chinese Tabular Data Based on Capsule Heterogeneous Graph Attention Network. Ruan Jian Xue Bao/Journal of Software, 2024, 35(9): 4324–4345 (in Chinese). http://www.jos.org.cn/1000-9825/6951.htm

## Fact Verification with Chinese Tabular Data Based on Capsule Heterogeneous Graph Attention Network

YANG Peng[1,2], ZHA Xian-Yu[1,2], ZHAO Guang-Zhen[1,2], LIN Xi[3]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

[2](Key Laboratory of Computer Network and Information Integration, Ministry of Education (Southeast University), Nanjing 211189, China)

[3](College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China)

**Abstract**: Fact verification is intended to check whether a textual statement is supported by a given piece of evidence. Due to the structural dependence and implicit content of tables, the task of fact verification with tables as the evidence still faces many challenges.

* 基金项目: 国家自然科学基金 (62272100); 中国工程院院地合作项目 (JS2021ZT05); 中国工程院咨询项目 (2023-XY-09)
收稿时间: 2022-10-24; 修改时间: 2023-03-06; 采用时间: 2023-03-31; jos 在线出版时间: 2023-08-23
CNKI 网络首发时间: 2023-08-28

Existing literature has either used logical expressions to parse statements based on tabular evidence or designed table-aware neural networks to encode statement-table pairs and thereby accomplish table-based fact verification tasks. However, these approaches fail to fully utilize the implicit tabular information behind the statements, which leads to the degraded inference performance of the model. Moreover, Chinese statements based on tabular evidence have more complex syntax and semantics, which also adds to the difficulties in model inference. For this reason, the study proposes a method of fact verification with Chinese tabular data based on the capsule heterogeneous graph attention network (CapsHAN). This method can fully understand the structure and semantics of statements. On this basis, the tabular information implied by the statements is mined and utilized to effectively improve the accuracy of table-based fact verification tasks. Specifically, a heterogeneous graph is constructed by performing syntactic dependency parsing and named entity recognition of statements. Subsequently, the graph is learned and understood by the heterogeneous graph attention network and the capsule graph neural network, and the obtained textual representation of the statements is sliced together with the textual representation of the encoded tables. Finally, the result is predicted. Further, this study also attempts to address the problem that the datasets of fact verification based on Chinese tables are scarce and thus unable to support the performance evaluation of table-based fact verification methods. For this purpose, the study transforms the mainstream English table-based fact verification datasets TABFACT and INFOTABS into Chinese and constructs a dataset that is based on the uniform content label (UCL) national standard and specifically tailored to the characteristics of Chinese tabular data. This dataset, namely, UCLDS, takes Wikipedia infoboxes as evidence of manually annotated natural language statements and labels them into three classes: entailed, contradictory, and neutral. UCLDS outperforms the traditional datasets TABFACT and INFOTABS in supporting both single-table and multi-table inference. The experimental results on the above three Chinese benchmark datasets show that the proposed model outperforms the baseline model invariably, demonstrating its superiority for Chinese table-based fact verification tasks.

**Key words**: table-based fact verification; heterogeneous graph attention network (HAN); capsule graph neural network (CapsGNN); dependency parsing; named entity recognition

随着在线数字内容的激增, 错误信息也随之增加, 当人们面对错综复杂的信息时, 为了辨别信息的真假, 需要对这些信息进行验证[1]. 在自然语言理解 (natural language understanding, NLU) 和语义表征的研究中, 验证一句文本陈述是否符合给定的事实证据是一个基础任务. 现有的工作[2−5]主要集中在对非结构化的文本数据进行事实验证, 它们用到的证据都属于纯文本信息. 由于半结构化和结构化数据在结构上具有依赖性、在内容上具有隐含性, 比如表格、图表、数据库等, 基于这些数据形式的事实验证任务面临很大的挑战.

最近, Chen 等人[6]和 Gupta 等人[7]提出了基于半结构化数据的事实验证数据集来帮助解决这一问题. 现有的工作[6,8−13]通常将基于表格的事实验证任务视为一个自然语言推理 (natural language inference, NLI) 问题, 其中陈述等于假设, 而目标表格等于前提. Chen 等人[6]通过行或列扫描将表格内容转换为连续句, 然后使用 BERT[14]编码陈述-表格进行分类, 从而测试陈述是否支持该表. 而大多数现有的方法[8−10]利用逻辑表达式作为先验信息来表示陈述, 然后通过使用一个图神经网络 (graph neural network, GNN)[15]来学习陈述的隐含关系, 他们试图通过理解陈述的逻辑实现这一任务. 近年来, 人们努力将表格的结构知识纳入到 BERT 风格的神经网络中[11−13], 并引入与表格相关的语料库以进一步理解表格, 他们试图通过理解表格的结构和内容实现这一任务.

尽管以前的工作取得了一定的成果, 但对于基于目标表格验证陈述这一任务仍然存在一些问题. 首先, 虽然像 BERT 这样的预训练语言模型在各种自然语言理解任务上取得了显著的表现, 但当它遇到具有复杂逻辑推理特征的句子时, 如最高级关系、比较关系、聚合关系等, 它的效果往往不尽如人意. 其次, 现有的方法[8−10]提出的逻辑表达式过于庞杂, 利用这些逻辑表达式表示陈述时, 易使其语义复杂化. 这就要求陈述必须包含表格中几乎全部的信息才能进行有效的推理, 而陈述本身大概率会缺失表格中的一些信息, 所以依赖于逻辑表达式实现这一任务是不可靠的. 最后, 先前的工作[11−13]过于注重表格的特征而没有充分利用陈述的特征, 这可能会忽略陈述中所蕴含的潜在证据, 而中文陈述具有更加丰富的语法结构和语义信息, 发现其蕴含的潜在证据至关重要. 为了解决上述问题, 本文提出基于胶囊异构图注意力网络的中文表格型数据事实验证方法, 其核心是胶囊异构图注意力网络框架 (capsule heterogeneous graph attention network, CapsHAN), 它可以充分理解陈述的结构和语义来挖掘陈述所蕴含的潜在证据, 以实现细粒度的推理. 具体而言, 本文首先通过对陈述进行依存句法分析和命名实体识别来构建异构图, 接着对该图采用异构图注意力网络 (heterogeneous graph attention network, HAN)[16]和胶囊图神经网络 (capsule graph neural network, CapsGNN)[17]进行学习和理解, 然后将得到的陈述文本表示与经过编码的表格文本表示进行拼接, 最后完成结果的预测.

数据集是用于检测基于表格的事实验证任务准确性的重要工具. 近年来, 随着 TABFACT 和 INFOTABS 数据集的出现, 基于英文表格的事实验证发展迅速. 中文作为一门历史悠久的语言, 其应用十分广泛并且含有非常丰富的语义信息. 与此同时, 国内的在线内容中存在很多表格形式的信息, 对这些信息进行辨别和验证成为一个不可避免的问题. 因此, 研究基于中文表格的事实验证具有十分重大的意义. 然而, 现有的中文表格型事实验证数据集匮乏, 难以支持基于中文表格的事实验证方法性能的评价. 为此, 本文首先对主流 TABFACT[6]和 INFOTABS[7]表格事实验证英文数据集进行了中文转化, 图 1 给出了 TABFACT 和 INFOTABS 中文数据集的一个实例. 然而这种通过中文转化所得数据集的语言习惯不符合中文的风格, 而且内容不包含国内的热门信息. 基于上述情况, 本文专门针对中文表格型数据的特点构建了基于 UCL 国家标准[18]的数据集 UCLDS, 该数据集将 2 979 个维基百科信息框作为 19 374 个人工注释的自然语言陈述的证据, 并被标记为蕴含、反驳或中立. UCLDS 在同时支持单表和多表推理方面比传统 TABFACT 和 INFOTABS 数据集更胜一筹.

| 国家 | 货币名称 | 通货膨胀率最高的月份 | 最高月通胀率 | 等效的每日通货膨胀率 | 价格翻倍所需的时间 |
|---|---|---|---|---|---|
| 匈牙利 | 匈牙利福林 | 1946 年 7 月 | 4 190 000 | 207.19% | 15 小时 |
| 津巴布韦 | 津巴布韦元 | 2008 年 11 月 | 796 000 000 | 98.01% | 24.7 小时 |
| 南斯拉夫 | 南斯拉夫第纳尔 | 1994 年 1 月 | 31 130 000 | 64.63% | 1.4 天 |
| 塞尔维亚共和国 | 塞尔维亚第纳尔 | 1994 年 1 月 | 2 970 000 | 64.3% | 1.4 天 |
| 德国 | 德国纸标 | 1923 年 10 月 | 29 500 | 20.87% | 3.7 天 |

表格: 恶性通货膨胀

陈述: 匈牙利的等效每日通货膨胀率是所有国家中最高的

标签: 蕴含

(a) TABFACT 数据集的一个实例

| 《在彩虹中》 | |
|---|---|
| 发布 | 2007 年 10 月 10 日 |
| 录制 | 2005 年 2 月5 日–2007 年 6 月 8 日 |
| 表演者 | Radiohead |
| 类型 | 另类摇滚, 流行摇滚, 艺术流行 |
| 长度 | 42:39 |
| 标签 | 自我释放, Xurbia Xendless 发行 |
| 制作人 | 奈杰尔·戈德里奇 |

陈述: 专辑《在彩虹中》时长不超过 40 分钟

标签: 反驳

(b) INFOTABS 数据集的一个实例

图 1 中文表格事实验证任务的实例

本文在 TABFACT、INFOTABS 和 UCLDS 这 3 个基准数据集上进行实验, 实验结果表明, 尽管 3 个数据集各不相同, CapsHAN 框架都实现了显著的性能改进. 本文的贡献可以概括为以下 4 点.

(1) 现有的工作或者理解陈述的方式有缺陷, 或者只理解表格的信息而忽略了陈述. 为此, 本文提出了一种通用的基于表格的事实验证框架 CapsHAN, 它从结构和语义两个角度来深刻理解陈述从而挖掘陈述中隐藏的表格信息, 以此实现细粒度的推理.

(2) 本文提出了一种基于自然语言的异构图构建方法, 通过对自然语句进行依存句法分析和命名实体识别构建异构图, 从而让模型更好的理解自然语句的结构和语义. 消融实验证明了所构建的异构图在理解中文陈述的结构和语义方面的合理性和有效性.

(3) 为解决现有中文表格型事实验证数据集匮乏的难题, 本文将主流数据集 TABFACT 和 INFOTABS 进行了中文转换, 并构建了基于 UCL 国家标准的中文表格事实验证数据集 UCLDS, 它首次提出了多表推理概念, 对半结构化数据的事实验证任务提出了更高的要求.

(4) 实验结果表明, 本文的方法在 TABFACT 的 Test_complex 子集上比最优的基线模型提升了 4.05%, 在 INFOTABS 的 $\alpha_3$Test 子集上比最优的基线模型提升了 3.22%, 在 UCLDS 的 $\alpha_2$Test 子集上比最优的基线模型提升了 4.73%, 证明了该方法的优越性.

本文第 1 节介绍基于表格的事实验证和统一内容标签的研究现状. 第 2 节介绍本文构建的胶囊异构图注意力网络模型. 第 3 节介绍本文所构建的数据集, 包括 TABFACT、INFOTABS 和 UCLDS. 第 4 节通过对比实验验证了所提模型的有效性. 最后总结全文.

## 1 相关工作

### 1.1 基于表格的事实验证

基于表格的事实验证是自然语言理解领域中的一项重要任务. Chen 等人[6]提出了一种用于基准测试的常规

半结构化表格数据集 TABFACT, 并使用 BERT 模型对陈述-表格对进行编码来实现这一任务. 此后, 一系列工作[8−10]联合使用了潜在程序算法 (latent program algorithm, LPA)[6]、图神经网络和 BERT 生成的语句的逻辑表达式来进行基于表格的事实验证, 这是因为逻辑表达式可以为理解语句的语义提供许多先验信息. 例如 Yang 等人[10]使用一个程序选择模块来选择最佳的逻辑表达式, 然后使用一个图注意力网络 (graph attention network, GAT)[19]学习程序树上的推理关系从而理解陈述的语义来挖掘其潜在的证据. 然而传统逻辑表达式驱动的事实验证模型的主要局限是生成的特定逻辑符号会对理解陈述的语义产生干扰, 因此很难从表中找到足够的证据. 最近, 人们研究了表格感知模型[11−13], 这种模型试图通过利用表格的结构特征来取代逻辑表达式. Zhang 等人[11]将表格的结构信息注入自注意力模块中, 同时添加额外的汇总行, 以增强模型的符号推理能力. 另一个重要工作是信息框 (information box, Infobox) 风格的半结构化表格事实验证任务. Gupta 等人[7]提出了此风格的基准数据集 INFOTABS, 它包含多条行记录, 每条记录由一个键-值对组成. 由于信息框风格的表格具有结构性差、信息分散等特点, Neeraja 等人[20]利用 RoBERTa[21]预训练模型引入了和表格相关的其他知识, 同时删除了表格中不相关的信息来解决这一问题. 这些现有的工作大多关注陈述-表格对或者表格本身, 没有充分利用陈述中隐含的表格信息进行推理.

与上述工作不同, 本文提出基于胶囊异构图注意力网络的中文表格型数据事实验证方法, 该方法注重于加强陈述部分的理解. 具体而言, 本文为陈述构建异构图, 并使用 HAN 和 CapsGNN 对陈述的语义和结构进行更加深入的理解.

### 1.2 统一内容标签

目前, 现有的中文表格型事实验证数据集匮乏, 难以支持基于中文表格的事实验证方法性能的评价. 本文对英文数据集进行中文转化得到中文数据集, 然而以这种方式得到的数据集的语言习惯不符合中文的风格, 而且内容不包含国内的热门信息. 文献 [18] 提出了一种描述信息资源的元数据, 统一内容标签 (uniform content label, UCL), 它能够描述内容资源的丰富语义信息, 也可以支持基于语义的内容组织和管理. 因此由 UCL 标引的数据可以为数据集提供丰富的表格信息. UCL 现有工作[22−24]大多利用 UCL 标引特定数据构建 UCL 知识空间或知识图谱, 然后利用该知识空间或知识图谱进行搜索、内容解析或者情感分析等下游工作. 例如, 汪巍[22]利用 UCL 标引热门新闻数据构建 UCL 知识空间, 然后基于 UCL 知识空间开发相应的内容解析原型系统来向用户提供内容解析服务. 然而这些工作都没有充分利用 UCL 知识空间结构的异质性以及内容的丰富性.

与上述工作不同, 本文首次利用 UCL 国家标准和知识空间构建基于 UCL 的中文事实验证数据集, 该数据集提出的多表推理充分利用 UCL 知识空间中节点、边的多样性, 以及 UCL 所标引数据在内容上的丰富性.

## 2 胶囊异构图注意力网络模型

CapsHAN 模型的示意图如图 2 所示, 本节将介绍该模型的整体框架并进行详细说明.

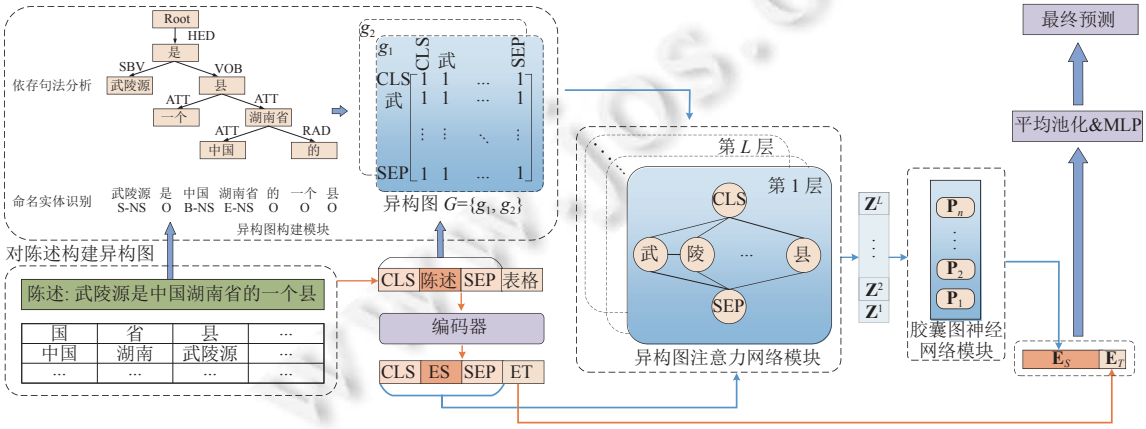



图 2 CapsHAN 模型结构示意图

### 2.1 任务定义

本文把数据集中的每个实例表示成 $(S,T,L)$, 其中, 表格 $T=\{T_k^{(i,j)}|0\leqslant k\leqslant m,0\leqslant i\leqslant R,0\leqslant j\leqslant C\}$, $m$ 代表一个数据集实例中的表格数量, $R$ 和 $C$ 分别为每个表格的行数和列数, $T_k^{(i,j)}$ 表示第 $k$ 个表格中第 $i$ 行第 $j$ 列的单元格内容; 陈述 $S=\{s_w|1\leqslant w\leqslant n\}$, $n$ 代表陈述中的词语数量, $s_w$ 表示陈述中第 $w$ 个词语. $L$ 表示标签, 在不同的数据集中, $L$ 有不同的取值. 在 TABFACT 数据集中, 标签 $L\in\{0,1\}$, $L=1$ 表示陈述 $S$ 被表格 $T$ 支持, $L=0$ 表示陈述 $S$ 被表格 $T$ 反驳. 在 INFOTABS 和 UCLDS 数据集中, 标签 $L\in\{0,1,2\}$, $L=1$ 表示陈述 $S$ 被表格 $T$ 支持, $L=0$ 表示陈述 $S$ 被表格 $T$ 反驳, $L=2$ 表示陈述 $S$ 与表格 $T$ 无关. 本文的目标是给定 $(S,T)$, 预测正确的标签 $L$.

### 2.2 异构图构建模块

异构图是一种存在不同类型节点和边的图, 即节点和边至少有一个具有多种类型, 而连接这些节点的路径被称为元路径. 本文对数据集中的陈述部分构建异构图以加深模型对陈述的理解, 图中节点为陈述经过中文预训练模型分词器得到的词语 (token). 该异构图根据边类型的不同分为依存句法分析图和命名实体连接图.

#### 2.2.1 依存句法分析图

依存句法分析 (dependency parsing, DP) 是自然语言处理中的关键技术之一, 其基本任务是确定句子的句法结构或者句子中词汇间的依存关系[25]. 主要包括两方面的内容, 一是确定语言的语法体系, 即对语言中合法句子的语法结构给予形式化定义; 二是依存句法分析技术, 即根据给定的语法体系, 自动推导出句子的句法结构, 分析句子所包含的句法单位以及这些句法单位之间的依存关系[26]. 依存句法分析树 (称 DP 树) 则将句法单位之间的依存关系以树的形式表示. 本文的依存句法分析采用哈尔滨工业大学语言技术平台 (language technology platform, LTP)[27]进行, LTP 共定义了 14 种依存关系, 如表 1 所示.

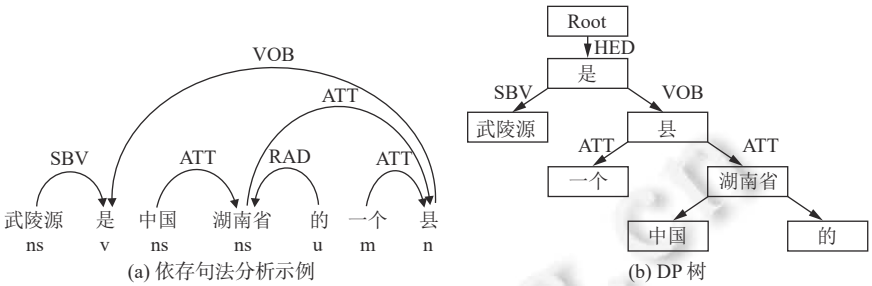
表 1 LTP 中依存关系名、含义及对应边的权值

| 标记 | 解释 | 边的权值 | 标记 | 解释 | 边的权值 |
|---|---|---|---|---|---|
| HED | 核心关系 | 2 | ADV | 状中结构 | 9 |
| SBV | 主谓关系 | 3 | CMP | 动补关系 | 10 |
| VOB | 动宾关系 | 4 | COO | 并列关系 | 11 |
| IOB | 间宾关系 | 5 | POB | 介宾关系 | 12 |
| FOB | 前置宾语 | 6 | LAD | 左附加关系 | 13 |
| DBL | 兼语 | 7 | RAD | 右附加关系 | 14 |
| ATT | 定中关系 | 8 | IS | 独立结构 | 15 |

语句 $S1$ “武陵源是中国湖南省的一个县”, 其依存句法分析结果如图 3(a) 所示, DP 树如图 3(b) 所示. 在图 3(a) 中, ns, v, u, m, n 分别代表地点名词、动词、助词、数量词和名词; 在图 3(b) 中, “是”与父节点 Root 关系为 HED, 是本语句核心词, 每一个节点代表一个句法单位, 节点之间的边代表句法依存关系.

本文中的分词器采用哈工大讯飞联合实验室 (Joint Laboratory of HIT and iFLYTEK Research, HFL) 提出的中文预训练模型 RoBERTa-wwm-ext-large-Chinese[28]带有的分词器, 语句 $S1$ 的分词结果为 $S1$ =[‘CLS’, ‘武’, ‘陵’, ‘源’, ‘是’, ‘中’, ‘国’, ‘湖’, ‘南’, ‘省’, ‘的’, ‘一’, ‘个’, ‘县’, ‘SEP’]. 其中, CLS 字符位于句子的开头, 表示整个句子的含义, SEP 字符位于句子的末尾作为与后续表格内容的分隔符. 分词后的每个词语都是依存句法分析图中的一个节点, 如果依存句法分析中的两个句法单位之间存在着某种依存关系, 则这两个句法单位经过分词后的所有词语两两之间会连上带有权值的边. 需要注意的是, CapsHAN 模型通过区分不同类型的依存句法关系来对陈述的结构进行更细粒度的理解, 本文按照顺序对所有的依存关系进行编号作为边的权值, 这些不同的权值仅用来区分不同的依存句法关系, 而并不代表重要性、频率等属性的具体含义, 其对应关系如表 2 所示. 同一个句法单位内部如果存在分词, 那么该句法单位经过分词后的词语之间会连上权值为 1 的边. 在语句 $S1$ 中, 句法单位“中国”和“湖南省”存在定中关系, 两者分词后的结果分别为 [‘中’, ‘国’] 和 [‘湖’, ‘南’, ‘省’], 由表 2 可知, 定中关系边的权值为 8, 那么‘中’与‘湖’会连上权值为 8 的边, 在邻接矩阵中表现为 $g_1$ [‘中’][‘湖’]=8, 由于本 DP 图为无向图, 其对称位置 $g_1$

['湖']['中']=8. 而'中'与'国'由于是属于同一个句法单位内部的分词, 依据规则会连上权值为 1 的边, 即 $g_1$['中']['国'] = $g_1$['国']['中']=1. 将 $S1$ 以此方法建立的依存句法分析图对应的邻接矩阵如图 4(a) 所示. 构建该依存句法分析图的目的是让模型从句法依存关系的角度更好地理解陈述部分的结构, 从而达到更好的推理效果, 相应实验效果对比将在第 4.4 节展示.




(a) 依存句法分析示例 (b) DP 树

图 3 依存句法分析与 DP 树

表 2 LTP 中命名实体标注及其释义

| 标记 | 含义 |
|---|---|
| O | 这个词不是命名实体 |
| S | 这个词单独构成一个命名实体 |
| B | 这个词为一个命名实体的开始 |
| I | 这个词为一个命名实体的中间 |
| E | 这个词为一个命名实体的结尾 |

| | CLS | 武 | 陵 | 源 | 是 | 中 | 国 | 湖 | 南 | 省 | 的 | 一 | 个 | 县 | SEP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLS | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 武 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 陵 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 源 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 是 | 1 | 3 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 |
| 中 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 1 |
| 国 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 1 |
| 湖 | 1 | 0 | 0 | 0 | 0 | 8 | 8 | 1 | 1 | 1 | 14 | 0 | 0 | 8 | 1 |
| 南 | 1 | 0 | 0 | 0 | 0 | 8 | 8 | 1 | 1 | 1 | 14 | 0 | 0 | 8 | 1 |
| 省 | 1 | 0 | 0 | 0 | 0 | 8 | 8 | 1 | 1 | 1 | 14 | 0 | 0 | 8 | 1 |
| 的 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 14 | 14 | 1 | 0 | 0 | 0 | 1 |
| 一 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8 | 1 |
| 个 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8 | 1 |
| 县 | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 8 | 8 | 8 | 0 | 8 | 8 | 1 | 1 |
| SEP | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

(a) 依存句法分析图

| | CLS | 武 | 陵 | 源 | 是 | 中 | 国 | 湖 | 南 | 省 | 的 | 一 | 个 | 县 | SEP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLS | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 武 | 1 | 1 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 陵 | 1 | 4 | 1 | 4 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 源 | 1 | 4 | 4 | 1 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 是 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 中 | 1 | 4 | 4 | 4 | 0 | 1 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 国 | 1 | 4 | 4 | 4 | 0 | 4 | 1 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 湖 | 1 | 4 | 4 | 4 | 0 | 4 | 4 | 1 | 4 | 4 | 0 | 0 | 0 | 0 | 1 |
| 南 | 1 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 1 | 4 | 0 | 0 | 0 | 0 | 1 |
| 省 | 1 | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 1 |
| 的 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 一 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 个 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 县 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| SEP | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

(b) 命名实体连接图

图 4 语句 $S1$ 所构建的异构图

### 2.2.2 命名实体连接图

命名实体 (named entity, NE) 是人名、机构名、地名以及其他所有以名称为标识的实体, 在自然语言中传递着关键信息, 在一个语句中具有重要的地位, 是信息处理的重点与难点[29]. 命名实体识别 (named entity recognition, NER) 是指识别文本中的命名实体, 是信息提取、问答系统、句法分析、机器翻译、面向 semantic Web 的元数据标注等应用领域的重要基础工具, 在自然语言处理技术走向实用化的过程中占有重要地位. 本文的命名实体识别

同样采用 LTP 进行, LTP 共定义了 3 种命名实体, 分别为人名 (Nh)、机构名 (Ni) 和地名 (Ns). 在 LTP 中, NER 模块的标注结果采用 O-S-B-I-E 标注形式, 含义如表 2 所示.

语句 $S1$ 的命名实体识别结果如图 5 所示. 其中, “武陵源”自身单独构成了一个命名实体, 而“中国”和“湖南省”两个词语一起构成了一个命名实体. 同样地, 经过分词器分词后的各个词语都是命名实体连接图中的一个节点, 与依存句法分析图类似, 不同命名实体之间以及同一个命名实体内部经过分词后的词语两两之间会连上相应权值的边. CapsHAN 模型同样能够依据不同类型的命名实体来对陈述的语义进行更细粒度的理解, 所以规定不同的命名实体所连边的权值不同, 此处不再赘述. 根据边两端节点所属的命名实体类型不同可将边的权值分为 6 种, 其具体的对应关系如表 3 所示.

| 武陵源 | 是 | 中国 | 湖南省 | 的 | 一个 | 县 |
|---|---|---|---|---|---|---|
| S-Ns | O | B-Ns | E-Ns | O | O | O |

图 5 命名实体识别示例

表 3 命名实体对及对应边的权值

| 命名实体对 | 边的权值 |
|---|---|
| Nh-Nh | 2 |
| Ni-Ni | 3 |
| Ns-Ns | 4 |
| Nh-Ni | 5 |
| Nh-Ns | 6 |
| Ni-Ns | 7 |

以 $S1$ 为例, “武陵源”和“中国湖南省”都是地名类实体, 二者分词结果为 [‘武’, ‘陵’, ‘源’] 和 [‘中’, ‘国’, ‘湖’, ‘南’, ‘省’], 地点实体之间所连边的权值是 4, 那么‘武’和‘中’之间将连上权值为 4 的边, 邻接矩阵表现为 $g_2$ [‘武’][‘中’] = $g_2$ [‘中’][‘武’]=4, 而同一个命名实体内部分词如‘武’和‘陵’将连上权值为 1 的边, 即 $g_2$ [‘武’][‘陵’] = $g_2$ [‘陵’][‘武’]=1. 将 $S1$ 以此方法建立的命名实体连接图对应的邻接矩阵如图 4(b) 所示. 由于识别命名实体对于理解自然语句至关重要, 所以构建命名实体连接图是为了让模型从命名实体的角度更好地理解陈述部分的语义, 从而达到更好的推理效果, 相应实验效果对比将在第 4.4 节展示.

#### 2.2.3 构建基于陈述的异构图

本文针对陈述 $S$ 所构建的异构图 $G$ 由相同类型的节点和不同类型的边构成, 边可分为依存句法分析类型的边以及命名实体连接类型的边, 即相邻节点之间存在两种元路径. 如图 4 所示, “国”节点 $N1$ 到“省”节点 $N2$ 可通过 $N1$ —ATT— $N2$ 路径到达, 也可以通过 $N1$ —Ns-Ns— $N2$ 路径到达. 具体过程如下:

$$G = \{g_1, g_2\} = \text{Graph}(DP(S), NER(S)) \tag{1}$$

其中, $g_1$ 和 $g_2$ 分别表示依存句法分析图和命名实体连接图, Graph 表示构建异构图的操作, $DP$ 和 $NER$ 分别表示依存句法分析和命名实体识别操作. 通过理解该异构图, 模型可以从句法依存关系和命名实体的角度全面地理解陈述的结构和语义, 从而提升模型的推理能力, 达到更好的推理效果, 相应实验效果对比将在第 4.4 节展示.

### 2.3 编码模块

产生合适的文本表示对于自然语言处理 (natural language processing, NLP) 任务至关重要. 本文采用 HFL 的 RoBERTa-wwm-ext-large-Chinese 模型作为编码模块. 首先将表格 $T$ 按水平方向展平并拼接起来得到 $Seq_T$ , 接着将 $Seq_T$ 送入编码模块后得到表格的文本表示 $\mathbf{E}_T$ , 然后将陈述 $S$ 送入编码模块得到文本表示 $\mathbf{E}_S = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n] \in \mathbb{R}^{n \times d}$. 其中, $d$ 表示每个词语的特征维数, $n$ 表示陈述中的词语数量, $\mathbf{h}_i$ 表示第 $i$ 个词语的文本表示, 同时也将作为 HAN 中第 $i$ 个词语节点的特征表示.

### 2.4 异构图注意力网络模块

本模块依据异构图注意力网络构建, 目的是得到陈述 $S$ 更深层次的文本表示. 该模块遵循层次注意力结构: 节点级别注意力 (node attention)→ 语义级别注意力 (semantic attention). 该模块的总体框架如图 6 所示. 首先, 节点级别注意力学习每个节点基于元路径邻居的重要性, 并将它们聚合以获得语义级别的节点嵌入; 然后, 语义级别注意力判断每个元路径的差异, 并为特定的任务获得特定语义节点嵌入的最佳加权组合.

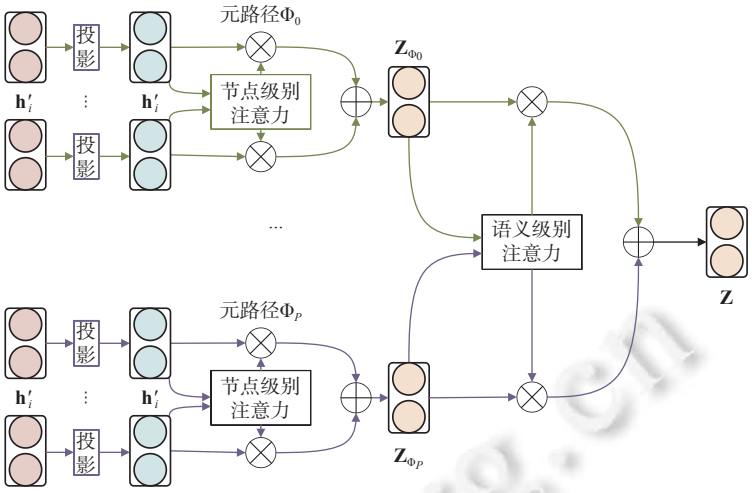



图 6 异构图注意力网络结构示意图

2.4.1 节点级别注意力机制

节点级别注意力是一种聚合异构图中节点邻居信息以形成新的节点表示的一种注意力机制, 它可以学习异构图中每个节点基于元路径邻居的重要性, 并聚合这些有意义的邻居表示以形成节点嵌入. 在异构图中, 由于节点的异质性, 不同类型的节点具有不同的特征空间. 因此, 对于不同类型的节点 (例如具有 $\phi_i$ 类型的节点), HAN 设计了一种特定类型的转换矩阵 $\mathbf{M}_{\phi_i}$, 用来将不同类型的节点特征投影到相同的特征空间中. 投影过程如下:

$$\mathbf{h}'_i = \mathbf{M}_{\phi_i} \cdot h_i \tag{2}$$

其中, $\mathbf{h}'_i$ 和 $\mathbf{h}_i$ 分别是节点 $i$ 的投影特征和原始特征. 通过特定类型的投影操作, 节点级别注意力可以处理任意类型的节点. 之后, 模块利用自注意力机制 (self-attention) 来学习各种节点之间的权重. 给定一个通过元路径 $\Phi$ 连接的节点对 $(i, j)$, 节点级别注意力 $e_{ij}^{\Phi}$ 表示对于节点 $i$ 来说节点 $j$ 的重要性. 具体操作如下:

$$e_{ij}^{\Phi} = att_{\text{node}}(\mathbf{h}'_i, \mathbf{h}'_j; \Phi) \tag{3}$$

其中, $att_{\text{node}}$ 表示的是计算节点级注意力的神经网络. 在获得基于元路径的节点对之间的重要性之后, 模块通过 *Softmax* 函数对它们进行归一化以得到相应的注意力权重系数 $\alpha_{ij}^{\Phi}$:

$$\alpha_{ij}^{\Phi} = Softmax(e_{ij}^{\Phi}) = \frac{\exp(\sigma(\mathbf{a}_{\Phi}^{\mathrm{T}} \cdot [\mathbf{h}'_i \| \mathbf{h}'_j]))}{\sum_{k \in N_i^{\Phi}} \exp(\sigma(\mathbf{a}_{\Phi}^{\mathrm{T}} \cdot [\mathbf{h}'_i \| \mathbf{h}'_k]))} \tag{4}$$

其中, $\sigma$ 表示激活函数, $\|$表示拼接操作, $N_i^{\Phi}$ 表示节点 $i$ 的所有邻居节点, 同时 $\mathbf{a}_{\Phi}$ 表示基于元路径 $\Phi$ 的节点级注意力向量. 随后, 模块将邻居结点的投影特征根据注意力权重系数进行聚合, 从而得到节点 $i$ 基于元路径 $\Phi$ 的嵌入 $\mathbf{z}_i^{\Phi}$, 具体操作如下:

$$\mathbf{z}_i^{\Phi} = \sigma\left(\sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}'_j\right) \tag{5}$$

此后, 模块引入多头注意力机制. 具体来说, 重复应用节点级别注意力 $K$ 次后将学习到的嵌入进行拼接, 得到特定语义的节点嵌入, 具体操作如下:

$$\mathbf{z}_i^{\Phi} = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}'_j\right) \tag{6}$$

给定元路径集 $\{\Phi_0, \Phi_1, \ldots, \Phi_P\}$, 将节点级注意力机制应用于节点特征之后, 可以得到 $P$ 组特定语义的节点嵌入 $\{\mathbf{Z}_{\Phi_0}, \mathbf{Z}_{\Phi_1}, \ldots, \mathbf{Z}_{\Phi_P}\}$.

2.4.2 语义级别注意力机制

为了解决异构图中元路径选择和语义融合的挑战, HAN 提出了一种语义级别注意力机制, 这种注意力机制可以自动学习不同元路径的重要性并将它们融合到特定任务中. 将从节点级别注意力中学习到的特定语义节点嵌入组合作为输入, 每个元路径 $\{\beta_{\Phi_0},\beta_{\Phi_1},\ldots,\beta_{\Phi_P}\}$ 的学习权重可以通过如下操作得到:

$$\{\beta_{\Phi_0},\beta_{\Phi_1},\ldots,\beta_{\Phi_P}\}=att_{\text{sem}}(\{\mathbf{Z}_{\Phi_0},\mathbf{Z}_{\Phi_1},\ldots,\mathbf{Z}_{\Phi_P}\}) \tag{7}$$

其中, $att_{\text{sem}}$ 表示的是计算语义级别注意力的神经网络.

为了学习每条元路径的重要性, 该模块首先通过非线性变换来转换特定语义的节点嵌入, 然后将转换后的嵌入与语义级别注意力向量 $\mathbf{q}$ 求相似度, 以此来衡量该嵌入的重要性. 此外, 该模块平均了每条元路径内所有特定语义节点嵌入的重要性来作为该元路径的重要性. 综上, 每条元路径的重要性 $w_{\Phi_i}$ 计算如下:

$$w_{\Phi_i}=\frac{1}{|v|}\sum_{i\in v}\mathbf{q}^{\mathrm{T}}\cdot\tanh(\mathbf{W}\cdot\mathbf{z}_i^{\Phi}+\mathbf{b}) \tag{8}$$

其中, $\mathbf{W}$ 表示权重矩阵, $\mathbf{b}$ 表示偏置向量, $v$ 表示元路径 $\Phi_i$ 上的所有节点, tanh 表示激活函数. 在获得每条元路径的重要性之后, 模块通过 *Softmax* 函数对它们进行归一化以得到相应的注意力权重系数 $\beta_{\Phi_i}$:

$$\beta_{\Phi_i}=\frac{\exp(w_{\Phi_i})}{\sum_{i=1}^{P}\exp(w_{\Phi_i})} \tag{9}$$

利用学习到的权重系数, 该模块可以融合这些特定语义的嵌入以获得最终的嵌入 $\mathbf{Z}$ 如下:

$$\mathbf{Z}=\sum_{i=1}^{P}\beta_{\Phi_i}\cdot\mathbf{Z}_{\Phi_i} \tag{10}$$

在本文中, 该模块的最终输出 $\mathbf{Z}\in\mathbb{R}^{n\times d}$ 表示的是陈述部分 $S$ 聚合了所有词语节点信息后得到的更深层次嵌入.

2.4.3 模块的输入与输出

综上所述, 陈述 $S$ 经过预训练语言模型后的文本表示 $\mathbf{E}_S=[\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n]$ 输入到本模块中作为异构图的节点特征信息, 同时 $G=\{g_1,g_2\}$ 也输入到本模块中作为异构图的结构信息. 本模块利用 $G$ 提供的结构信息产生 $P$ 条不同的元路径, 然后对节点的特征表示先后应用节点级别注意力和语义级别注意力得到最终的输出 $\mathbf{Z}$. 向量 $\mathbf{Z}$ 是陈述 $S$ 更深层次的文本表示, 同时也表示异构图中所有词语节点特征的拼接, 具体过程如下:

$$\mathbf{Z}=[\mathbf{Z}_{h_1},\mathbf{Z}_{h_2},\ldots,\mathbf{Z}_{h_n}]=\text{HAN}([\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n];\{g_1,g_2\}) \tag{11}$$

其中, HAN 表示本异构图注意力网络模块, $\mathbf{Z}_{h_i}$ 表示节点 $i$ 最终的特征表示. 同样地, 在 $L$ 层的异构图注意力网络中, 本模块将所有层的输出进行拼接得到 $\mathbf{X}$ 作为下一模块的输入:

$$\mathbf{X}=\mathop{\Vert}_{l=1}^{L}(\mathbf{Z}^l)=\mathop{\Vert}_{l=1}^{L}(\mathop{\Vert}_{i=1}^{n}\mathbf{Z}_{h_i}^l) \tag{12}$$

### 2.5 胶囊图神经网络模块

本模块通过改进胶囊图神经网络来构建, 目的是使用从 HAN 模块提取的节点特征来生成更高质量的图嵌入, 即陈述 $S$ 更高质量的文本表示. 具体而言, 首先利用 HAN 模块提取的节点特征构建基础节点胶囊, 然后应用动态路由机制生成高级图胶囊以及高级节点胶囊. 后文图 7 展示了本模块的总体框架.

2.5.1 基础节点胶囊

本功能模块将 HAN 模块中所有层输出的拼接作为输入来构建基础节点胶囊 $\mathbf{C}=[[\mathbf{c}_{(1,1)},\ldots,\mathbf{c}_{(1,L)}],\ldots,[\mathbf{c}_{(n,1)},\ldots,\mathbf{c}_{(n,L)}]]\in\mathbb{R}^{n\times L\times d}$. 具体过程如下:

$$\mathbf{c}_{(i,l)}=pri_{\text{caps}}(\mathbf{Z}_{h_i}^l) \tag{13}$$

其中, $\mathbf{c}_{(i,l)}\in\mathbb{R}^d$ 表示节点 $i$ 的第 $l$ 个胶囊的特征, $pri_{\text{caps}}$ 表示用于构建基础节点胶囊的卷积神经网络.

2.5.2 高级图胶囊

获得基础节点胶囊后, 本模块应用动态路由机制生成图胶囊. 输入基础节点胶囊 $\mathbf{C}$, 得到一组图胶囊嵌入 $\mathbf{H}\in\mathbb{R}^{p\times d}$, 其中 $p$ 表示图胶囊的个数, 每个图胶囊从不同方面反映了异构图 $G$ 的属性. 胶囊的长度反映了属性存在

的概率, 角度反映了属性的细节. 在使用节点胶囊生成图胶囊之前, 本模块使用注意力机制对基本节点胶囊进行缩放, 这样做的目的是为了减小 $p$ 对 $n$ 的依赖以生成更高质量的图胶囊. 具体如下:

$$scaled(\mathbf{c}_{(i,l)}) = \frac{F_{\text{attn}}(\mathbf{c}_i^*)}{\sum_{i=1}^{n} F_{\text{attn}}(\mathbf{c}_i^*)} \cdot \mathbf{c}_{(i,l)} \tag{14}$$

其中, $\mathbf{c}_i^* \in \mathbb{R}^{l \cdot d}$ 表示通过拼接节点 $i$ 的所有胶囊得到的向量, $F_{\text{attn}}$ 是用来计算注意力向量的两层全连接网络, $scaled(\mathbf{c}_{(i,l)}) \in \mathbb{R}^d$ 表示缩放后的基础节点胶囊. 在对所有的基础节点胶囊应用注意力机制进行缩放操作之后, 模块对缩放后的节点胶囊应用胶囊网络 (capsule network, CapsNet)[30]中的动态路由机制 (dynamic routing) 得到图胶囊. 此过程如下:

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_p] = Routing(scaled(\mathbf{C})) \tag{15}$$

其中, $\mathbf{H}_i \in \mathbb{R}^d$ 表示第 $i$ 个图胶囊, $Routing$ 表示动态路由机制.

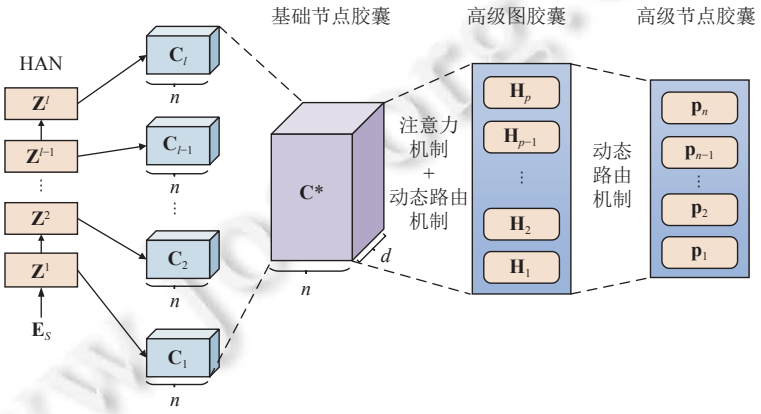



图 7 胶囊图神经网络结构示意图

2.5.3 高级节点胶囊与重构损失

在图胶囊上再次应用动态路由机制生成最终的高级节点胶囊 $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n] \in \mathbb{R}^{n \times d}$, 即本模块输出的陈述 $S$ 最终文本表示, $p_i$ 表示第 $i$ 个词语的文本表示. 该模块沿用了 CapsGNN 中的重建损失来控制所输出文本表示的质量, 此过程如下:

$$Loss_r = Recons_{loss}(\mathbf{Z}^L, \mathbf{P}) \tag{16}$$

其中, $Loss_r$ 表示得到的重构损失, $Recons_{loss}$ 表示计算重构损失的两层全连接网络.

2.5.4 模块的输入与输出

综上所述, 陈述 $S$ 在 HAN 模块中每一层输出的拼接 $\mathbf{X}$ 作为本模块的输入提供节点特征信息, 本模块首先利用这些信息构建基础节点胶囊, 然后对基础节点胶囊先后应用注意力机制和动态路由机制得到图胶囊, 最后再次利用动态路由机制得到高级节点胶囊, 即最终的输出 $\mathbf{P}$. 向量 $\mathbf{P}$ 是陈述 $S$ 更高质量的文本表示, 具体如下:

$$\mathbf{E}'_S = \mathbf{P} = \text{CapsGNN}(\mathbf{X}) \tag{17}$$

其中, CapsGNN 即本胶囊图神经网络模块, $\mathbf{E}'_S$ 表示陈述 $S$ 最终的文本表示.

### 2.6 训练与测试模块

本模块将陈述最终的文本表示和表格的文本表示拼接后进行训练和标签预测. 首先, 模块将 CapsGNN 模块输出的陈述 $S$ 最终文本表示 $\mathbf{E}'_S$ 和表格的文本表示 $\mathbf{E}_T$ 进行拼接得到 $\mathbf{E}'$. 随后, 模块对 $\mathbf{E}'$ 进行平均池化得到 $\mathbf{E}^* \in \mathbb{R}^d$ 作为一个数据集实例的最终表示. 此过程如下:

$$\mathbf{E}^* = Mean_{\text{pooling}}(\mathbf{E}') = Mean_{\text{pooling}}(\mathbf{E}'_S \| \mathbf{E}_T) \tag{18}$$

其中, $Mean_{\text{pooling}}$ 表示平均池化操作. $\mathbf{E}^*$ 经过一层全连接网络得到最终用于分类的向量 $\mathbf{E}_C^* \in \mathbb{R}^c$, $c$ 表示类别数, 在 TABFACT 数据集里 $c$=2, 在 INFOTABS 和 UCLDS 数据集中 $c$=3.

2.6.1 损失函数

CapsHAN 使用的损失函数为交叉熵损失和重构损失的加权和:

$$Loss = \text{CrossEntropy}(\mathbf{E}_C^*, L) + \lambda Loss_r \tag{19}$$

其中, CrossEntropy 表示交叉熵损失, $\lambda$ 为重构损失系数. 本文通过最小化损失函数 $Loss$ 来训练所提的模型.

2.6.2 标签预测

CapsHAN 利用 $\mathbf{E}_C^*$ 经过 *Softmax* 函数进行标签预测:

$$L^* = \text{argmax}(Softmax(\mathbf{E}_C^*)) \tag{20}$$

其中, $L^*$ 表示预测的标签, argmax 表示取最大值索引的操作.

胶囊异构图注意力网络模型的整体过程如算法 1 所示.

**算法 1**. CapsHAN.

输入: 陈述 $S$, 表格 $T$, 标签 $L$;
输出: 损失函数 $Loss$; 预测标签 $L^*$.

1. WHILE 数据集没有被完全采样 DO:
2. 从数据集中采样一个大小为 $N$ 的训练批次 $(S_N, T_N, L_N)$
3. FOR $i = 0 \to N-1$ DO:
4. 对陈述 $S_i$ 构造异构图: $G_i = \text{Graph}(DP(S_i), NER(S_i))$
5. 计算表格 $T_i$ 的文本表示: $\mathbf{E}_{T_i} = Enconder(T_i)$
6. 计算陈述 $S_i$ 的文本表示: $\mathbf{E}_{S_i} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n]_i = Encoder(S_i)$
7. 通过 HAN 模块计算陈述 $S_i$ 的强化特征表示:
$$\mathbf{Z}_i = [\mathbf{Z}_{h_1}, \mathbf{Z}_{h_2}, \ldots, \mathbf{Z}_{h_n}]_i = \text{HAN}(\mathbf{E}_{S_i}, G_i)$$
8. 将 HAN 模块每一层的拼接 $\mathbf{X}_i$ 作为 CapsGNN 模块的输入:
$$\mathbf{X}_i = \mathop{\Vert}_{l=1}^{L}(\mathbf{Z}_i^l) = \mathop{\Vert}_{l=1}^{L}(\mathop{\Vert}_{i=1}^{n}\mathbf{Z}_{h_i}^l)_i$$
9. 通过 CapsGNN 模块计算陈述 $S_i$ 的高质量特征表示: $\mathbf{E}'_{S_i} = \text{CapsGNN}(\mathbf{X}_i)$
10. 对 $\mathbf{E}'_{S_i}$ 与 $\mathbf{Z}_i$ 计算重构损失: $Loss_r = Recons_{loss}(\mathbf{Z}_i, \mathbf{E}'_{S_i})$
11. 将 $\mathbf{E}'_{S_i}$ 与 $\mathbf{E}_{T_i}$ 拼接后进行平均池化得到一个数据集实例的最终表示:
$$\mathbf{E}'_i = Mean_{\text{pooling}}(\mathbf{E}'_{S_i} \| \mathbf{E}_{T_i})$$
12. END FOR
13. 对一个批次中所有实例的表示进行拼接: $\mathbf{E}^* = [\mathbf{E}_0^*, \mathbf{E}_1^*, \ldots, \mathbf{E}_{N-1}^*]$
14. $\mathbf{E}^*$ 经过一层全连接网络后得到 $\mathbf{E}_C^*$
15. 最小化损失函数: $Loss = \text{CrossEntropy}(\mathbf{E}_C^*, L_N) + \lambda Loss_r$
16. 预测标签: $L^* = \text{argmax}(Softmax(\mathbf{E}_C^*))$
17. END WHILE
18. return $Loss$, $L^*$

## 3 中文表格型事实验证数据集构建

数据集是支撑事实验证任务研究的重要工具. 近年来, TABFACT、INFOTABS 等高质量英文事实验证数据集相继出现, 基于英文表格的事实验证发展迅猛. 中文作为世界上为数不多的象形文字, 其历史悠久且含有十分丰

富的语义, 与此同时, 中文也是世界上使用人数最多的语言, 它具有十分广泛的应用. 当今正处于互联网时代, 网络在线内容中存在着很多表格形式的中文信息, 对这些信息进行辨别和验证必不可少. 因此, 研究基于中文表格的事实验证具有十分重大的意义. 然而, 现有的中文表格型事实验证数据集匮乏, 难以支持这一方向的研究. 为此, 本文首先提出了对英文数据集进行中文转化构建数据集, 然后这种通过中文转化得来的数据集存在一些缺陷. 一方面, 数据集陈述部分的语言习惯更符合英文的风格而不符合中文的风格, 这会对中文预训练模型的推理造成障碍; 另一方面, 数据集表格部分的内容包含了国外各种热点事件和任务, 但几乎没有包含国内的热点信息, 这不利于将这些数据集的研究结果应用于国内谣言、假新闻的检测中. 基于上述情况, 本文专门针对中文表格型数据的特点构建了基于 UCL 国家标准和 UCL 知识空间的数据集 UCLDS, 这样做的目的有如下两点: 首先, UCL 具有普遍性、规范性等特点, 它能够描述内容资源的丰富语义信息, 也可以支持基于语义的内容组织和管理, 因此由 UCL 标引的数据可以为数据集提供丰富的表格信息; 其次, UCL 知识空间中每个节点的信息都来自于中文维基百科和各大门户网站的热门新闻, 这保证了数据集内容的实时性和丰富性. 这两种方法具体描述如下.

### 3.1 英文数据集的中文转化

TABFACT 是一个包含 1.6 万个维基百科表格作为事实证据和 11.8 万条人工标注的自然语言陈述的事实验证数据集. 一方面, 该数据集在形式上与纯文本数据不同, 采用的是表格形式的数据. 另一方面, 该数据集与一般的事实验证数据集不同, 要求模型同时具有语义推理和符号推理的能力. 本文所提出的 TABFACT 数据集是基于原版数据集进行中文转换以及数据清洗所得到的中文数据集, 该中文数据集保留了英文数据集的格式和内容, 但是去除了原本数据集中非中英文字符的噪声干扰, 强化了中文预训练模型理解该数据集的能力, 相应的去噪效果对比将在第 4.4 节展示. 图 1(a) 为该数据集的一个实例.

INFOTABS 是一个包含 2 540 个不同的维基百科信息框作为事实证据和 2.3 万条人工标注的自然语言陈述的半结构事实验证数据集, 与 TABFACT 不同的是, INFOTABS 的标签分为“蕴含”“反驳”和“中立”这 3 种. 本文中所提出的 INFOTABS 数据集是基于英文数据集进行中文转换和数据清洗所得到的中文数据集, 该中文数据集保留了原数据集的格式与内容. 图 1(b) 为该数据集的一个实例.

### 3.2 基于 UCL 国家标准构建数据集

#### 3.2.1 UCL 国际标准与知识空间

统一内容定位符 (uniform resource locator, URL) 是互联网中基础性、核心性和通用性的内容知识标准, 内容资源普遍采用 URL 进行标识. 但是, 以面向地址理念设计的 URL, 原理上无法描述内容资源的丰富语义信息, 也难以支持基于语义的内容组织和管理, 由此带来内容资源难找难管、混乱失序和不可信等问题. 为解决上述问题, 文献 [18] 提出了统一内容标签, 该标签从互联网中内容资源难找、难管和失序等问题的根本症结入手, 采取内容驱动的理念对内容标识进行全新设计, 形成生产、消费和管理三位一体的内容大数据创新标识体系.

UCL 一般以数据包的形式存在, 一个 UCL 数据包由具有一定次序的多个域组成, UCL 包可用于描述任意内容, 这种对内容进行的描述也被称为索引或标识. 一般来说, 一个 UCL 包可被分为前后两个部分: UCL 代码部分 (UCL code) 和 UCL 属性部分 (UCL properties). UCL 代码部分一般包含多个 UCL 代码域, UCL 属性部分通常包含多个 UCL 属性域. UCL 数据包可以根据其实际应用进行灵活的裁剪或扩展.

UCL 代码部分的基本长度是 32 字节, 它们被称为基本 UCL 代码. 除了基本 UCL 代码以外, UCL 代码部分在需要时可以进行灵活的扩展, 需要注意的是, 扩展部分的长度应为 16 字节的整数倍, 这被称为扩展 UCL 代码, UCL 代码部分的信息实则蕴含了面向读者的内容导引. UCL 属性部分记录与内容相关的多个属性信息, 每个具体属性称为一个 UCL 属性元素, 每个 UCL 属性元素由一个 UCL 属性元素域定义, 性质或功能相近的若干 UCL 属性元素构成一个 UCL 属性集合, 每个 UCL 属性集合由一个 UCL 属性集合头部域和紧接其后的多个连续存放的 UCL 属性元素域构成[18]. 目前已经定义的两个 UCL 属性集合是: 内容描述属性集合 (CDPS) 和内容管理属性集合 (CGPS). 与 UCL 代码部分相对应, UCL 属性部分包含了语义信息和管理信息, 语义信息体现作者的意图, 关系信息支持内容的依法管理.

由于 UCL 在原理上能够描述内容资源的丰富语义信息, 也可以支持基于语义的内容组织和管理, 越来越多的工作[22−24]开始构建基于 UCL 的知识库. 传统的知识库仅包含实体和实体间的关系, 而用于 UCL 知识发现的知识库不仅有实体及其关系, 还含有 UCL 信息及 UCL 与实体之间的语义关系, 因此, 为了与传统知识库进行区分, 称这种基于实体的海量 UCL 的知识组织和发现的知识库为 UCL 知识空间[22]. UCL 知识空间包含两大要素: UCL 和实体. UCL 是对互联网热门信息的标引, 实体是知识空间中知识的基础. UCL 知识空间通过利用 UCL 与实体的关系实现 UCL 相互之间的语义关联. 为了构建 UCL 知识空间, 汪巍[22]提出了一种基于命名实体的 UCL 知识空间构建方法. 作者将 UCL 知识空间的构建整体划分为 3 层: 信息采集层、信息处理层和信息关联层.

在信息采集层, 作者利用 Scrapy 爬虫框架来对网页进行爬取, 利用从网络上获得的百度百科离线数据中提取到的词条链接爬取百度百科词条, 进而得到较为完整的百度百科词条库. 此外, 从主流门户网站爬取热门新闻得到原始网页文档库.

对于爬取得到的百度百科词条原始数据, 作者将其与离线数据结合并采用信息抽取技术提取词条相关信息. 接下来, 统计分析百科词条中的数据生成实体名称映射词典和关系映射词典, 进而得到实体逻辑关联图谱. 之后, 采用网页信息自动抽取技术提取原始网页文档中的新闻信息关键内容, 然后使用 UCL 标引器对内容进行标引得到 UCL 标签并入库.

在信息关联层, 作者首先利用 HanLP 工具抽取 UCL 中的命名实体, 并通过计算得到实体在 UCL 中的语义权重信息. 接着, 采用基于语境相似的实体消歧技术进行实体链接, 该步骤的目的是将从 UCL 中抽取得到的实体和基于百度百科词条得到的实体进行对齐. 最后将 UCL 入库并建立其与对应实体之间的关联关系, 从而构建出完整的 UCL 知识空间. 以此方法构建的 UCL 知识空间储存模型如图 8 所示.

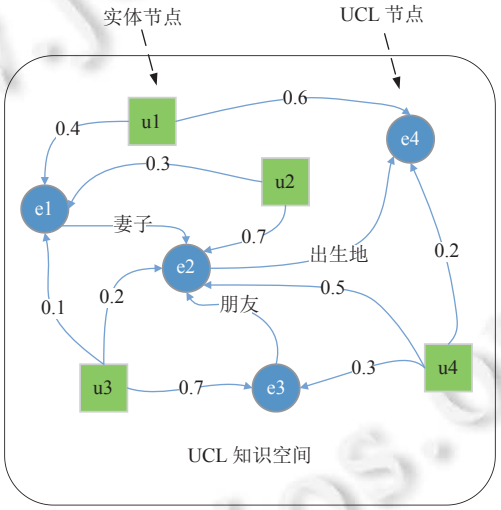



图 8 UCL 知识空间存储模型示意图

### 3.2.2 多表推理

如第 2.1 节所述, 给定一个数据集实例 $(S,T,L)$, $S$ 和 $L$ 分别代表陈述和标签, $T=\{T_1,T_2,\ldots,T_n\}$ 代表作为证据的表格. 当 $|T|>1$, 称这种由 $(S,T)$ 推理得到 $L$ 的过程称为多表推理; $|T|=1$ 时则称之为单表推理. TABFACT 和 INFOTABS 数据集都是以单个维基百科表格或者信息框作为推理的最小单位, 它们仅支持单表推理. 然而, 随着互联网的迅速发展, 很多在线的表格信息中存在外链形式的数据, 这些数据可能和别的表格信息相关联, 而在进行谣言、假新闻的检测时也需要对这些形式的数据进行验证, 这就要求模型具有推理多个相关联表格的能力. 为此, 本文利用 UCL 国家标准和 UCL 知识空间构建了 UCLDS 数据集, 这样做是因为 UCL 知识空间中一个实体节点可能和多个 UCL 节点有关联, 而将这些相关联的 UCL 节点进行有选择地合并可以得到多表形式的数据, 这样数

据集可以在保证表格 UCL 格式的基础上支撑模型进行多个表格的推理. 如图 8 所示, e2 节点与 e1 节点之间存在“夫妻”关系, 与 e3 节点之间存在“朋友”关系, 与 e4 节点之间存在“出生地”的关联. 由此可知, 在多表推理中, 表格与表格之间的关系错综复杂, 进而可以推断出多表推理存在很大的难度.

3.2.3 UCLDS 数据集的构建

在图 8 中, 与 INFOTABS 数据集类似, 本文将 UCL 知识空间中每一个 UCL 节点的内容视作一个由键-值对构成的半结构化表格. 而对于陈述部分, 本文采用人工的方式针对每个实体节点及其连接的 UCL 节点编写对应的自然语句作为陈述. 值得注意的是, 如果一个实体节点与多个 UCL 节点有关联, 本文将选取与其语义相似度最高的 3 个 UCL 节点合并成为一个半结构化表格, 即保证一个实体节点对应一个 UCL 表格. 涉及这类表格的推理即为多表推理. 与之相对应, 如果一个实体节点只和一个 UCL 节点有关联, 那么这种推理就是单表推理. 在多表推理中, 陈述的最小单位依然是合并后的 UCL 表格及其相应的实体节点. 综上, 本文所构建的 UCLDS 是一个同时支持单表推理和多表推理的数据集.

对于 UCLDS 数据集的分割, 本文将其分为训练集、验证集和测试集. 在 UCL 训练集中, 对于每一个 UCL 表格和实体节点, 统一设置 6 条陈述, 在难度上分别设置 3 条简单级别和 3 条困难级别, 对于每个难度级别, 标签均匀分布, 分别是 1 条蕴含, 1 条反驳以及 1 条中立. 陈述的具体推理类别详见附录 A. 验证集与训练集类似, 此处不再赘述. 参考 INFOTABS 数据集, 本文采用多方面评估的方式将测试集分成 3 个, 分别为 $\alpha_1$Test 、 $\alpha_2$Test 和 $\alpha_3$Test .

UCLDS 数据集的常规测试集 ( $\alpha_1$Test ) 中陈述和表格的词汇组成方面与训练集是相似的.

对抗性测试集 ( $\alpha_2$Test ) 与训练集具有相似的表格, 而其陈述会在训练集陈述的基础上进行较少的否定改动而得到, 其相应的标签也随之改动. 具体而言, 对于训练集中某一个 UCL 表格的 6 条陈述, 对 2 条标签为“蕴含”的陈述进行否定改动得到 $\alpha_2$Test 的陈述, 标签改为“反驳”; 相应地, 对 2 条标签为“反驳”的陈述进行否定改动得到 $\alpha_2$Test 的陈述, 标签改为“蕴含”; 而对于标签为“中立”的陈述, 本文将继续进行否定改动, 但标签依然是“中立”. 设置对抗性测试集的目的是为了防止模型只学会陈述表面的规律, 而没有真正学会推理.

复杂测试集 ( $\alpha_3$Test ) 的表格在词汇组成方面与训练集相似, 在陈述的设置上, 本文为每一个表格编写 6 条困难级别的陈述, 标签依然采用均匀分布, 分别是 2 条蕴含, 2 条反驳和 2 条中立. 设置该测试集的目的是为了测试模型在高难度推理中的泛化能力.

UCLDS 数据集共有 2 979 个 UCL 半结构化表格和 19 374 条陈述. 在数据集的分割上, 训练集拥有 2 229 个表格, 13 374 条陈述; 验证集和 3 个测试集分别拥有 250 个表格和 1 500 条陈述. 图 9 为该数据集多表推理的一个实例, 其中由“相关题目”标注的即为合并的 UCL 节点.

| 题目 | 云南日报 |
|---|---|
| 类型 | 日报 |
| 创刊日 | 1950 年 3 月 4 日 |
| 语言 | 简体中文 |
| 总部 | 云南省昆明市西山区日新路 516 号云报传媒广场 |
| 网站 | 云南日报数字报、云南日报网 |
| 相关题目 | 简体中文 |
| 使用日期 | 1956 年至今 |
| 母书写系统 | 汉字、甲骨文、金文、篆书、隶书、繁体字、简体字 |
| 姊妹书写系统 | 日本汉字、朝鲜汉字、契丹文、西夏文、喃字 |

陈述: 《云南日报》的编写语言从建国开始就已在使用
标签: 反驳

图 9 UCLDS 数据集的一个实例

为了控制数据集的质量, 本文对 UCLDS 数据集进行了质量验证, 从验证集和测试集中各抽取 200 个陈述-表格对, 并将每对重新分配给 5 个单独的标注人员进行标签的预测. 此次质量验证采用科恩 Kappa 分数[31]进行标注

者间信度的计算, 其结果如表 4 所示. 从表中可以看出, 验证集和测试集的科恩 Kappa 分数在 0.78–0.83 之间, 这意味着人们对这一任务的意见具有很高的一致性, 同时也反映了该数据集是可信的. 此外, 从表中可以得知多数协议 (5 位标注人员中有 3 位同意) 的范围在 95%–97%, 这同样表示了该数据集是一个经过验证合格的事实验证数据集. 对于所有的分割, 人类准确率在 81%–88%, 其数值处于标注者间信度和多数协议之间, 考虑到任务的难度, 人类准确率处于一个合理的范围之内.

表 4 UCLDS 数据集质量验证统计数据

| UCLDS数据集 | 科恩Kappa分数 | 人类准确率 (%) | 多数协议 (%) |
|---|---|---|---|
| Val | 0.82 | 86.78 | 97.23 |
| $\alpha_1$Test | 0.83 | 88.36 | 98.02 |
| $\alpha_2$Test | 0.78 | 81.04 | 95.33 |
| $\alpha_3$Test | 0.79 | 82.88 | 96.77 |

## 4 实 验

### 4.1 数据集和指标

本文所提出的模型在 3 个表格事实验证基准数据集上验证, 分别为 TABFACT、INFOTABS 和 UCLDS 数据集. 在 TABFACT 和 INFOTABS 数据集中, 一条陈述对应一个表格, 而一个表格对应多条陈述, 而在 UCLDS 数据集中, 一条陈述可能对应多个表格. 3 个数据集中的每条陈述都被人工打上了相应的标签, 在 TABFACT 数据集中, 标签为“蕴含”或“反驳”; 在 INFOTABS 和 UCLDS 数据集中, 标签为“蕴含”“反驳”或“中立”. 对于 TABFACT 数据集, 除了常规的训练集、验证集和测试集, TABFACT 以简单和复杂两种方式提取了 Test_simple 和 Test_complex 子集. 而对于 INFOTABS 数据集, 它从相似分布、敌对分布和训练数据跨域的角度进行了多方面的评估, 所以它将测试集分为了 $\alpha_1$Test 、 $\alpha_2$Test 和 $\alpha_3$Test 这 3 个子集. 而对于 UCLDS 数据集, 它从相似分布、对抗分布和复杂推理的角度进行多方面的验证, 与 INFOTABS 类似, 它也将测试集分为了 $\alpha_1$Test 、 $\alpha_2$Test 和 $\alpha_3$Test 这 3 个子集. 这 3 个数据集的信息见表 5.

表 5 所有数据集的统计数据

| TABFACT | Train | Val | Test | Test_simple | Test_complex |
|---|---|---|---|---|---|
| 陈述 | 92 283 | 12 792 | 12 779 | 50 244 | 68 031 |
| 表格 | 13 182 | 1 696 | 1 695 | 9 189 | 7 392 |
| INFOTABS | Train | Val | $\alpha_1$Test | $\alpha_2$Test | $\alpha_3$Test |
| 陈述 | 16 538 | 1 800 | 1 800 | 1 800 | 1 800 |
| 表格 | 1 740 | 200 | 200 | 200 | 200 |
| UCLDS | Train | Val | $\alpha_1$Test | $\alpha_2$Test | $\alpha_3$Test |
| 陈述 | 13 374 | 1 500 | 1 500 | 1 500 | 1 500 |
| 表格 | 2 229 | 250 | 250 | 250 | 250 |

本文遵循之前的工作使用数据集, 并将准确性作为评估表格事实验证性能的指标. 具体如下:

$$acc = \frac{1}{N}\sum_{i=1}^{N} sum(f(S_i, T_i) = L_i) \tag{21}$$

其中, $N$ 表示样本总数, $sum$ 表示计数求和操作, $f()$ 表示由陈述 $S$ 和表格 $T$ 到标签 $L$ 的映射.

### 4.2 实验设置

本文的实验通过 PyTorch 框架执行, 优化器采用权值衰减为 1E–4、热身率为 0.05 的 AdamW, 在实验中根据验证集上的性能调整超参数. 对于 TABFACT 数据集, 本文使用 16 个注意力头、24 层的 RoBERTa-wwm-ext-

large-Chinese 模型在批大小为 9、初始学习率为 1E–5 的设置上运行了 10 次 epoch; 对于 INFOTABS 和 UCLDS 数据集, 使用相同的设置运行了 20 次 epoch. 实验中, 本文使用 8 个注意力头、2 种元路径、dropout 为 0.6 的 2 层 HAN 模型作为异构图注意力模块; 对于胶囊图神经网络模块, 本文使用图胶囊个数为 8 的 CapsGNN 模型.

### 4.3 基线模型

本次实验所采用的基线模型一共有 4 个, 它们分别代表了处理表格事实验证任务的主要方法. 具体如下.

Table-BERT[6]: 该模型将基于表格的事实验证任务视为 NLI 问题, 通过行或列扫描将表格内容转换为连续句, 然后使用 BERT 编码陈述-表格进行分类, 从而测试陈述是否支持该表.

RoBERTa-CH[28]: 该模型使用 RoBERTa-wwm-ext-large-Chinese 预训练模型编码陈述-表格对进行分类, 从而测试陈述是否支持该表.

ProgVGAT[10]: 该模型沿用了 LPA 的思想, 即利用 LPA 方法为每个陈述生成适当的逻辑表达式从逻辑上理解陈述的语义, 再进行后续的分类. 作者首先利用基于边距损失的程序选择模块生成最优逻辑表达式, 然后利用 GAT 对陈述、表格和逻辑表达式进行编码, 进行最终预测.

SAT[11]: 作者提出了一种基于结构感知的 Transformer 模型[32]. 该模型通过在自注意力层屏蔽部分标记, 从而将表结构特征嵌入到 Transformer 模型中.

### 4.4 结果和分析

#### 4.4.1 性能评估

表 6 展现了 3 个数据集上各种模型的总体表现. 如表 6 所示本文所提出的 CapsHAN 模型基本上优于基线模型且具有显著的余量. 在 TABFACT 数据集上, CapsHAN 比表现最好的基线模型 SAT 高出 0.08%–4.05%; 在 INFOTABS 数据集上, CapsHAN 比表现最好的基线模型 ProgVGAT 高出 1.55%–3.22%; 在 UCLDS 数据集上, CapsHAN 比表现最好的基线模型 ProgVGAT 高出 0.94%–4.73%. 这说明了异构图注意力网络和胶囊图神经网络可以提高模型的表现. 值得注意的是, 本文所提的模型在 TABFACT 数据集上的 Test_complex 子集和 INFOTABS 数据集上的 $\alpha_3$Test 子集上获得了最大的提升, 说明了它具有处理需要复杂逻辑推理的陈述的优势, 同时也说明了所构建的异构图可以帮助模型理解陈述的结构和语义. 而对于 UCLDS 数据集, 模型在 $\alpha_2$Test 子集上效果提升显著, 说明了它在处理对抗性逻辑推理 (否定推理) 方面具有显著的优势, 进一步说明所构建异构图可以加深模型对于陈述的理解.

表 6 主要实验结果 (%)

| 数据集 | 子集 | Table-BERT[6] | RoBERTa-CH[28] | ProgVGAT[10] | SAT[11] | Ours |
|---|---|---|---|---|---|---|
| TABFACT | Val | 57.84 | 66.97 | 66.37 | 69.81 | **69.89** |
| | Test | 57.21 | 66.35 | 66.70 | 69.90 | **72.80** |
| | Test_simple | 60.08 | 75.41 | 75.31 | 80.88 | **81.43** |
| | Test_complex | 55.81 | 62.00 | 62.52 | 64.52 | **68.57** |
| | Test_small | 57.59 | 67.92 | 68.13 | 72.61 | **76.12** |
| INFOTABS | Val | 67.17 | 68.61 | 69.39 | 65.89 | **71.61** |
| | $\alpha_1$Test | 66.50 | 67.28 | 69.17 | 65.50 | **70.72** |
| | $\alpha_2$Test | 53.06 | 56.17 | 58.33 | 52.78 | **61.28** |
| | $\alpha_3$Test | 52.83 | 55.39 | 57.39 | 53.22 | **60.61** |
| UCLDS | Val | 77.47 | 78.13 | 80.73 | 72.80 | **81.67** |
| | $\alpha_1$Test | 75.27 | 76.60 | 79.07 | 72.33 | **81.00** |
| | $\alpha_2$Test | 61.07 | 62.20 | 64.40 | 60.36 | **69.13** |
| | $\alpha_3$Test | 73.27 | 72.73 | 75.27 | 69.28 | **76.27** |

接下来, 本文将深入分析基线模型的表现. 在表 6 中, 对于生成逻辑表达式的方法, ProgVGAT 使用 LPA 方法为陈述生成了很多的逻辑表达式, 然后利用这些逻辑表达式与陈述、表格之间的图结构特征进行验证, 在 3 个数据集上其效果均优于 Table-BERT, 这说明为陈述生成准确的逻辑表达式对于提高验证表现具有一定的效果. 而对

于没有生成逻辑表达式的方法, 在 TABFACT 数据集上, Table-BERT (57.84%) 是第 1 个使用顺序文本来编码表格内容并利用 BERT 进行验证的基线模型. RoBERTa-CH (66.97%) 是在 Table-BERT 基础上利用 RoBEATa 模型替换了 BERT 模型进行验证的基线模型, 这表明了 RoBERTa 模型在表格事实验证任务上的表现优于 BERT 模型. SAT (69.81%) 在自我注意力层上使用相关标记来将行和列的信息汇总到每个单元格, 这表明了理解表格结构对于 BERT 类方法是有效的. 而在 INFOTABS 数据集上, SAT (65.89%) 模型表现不如 Table-BERT (67.17%) 的原因是 INFOTABS 的结构性远远低于 TABFACT, 其表格中的单元格通常由一些句子或段落组成而且内容比较分散, 这不利于 SAT 聚合行和列的信息导致其表现较差. 对于 UCLDS 数据集来说, 由于其和 INFOTABS 数据集的表格结构相似, SAT (72.80%) 模型表现同样不佳.

4.4.2 测试异构图的有效性

为了验证本文所构建的依存句法分析图、命名实体连接图和异构图的有效性, 本文做了如下消融实验.

GAT-DP: 该模型在 RoBERTa-CH 的基础上利用 GAT 捕捉依存句法分析图的结构特征, 目的是为了从句法依存关系的角度理解陈述的结构.

GAT-NER: 模型在 RoBERTa-CH 的基础上利用 GAT 捕捉命名实体连接图的结构特征, 目的是为了从命名实体的角度理解陈述的语义.

HAN: 模型在 RoBERTa-CH 的基础上利用 HAN 捕捉拥有依存句法和命名实体两种元路径的异构图的结构特征, 目的是为了从依存句法关系、命名实体两方面的角度理解陈述的结构和语义.

该消融实验结果如表 7 所示. 从表中可以发现 GAT-DP、GAT-NER 的表现都优于基线模型 RoBERTa-CH, 这验证了本文所构建的依存句法分析图和命名实体连接图在理解陈述的结构和语义上的有效性. 而 HAN 模型的表现优于 GAT-DP 和 GAT-NER, 这说明了融合了两种元路径的异构图在理解陈述上的能力优于单一的依存句法分析图和命名实体连接图.

表 7 异构图消融实验结果 (%)

| 数据集 | 子集 | RoBERTa-CH[28] | GAT-DP | GAT-NER | HAN |
|---|---|---|---|---|---|
| TABFACT | Val | 66.97 | 67.16 | 67.18 | **67.41** |
| | Test | 66.35 | 66.73 | 66.69 | **66.73** |
| | Test_simple | 75.41 | 74.19 | 75.63 | **76.04** |
| | Test_complex | 62.00 | 62.33 | 62.30 | **62.47** |
| | Test_small | 67.92 | 68.94 | 68.79 | **69.04** |
| INFOTABS | Val | 68.61 | 68.83 | 69.56 | **69.89** |
| | $\alpha_1$Test | 67.28 | 67.39 | 68.06 | **68.22** |
| | $\alpha_2$Test | 56.17 | 57.56 | 57.78 | **57.94** |
| | $\alpha_3$Test | 55.39 | 55.39 | 56.11 | **56.33** |
| UCLDS | Val | 78.13 | 81.13 | 81.00 | **81.27** |
| | $\alpha_1$Test | 76.60 | 80.00 | 79.80 | **80.67** |
| | $\alpha_2$Test | 62.20 | 63.07 | 63.67 | **64.07** |
| | $\alpha_3$Test | 72.73 | 75.47 | 75.27 | **75.47** |

综上所述, 本文所提出的融合了依存句法和命名实体两种元路径的异构图能够帮助模型从句法依存关系和命名实体的角度理解陈述的结构和语义, 从而提高模型的表现.

4.4.3 测试 HAN 与 CapsGNN 模块的有效性

为了验证本文所使用的 HAN 模块和 CapsGNN 模块的有效性, 本文做了如下消融实验.

CapsGNN-DP: 该模型在 RoBERTa-CH 的基础上将原始 CapsGNN 模型的基础节点胶囊层所使用的图卷积神经网络 (graph convolutional network, GCN)[33]更改为 GAT, 并利用依存句法分析图构建基础节点胶囊, 最后利用陈述的高质量文本表示进行预测.

CapsGNN-NER: 该模型在 RoBERTa-CH 的基础上将原始 CapsGNN 模型的基础节点胶囊层所使用的 GCN

更改为 GAT, 并利用命名实体连接图构建基础节点胶囊, 最后利用陈述的高质量文本表示进行预测.

该消融实验结果如表 8 所示.

表 8 HAN、CapsGNN 模块消融实验结果 (%)

| 数据集 | 子集 | RoBERTa-CH[28] | HAN | CapsGNN-DP | CapsGNN-NER | Ours |
|---|---|---|---|---|---|---|
| TABFACT | Val | 66.97 | 67.41 | 68.44 | 68.85 | **69.89** |
| | Test | 66.35 | 66.73 | 71.77 | 72.36 | **72.80** |
| | Test_simple | 75.41 | 76.04 | **81.84** | 81.31 | 81.43 |
| | Test_complex | 62.00 | 62.47 | 66.83 | 67.98 | **68.57** |
| | Test_small | 67.92 | 69.04 | 75.66 | **76.88** | 76.12 |
| INFOTABS | Val | 68.61 | 69.89 | 69.56 | 70.56 | **71.61** |
| | $\alpha_1$Test | 67.28 | 68.22 | 69.72 | 69.72 | **70.72** |
| | $\alpha_2$Test | 56.17 | 57.94 | 60.33 | **61.56** | 61.28 |
| | $\alpha_3$Test | 55.39 | 56.33 | 60.61 | 59.33 | **60.61** |
| UCLDS | Val | 78.13 | 81.27 | 80.40 | 81.20 | **81.67** |
| | $\alpha_1$Test | 76.60 | 80.67 | 79.00 | 78.47 | **81.00** |
| | $\alpha_2$Test | 62.20 | 64.07 | 64.53 | 63.93 | **69.13** |
| | $\alpha_3$Test | 72.73 | 75.47 | 75.20 | 75.93 | **76.27** |

从表 8 中可以发现所提模型的表现基本上优于没有使用胶囊图神经网络的 HAN 模型, 这说明了 CapsGNN 模块在提取高质量文本特征上的有效性. 同时所提模型的表现基本上优于没有使用异构图注意力网络模块的 CapsGNN 模型, 这证明了 HAN 模块在利用图结构提取文本特征方面的有效性.

#### 4.4.4 测试异构图中区分边的种类的有效性

为了验证异构图中区分边种类的有效性, 本文将异构图中两种图的边种类不加以区分, 即依存句法分析图和命名实体连接图中边的权值全部置为 1 以作为对比实验, 这样模型在理解陈述语义的时候对于所有依存句法类型和命名实体类型将不再区分, 该模型记为 CapsHAN-SameEage. 实验结果如表 9 所示.

表 9 边的种类对模型效果的影响 (%)

| 数据集 | 子集 | CapsHAN-SameEage | Ours |
|---|---|---|---|
| TABFACT | Val | 68.77 | **69.89** |
| | Test | 71.20 | **72.80** |
| | Test_simple | 80.31 | **81.43** |
| | Test_complex | 66.73 | **68.57** |
| | Test_small | 75.71 | **76.12** |
| INFOTABS | Val | 69.89 | **71.61** |
| | $\alpha_1$Test | 67.67 | **70.72** |
| | $\alpha_2$Test | 56.50 | **61.28** |
| | $\alpha_3$Test | 56.33 | **60.61** |
| UCLDS | Val | 81.13 | **81.67** |
| | $\alpha_1$Test | 79.67 | **81.00** |
| | $\alpha_2$Test | 65.93 | **69.13** |
| | $\alpha_3$Test | 75.07 | **76.27** |

从表 9 中可以看出, 区分边种类的模型表现优于没有区分的模型, 本文认为让模型在理解陈述时区分不同类型的句法依存关系和命名实体是有意义的, 由此证明了区分边种类的有效性.

#### 4.4.5 测试 HAN 模块中层数 $L$ 的影响

为了测试 HAN 模块中层数 $L$ 对实验结果的影响, 本文分别验证了 $L$ 为 1–5 时模型的表现, 结果如图 10 所示.

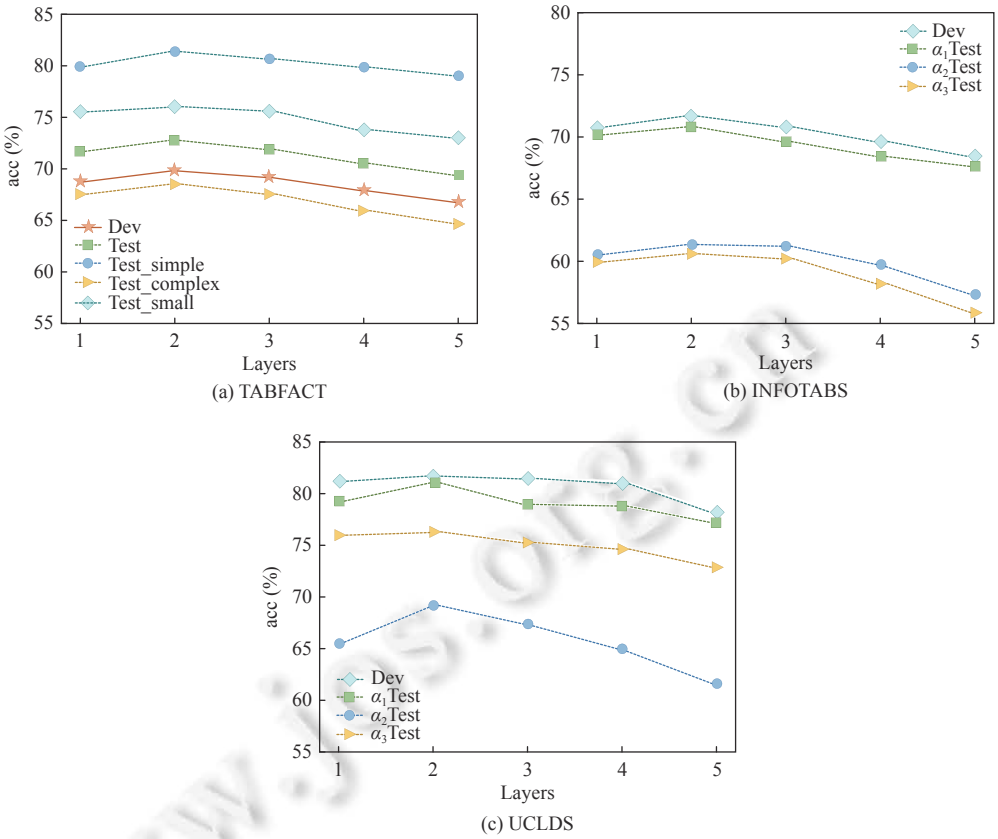



图 10 层数对于模型效果的影响

从图 10 中可以得知, 随着 HAN 模块层数 $L$ 的增加, 模型的性能先上升后下降, 这意味着合理增加异构图注意力网络层可以提高 CapsHAN 的推理能力. 但是如果层数设置的太深, 可能会导致梯度爆炸问题, 进而可以推断出较多层不足以进行推理.

#### 4.4.6 测试数据降噪的有效性

在对 TABFACT 英文数据集进行中文转换的时候, 经过对比发现数据集中存在一些非中英文的字符, 而这些字符会作为噪声干扰中文预训练模型在推理时的表现, 所以本文对这些字符进行了转换和删除. 表 10 为利用 RoBERTa-wwm-ext-large-Chinese 对降噪前后数据集进行事实验证的对比. 从表中可以看出降噪以后模型的表现明显提升, 这验证了对 TABFACT 中文数据集进行降噪的有效性.

表 10 TABFACT 降噪前后对比实验结果 (%)

| 降噪前/后 | Val | Test | Test_simple | Test_complex | Test_small |
|---|---|---|---|---|---|
| 降噪前 | 66.21 | 65.99 | 74.40 | 61.85 | 66.70 |
| 降噪后 | **66.97** | **66.35** | **75.41** | **62.00** | **67.92** |

## 5 结 论

本文研究了一个非常重要但目前尚未充分探索的问题: 基于中文表格型数据的事实验证. 本文提出了一种以陈述为中心的框架. 该框架采用 HAN 和 CapsGNN 来充分利用陈述的结构和语义, 以此为基于表格的事实验证提供细粒度的推理. 除此之外, 本文提出了一种基于依存句法分析和命名实体识别的异构图构建方法, 该方法通过加

深模型对于陈述的理解从而增强模型的推理能力. 实验结果表明, 该框架在所提的 TABFACT、INFOTABS 和 UCLDS 这 3 个中文基准数据集都取得了最好的表现, 同时证明了这 3 个数据集具有可学习性. 对于 UCLDS 数据集, 本文发现这种将单表和多表相结合的推理任务具有挑战性. 未来计划设计一种专门针对中文表格的预训练语言模型, 引入单元格级别的自注意力机制, 目的是将表格的结构信息融入到所生成的文本特征中, 用所设计的模型取代本文框架中的表格编码器, 在充分挖掘陈述隐含表格信息的同时理解表格的结构和内容, 以提供更细粒度的推理.

**References:**


[1] Deng ZY, Zhang M. Graph neural networks for table-based fact verification. Ruan Jian Xue Bao/Journal of Software, 2021, 32(3): 753−762 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6184.htm [doi: 10.13328/j.cnki.jos.006184]

[2] Chen CH, Cai F, Hu XJ, Chen WY, Chen HH. HHGN: A hierarchical reasoning-based heterogeneous graph neural network for fact verification. Information Processing & Management, 2021, 58(5): 102659. [doi: 10.1016/J.IPM.2021.102659]

[3] Zhao C, Xiong CY, Rosset C, Song X, Bennett P, Tiwary S. Transformer-XH: Multi-evidence reasoning with extra hop attention. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa, 2020.

[4] Si JS, Zhou DY, Li TZ, Shi XY, He YL. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. arXiv:2106.01191, 2021.

[5] Zhou J, Han X, Yang C, Liu ZY, Wang LF, Li CC, Sun MS. GEAR: Graph-based evidence aggregating and reasoning for fact verification. arXiv:1908.01843, 2019.

[6] Chen WH, Wang HM, Chen JS, Zhang YK, Wang H, Li SY, Zhou XY, Wang WY. TABFACT: A large-scale dataset for table-based fact verification. arXiv:1909.02164, 2020.

[7] Gupta V, Mehta M, Nokhiz P, Srikumar V. INFOTABS: Inference on tables as semi-structured data. arXiv:2005.06117, 2020.

[8] Zhong WJ, Tang DY, Feng ZY, Duan N, Zhou M, Gong M, Shou LJ, Jiang DX, Wang JH, Yin J. LogicalFactChecker: Leveraging logical operations for fact checking with graph module network. arXiv:2004.13659, 2020.

[9] Shi Q, Zhang Y, Yin QY, Liu T. Learn to combine linguistic and symbolic information for table-based fact verification. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 5335−5346. [doi: 10.18653/v1/2020.coling-main.466]

[10] Yang XY, Nie F, Feng YF, Liu Q, Chen ZG, Zhu XD. Program enhanced fact verification with verbalization and graph attention network. arXiv:2010.03084, 2021.

[11] Zhang HZ, Wang YY, Wang SR, Cao XZ, Zhang FZ, Wang ZY. Table fact verification with structure-aware transformer. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 1624–1629. [doi: 10.18653/v1/2020.emnlp-main.126]

[12] Eisenschlos JM, Krichene S, Müller T. Understanding tables with intermediate pre-training. arXiv:2010.00571, 2020.

[13] Yang XY, Zhu XD. Exploring decomposition for table-based fact verification. arXiv:2109.11020, 2021.

[14] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2019.

[15] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans. on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]

[16] Wang X, Ji HY, Shi C, Wang B, Ye YF, Cui P, Yu PS. Heterogeneous graph attention network. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 2022–2032. [doi: 10.1145/3308558.3313562]

[17] Zhang XY, Chen LH. Capsule graph neural network. In: Proc. of the 2019 Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019. 1–16.

[18] General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. GB/T 35304-2017 Uniform content label format specification. Beijing: Standards Press of China, 2017 (in Chinese).

[19] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv:1710.10903, 2018.

[20] Neeraja J, Gupta V, Srikumar V. Incorporating external knowledge to enhance tabular reasoning. arXiv:2104.04243, 2021.

[21] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

[22] Wang W. UCL knowledge space based on entity in dual-structural network [MS. Thesis]. Nanjing: Southeast University, 2018 (in

Chinese with English abstract).
[23] Chang XC. Web news search system based on UCL knowledge space [MS. Thesis]. Nanjing: Southeast University, 2021 (in Chinese with English abstract). [doi: 10.27014/d.cnki.gdnau.2021.002870]
[24] Yang HR. Social media sentiment analysis based on knowledge graph [MS. Thesis]. Nanjing: Southeast University, 2020 (in Chinese with English abstract). [doi: 10.27014/d.cnki.gdnau.2020.003347]
[25] Zong CQ. Statistical Natural Language Processing. 2nd ed., Beijing: Tsinghua University Press, 2013 (in Chinese).
[26] Wan QZ, Wan CX, Hu R, Liu DX. Chinese financial event extraction base on syntactic and semantic dependency parsing. Chinese Journal of Computers, 2021, 44(3): 508–530 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00508]
[27] Che WX, Li ZH, Liu T. LTP: A Chinese language technology platform. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics: Demonstrations. Beijing: Association for Computational Linguistics, 2010. 13–16.
[28] Cui YM, Che WX, Liu T, Qin B, Yang ZQ. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]
[29] Chen YF, Zong CQ, Su KY. Joint Chinese-English named entity recognition and alignment. Chinese Journal of Computers, 2011, 34(9): 1688–1696 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01688]
[30] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3859–3869.
[31] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Computational Linguistics, 2008, 34(4): 555–596. [doi: 10.1162/coli.07-034-R2]
[32] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000-6010.
[33] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907, 2017.

**附中文参考文献**:

[1] 邓哲也, 张铭. 用于表格事实检测的图神经网络模型. 软件学报, 2021, 32(3): 753−762. http://www.jos.org.cn/1000-9825/6184.htm [doi: 10.13328/j.cnki.jos.006184]
[18] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 35304-2017 统一内容标签格式规范. 北京: 中国标准出版社, 2017.
[22] 汪巍. 双结构网络中基于实体的UCL知识空间研究 [硕士学位论文]. 南京: 东南大学, 2018.
[23] 常欣辰. 基于UCL知识空间的网页新闻搜索系统研究与实现 [硕士学位论文]. 南京: 东南大学, 2021. [doi: 10.27014/d.cnki.gdnau.2021.002870]
[24] 杨浩然. 基于知识图谱的社交媒体情感分析研究 [硕士学位论文]. 南京: 东南大学, 2020. [doi: 10.27014/d.cnki.gdnau.2020.003347]
[25] 宗成庆. 统计自然语言处理. 第2版, 北京: 清华大学出版社, 2013.
[26] 万齐智, 万常选, 胡蓉, 刘德喜. 基于句法语义依存分析的中文金融事件抽取. 计算机学报, 2021, 44(3): 508–530. [doi: 10.11897/SP.J.1016.2021.00508]
[29] 陈钰枫, 宗成庆, 苏克毅. 汉英双语命名实体识别与对齐的交互式方法. 计算机学报, 2011, 34(9): 1688–1696. [doi: 10.3724/SP.J.1016.2011.01688]

## 附录 A　UCLDS 数据集的推理规则

为了提升 UCLDS 数据集的质量和挑战性, 本文在该数据集中引入了多种推理类型. 需要注意的是, 一个单一的前提-假设对可能与多种类型的推理有关. 如果在假设中使用了多次同一种推理类型, 本文只标记一次. 推理类型的分类具体如下.

(1) 简单查找 (simple lookup): 这是一个没有推理的简单情况, 假设是通过重申表格中的信息形成的.

(2) 多行推理 (multi-row reasoning): 这种推理涉及表格中的多行来进行推断.这就对推理提出了更高的要求: 如果没有多行, 就无法得出结论.

(3) 常识推理 (common sense reasoning): 这种推理与世界知识和常识有关. 由于常识分为很多种, 本文将对常识推理进行细分.

• 词语替换推理: 推理中的词语替换又分为常用词语替换和命名实体替换. 常用词语替换就是将表格中的词语替换成近义词,

例如词语“喜欢”可以替换成“中意”. 而对于命名实体替换, 就是将特定的命名实体替换成它的别称, 例如“东南大学”可以替换成“SEU”或者“东大”.

• 否定推理: 任何对于表格内容的否定都应该属于这一种推理类型. 无论是对词语语义的否定还是对表格内容所述事实的否定, 本文都将其归为否定推理.

• 世界知识推理: 此类推理中会涉及基本的世界知识, 例如, “太阳不会从西边升起”, 而表格的信息中并不包含这些知识.

• 数值关系推理: 任何涉及到关于数字的推理都属于这一类. 这种推理类型又分为聚合关系推理、最高级推理和比较推理. 聚合关系推理包括对表格中的数字进行计数、求均值等一系列数学运算操作; 最高级推理涉及对数字进行排名, 即诸如求最大值、最小值之类的操作; 比较推理主要是对数字进行数值比较, 例如大于、小于或等于.

• 时间关系推理: 任何涉及关于时间量的数值推理和时间知识的使用都属于这一类. 例如“120 分钟”等价于“2 小时”“9 点比 7 点更晚”等.

• 包含关系推理: 任何涉及包含关系的推理都属于这一类型. 例如电影《阿凡达》的主演有萨姆·沃辛顿、佐伊·索尔达娜和西格妮·韦弗等, 那么萨姆·沃辛顿包含于《阿凡达》的主演之中.

(4) 复合推理 (complex reasoning): 本文把涉及 3 种或 3 种以上在上文中提到的推理称为复合推理, 这种推理也将是最复杂、最具挑战性的推理类型.

**杨鹏**(1975－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为自然语言处理, 新型网络, 大数据治理.

**赵广振**(1992－), 男, 博士, CCF 学生会员, 主要研究领域为自然语言处理, 机器学习.

**查显宇**(1998－), 男, 硕士生, 主要研究领域为自然语言处理, 机器学习.

**林茜**(1999－), 女, 硕士生, 主要研究领域为自然语言处理, 计算机三维视觉.