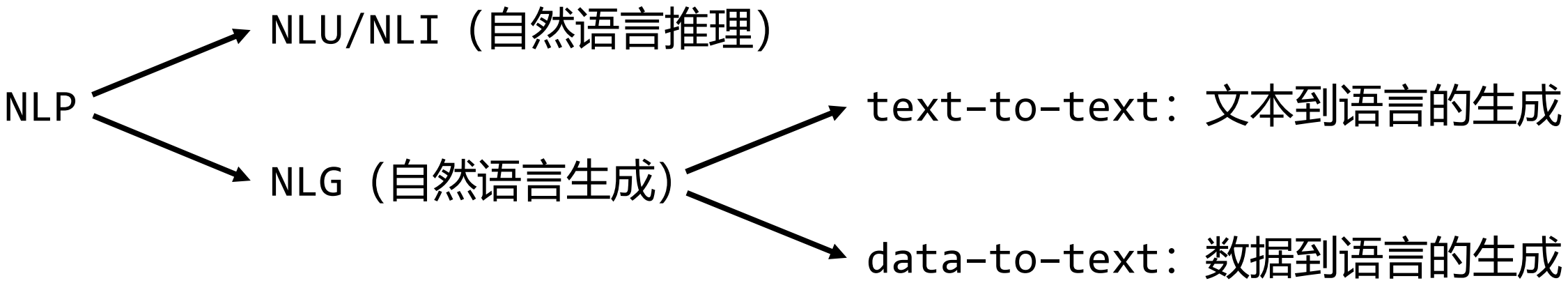
Table-to-Text

2022.7.8

石微微

目录

表格生成文本



- Text-to-text：文本摘要、翻译等
- Data-to-text：图像生成文本、表格生成文本等

表格生成文本——数据集

- 生成表面事实的文本
  - RotoWire

    输入：三个表格，包含两个球队整体表现的记录、有关球员在主队的表现以及球员在客队的表现

    输出：一段对比赛情况的描述
  - WikiBio

    输入：维基百科的infobox

    输出：传记的第一句话
  - WeatherGOV

    输入：温度、天空条件等的结构化表示的记录

    输出：天气预报

表格生成文本——数据集

- 受控文本生成
  - ToTTo

    输入：一个表格，表格标题信息，一组突出显示的单元格

    输出：描述突出显示的单元格的句子

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description:None

		RUSHING					RECEIVING				
YEAR	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1050	4.9	62	9	92	1016	11	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	-	**1991**	**8189**	**4.1**	**71**	**56**	**566**	**4911**	**8.7**	**73**	**17**

Target Text: Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

表格生成文本——数据集

- 带逻辑推理的文本生成
  - LogicNLG

    输入：一个表格

    输出：包含表格内容的一个带有逻辑推理的句子
  - Logic2Text

    输入：一个表格，一个逻辑形式

    输出：该逻辑形式对应的句子

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation
Sentence: Canada obtained 1 more gold medal than Mexico. **Sentence**: Canada obtained the most gold medals in the game.

Medal Table from Tournament

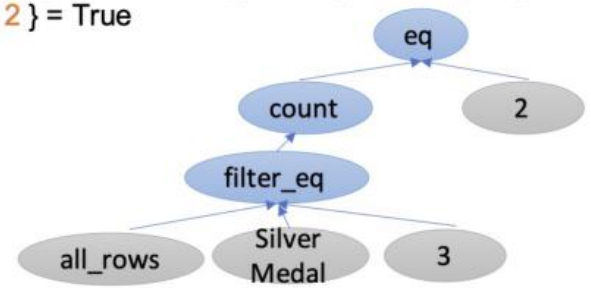
Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level generation

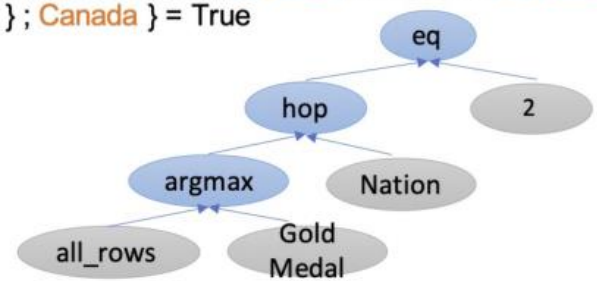
Sentence: Canada has got 3 gold medals in the tournament.
Sentence: Mexico got 3 silver medals and 1 bronze medal.

Logical-level generation with logical forms (Logic2text)

Sentence: There are 2 nations getting 3 silver medals in the game.
Logical form: eq { count { filter_eq { all_rows ; Silver Medal ; 3 } } ; 2 } = True



Sentence: Canada obtained the most gold medals in the game.
Logical form: eq { hop { argmax { all_rows ; Gold Medal } ; Nation } ; Canada } = True



表格生成文本——数据集

Data Set	Domin	Table	Examples	Sourse
WeatherGOV (2009)	Weather	22.1K	22.1K	www.weather.gov
WikiBio (2016)	Biography	728K	728K	Wikipedia
RotoWire (2017)	NBA	4.9K	4.9K	Rotowire
ToTTO (2020)	Open	136K	136K	Wikipedia
LogicNLG (2020)	Open	7.3K	37.0K	Wikipedia
Logic2Text (2020)	Open	5.5K	10.7K	Wikipedia

表格生成文本——评估方法

- 相似度评估：BLEU-1，2，3
- 逻辑保真度评估：
  - 基于IE抽取式的评估

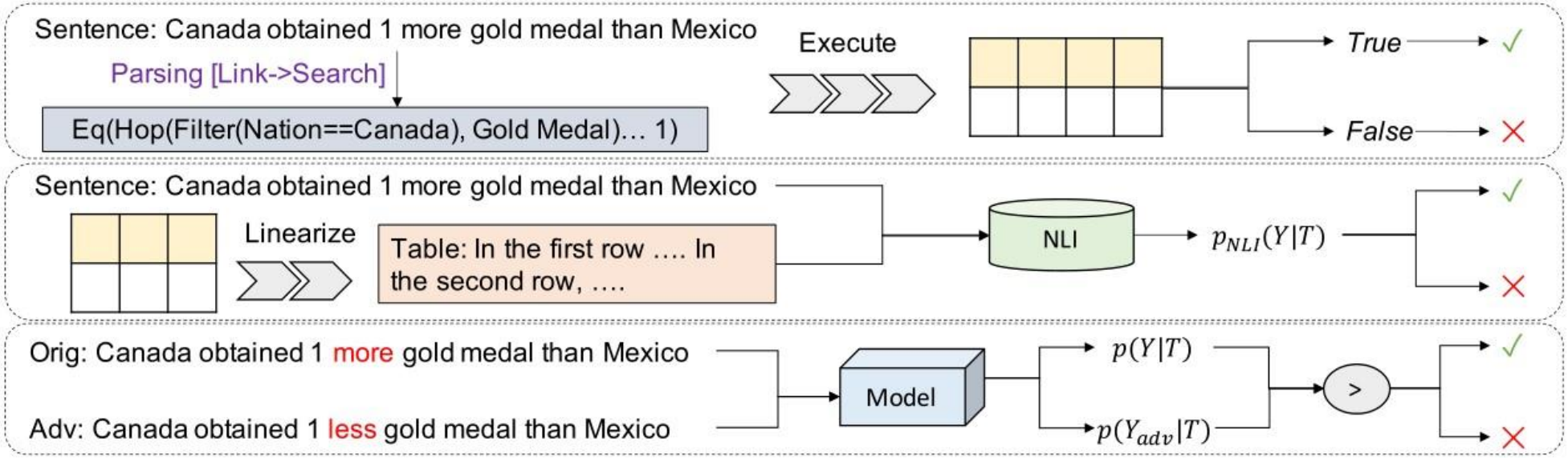
    存在问题：

    1．空抽取：句子不能被表述为（主语、谓语、宾语）结构，因此 IE 系统无法提取三元组进行验证。

    2．假阴性：句子是表中事实的逻辑组合（而不是表层形式），IE系统无法将其与表匹配。

表格生成文本——评估方法

- 相似度评估：BLEU-1，2，3
- 逻辑保真度评估：
  - 基于IE抽取式的评估
  - 基于解析的评估：直接从生成的句子中提取逻辑表达式，并在表中执行以验证其正确性
  - 基于NLI的评估：基于 TableBERT
  - 对抗性评估：如果原始句子的置信度高于其对抗性句子，我们将其视为成功的防御。



目录

1.表格生成文本

2.方法

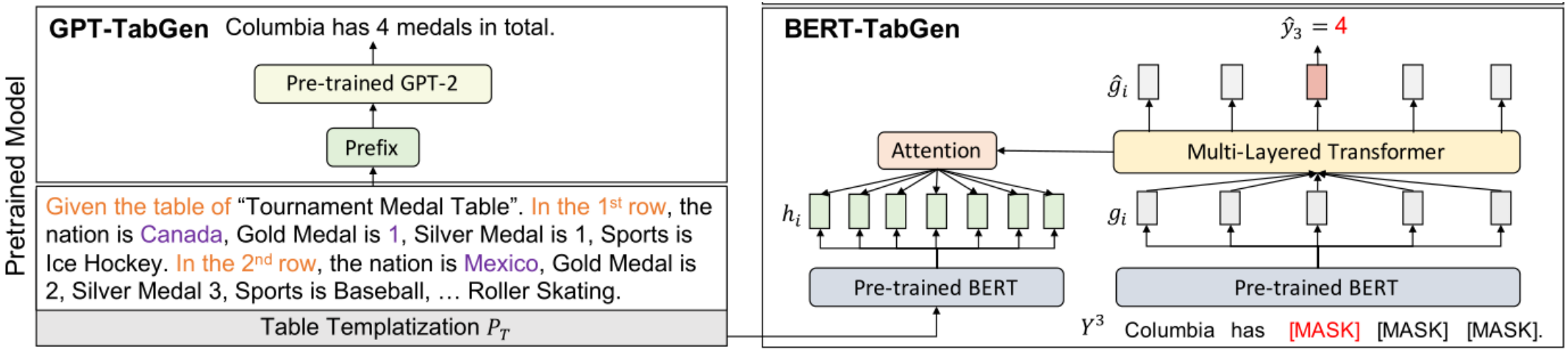
- 由粗到细
- 基于关系实体图
- 多样性增强
- 基于分解

3.我的论文

4.总结

由粗到细的方法

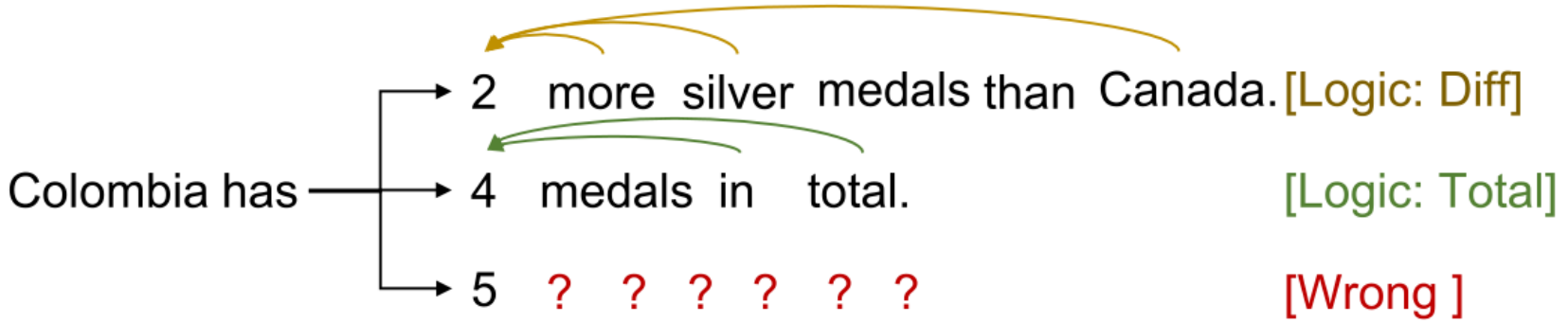
Logical Natural Language Generation from Open-Domain Tables(2020)

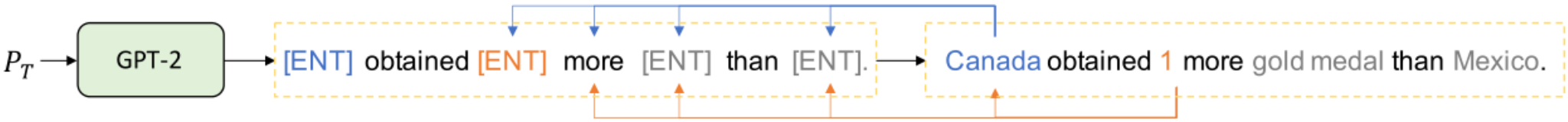


- 基于GPT-2。将表格线性化。然后直接将该句子作为输入提供给GPT-2生成句子。
- 基于BERT。将表格线性化。将该句子使用BERT模型编码。在第i个时间步后换为对应的生成句子并进行部分遮罩，然后用BERT编码遮罩词。对编码的遮罩词使用注意力层获得输出的隐藏状态并用于预测。

由粗到细的方法

Logical Natural Language Generation from Open-Domain Tables(2020)





- 第一阶段：生成一个确定全局逻辑结构的模板
- 第二阶段：根据第一阶段生成的模板生成最终的、有根据的句子

目录

1.表格生成文本

2.方法

- 由粗到细
- 基于关系实体图
- 多样性增强
- 基于分解

3.我的论文

4.总结

基于关系实体图的方法

Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation

Name	PTS	AST	REB	TOV	…
Celtics	115	23	53	15	…
Knicks	87	19	57	25	…

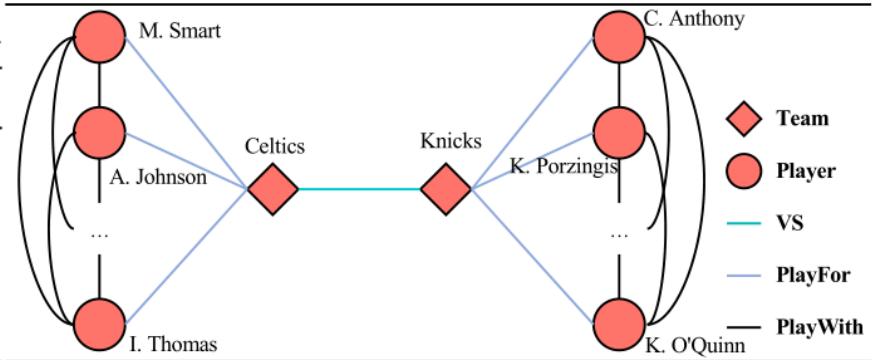
Name	MIN	PTS	AST	REB	…
Marcus Smart	34	12	10	6	…
Amir Johnson	21	2	1	3	…
Kelly Olynyk	30	19	3	7	…
…	…	…	…	…	…
Isaiah Thomas	28	29	4	3	…

Name	MIN	PTS	AST	REB	…
Carmelo Anthony	12	12	1	2	…
Kristaps Porzingis	23	14	1	6	…
Joakim Noah	22	9	2	10	…
…	…	…	…	…	…
Kyle O'Quinn	3	2	0	3	…

(a)

The Boston Celtics dominated the visiting New York Knicks, 115 - 87, on Friday night at TD Garden … Isaiah Thomas was huge for Boston (4 - 4) as he led the way offensively with 29 points on 9-of-17 shooting, in only 28 minutes... Avery Bradley and Marcus Smart both filled the stat sheet. Smart finished with 12 points, 10 assists, six rebounds and three steals, while Bradley notched 15 points, 10 rebounds, two assists and four steals… Kristaps Porzingis was the high-point man with 14 points, along with six rebounds and two blocks , in 23 minutes. Derrick Rose added 11 points , six assists and four rebounds , while the only other player to tally double-digits for the Knicks was Justin Holiday…

(b)



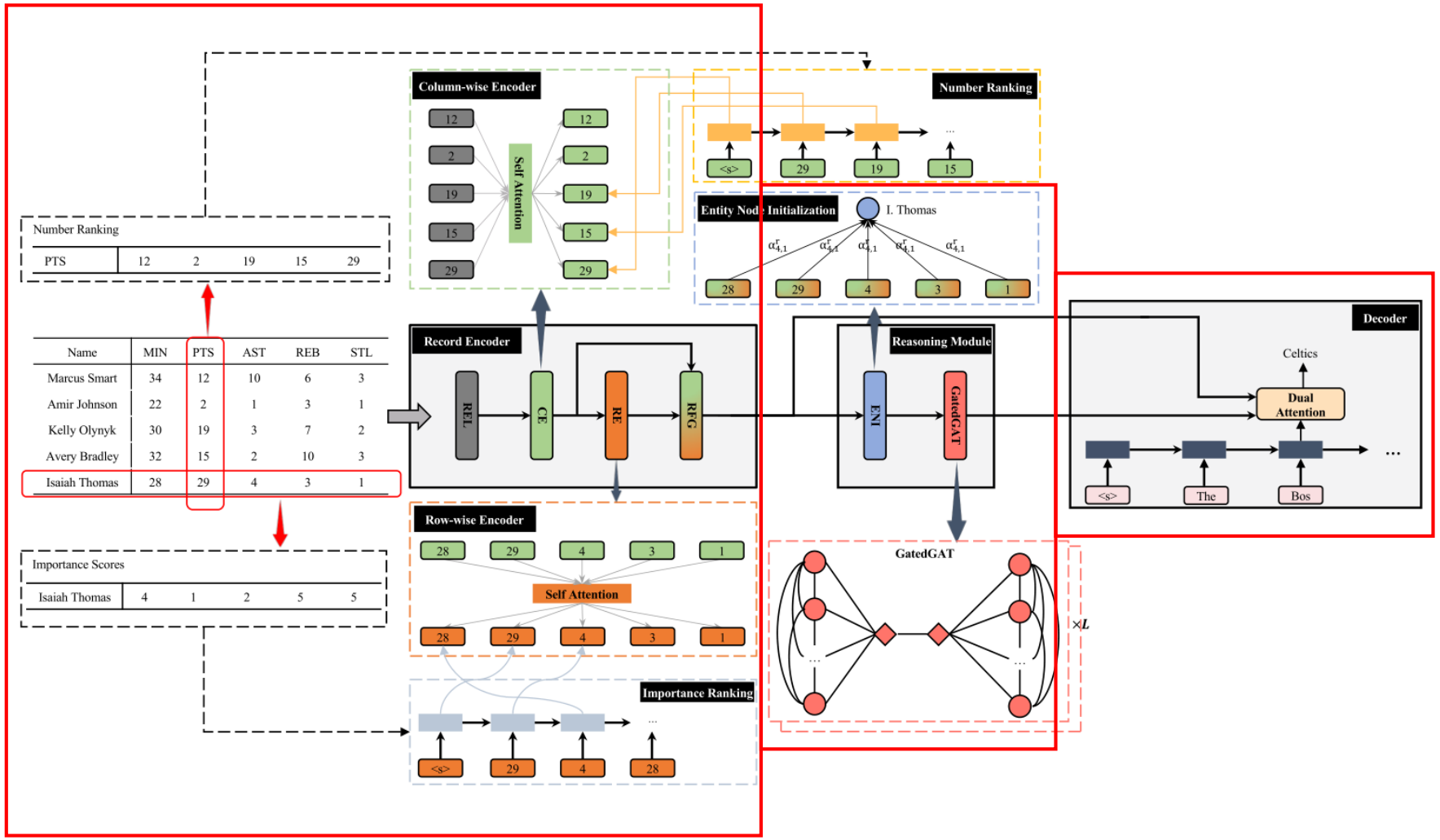
(c)

两种关系：

- 列维度中的数值大小关系
- 行维度中的相对重要性关系

基于关系实体图的方法

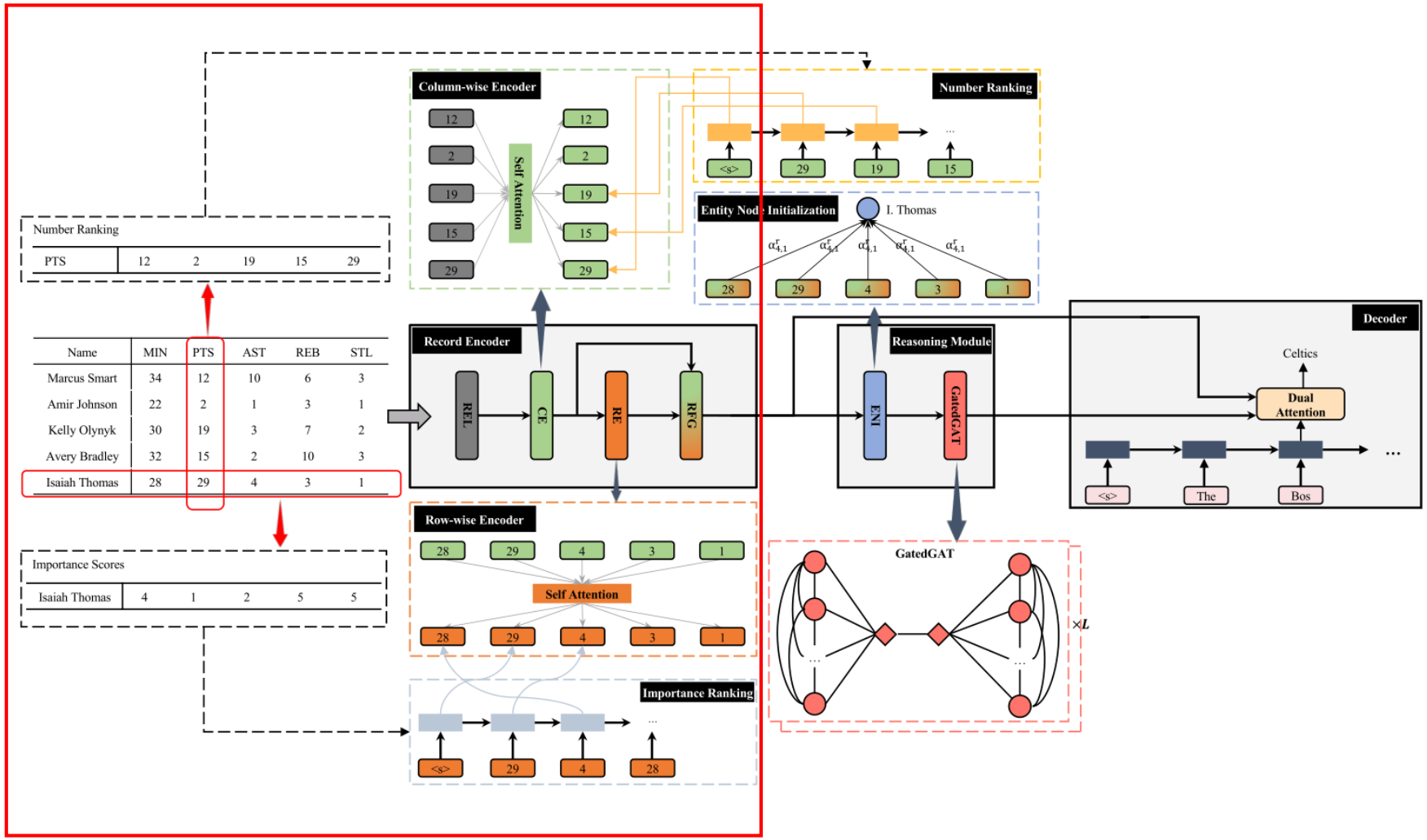
[Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation(2021)]



- 记录编码器
- 推理模块
- 解码器

基于关系实体图的方法

Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation

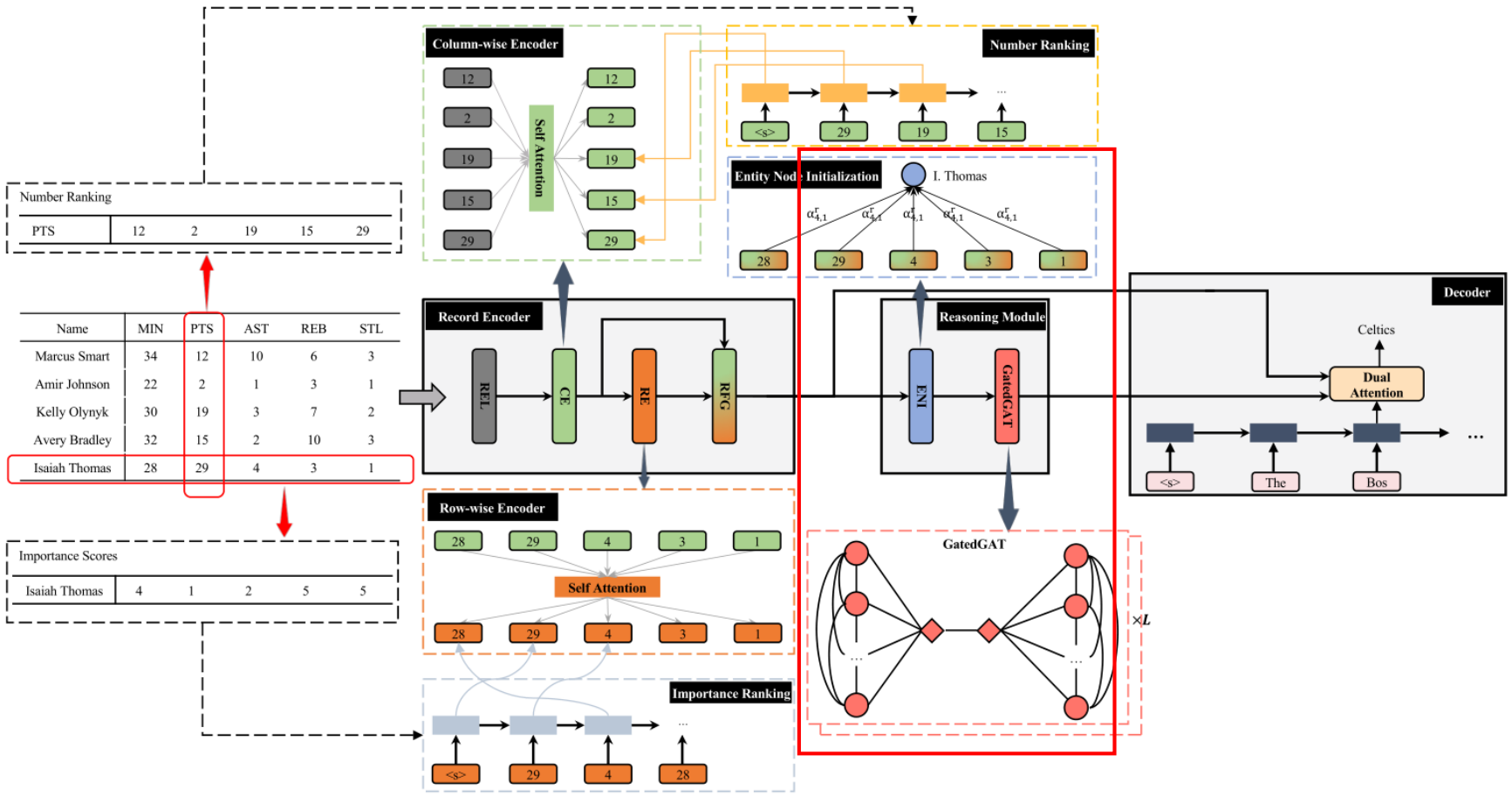


记录编码器：

- REL：记录编码
- CE：列式编码（引入大小排序NR）
- RE：行式编码（引入重要性排序IR）
- RFG：记录融合门

基于关系实体图的方法

Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation

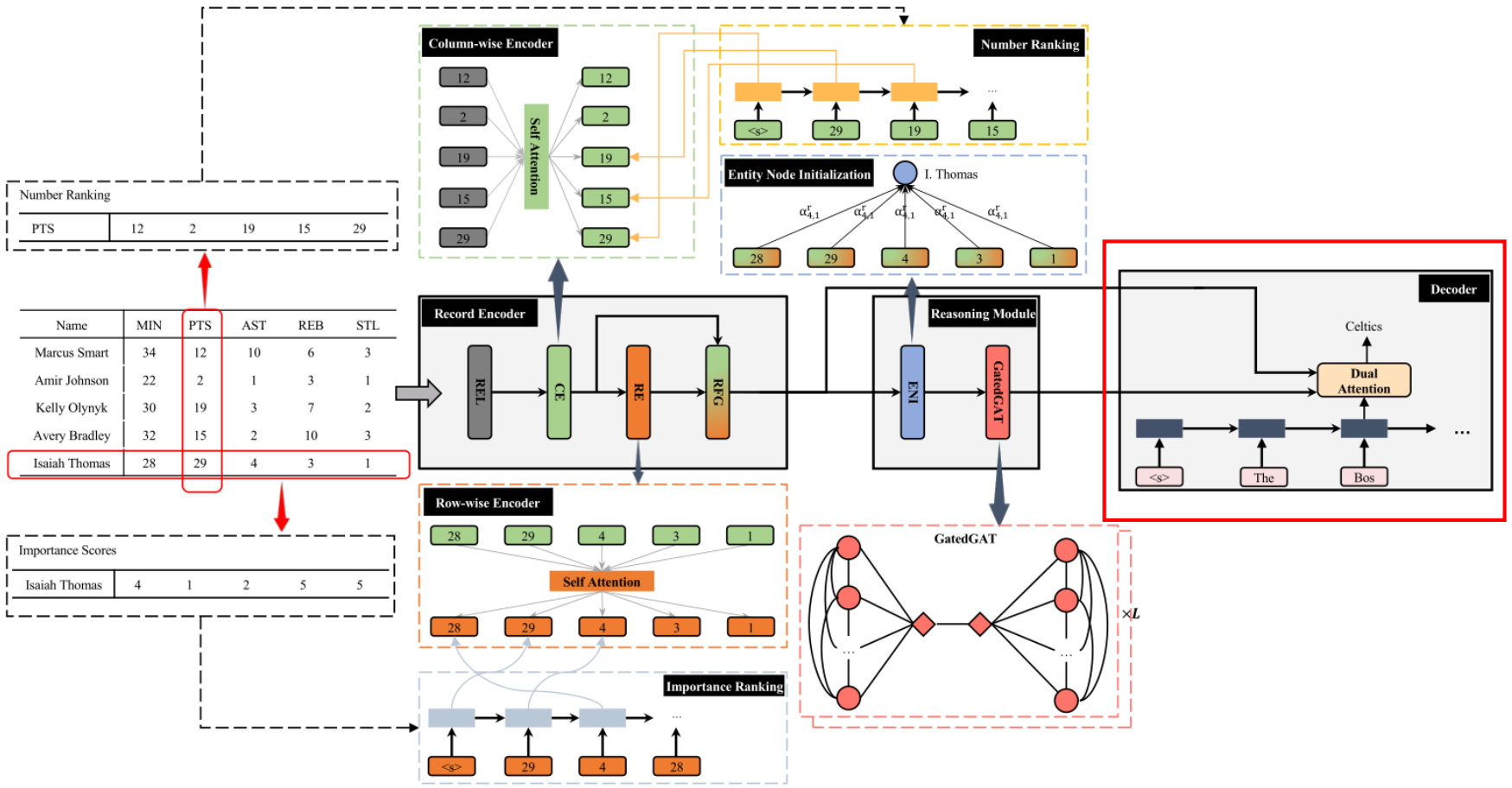


推理模块：

- ENI： 节点初始化模块
- GatedGAT： 图注意力模块

基于关系实体图的方法

Improving Encoder by Auxiliary Supervision Tasks for Table-to-Text Generation



解码器：

双重注意力机制

目录

多样性增强的方法

Diversity Enhanced Table-to-Text Generation via Type Control(2022)

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation
Sentence: Canada obtained 1 more gold medal than Mexico. **Sentence**: Canada obtained the most gold medals in the game.

1.要生成哪种逻辑类型的句子？比较？求和？最高级？

2.对于一种逻辑类型，要用哪些数据生成这样的句子？

多样性增强的方法

Diversity Enhanced Table-to-Text Generation via Type Control(2022)

Worldwide cheese market cap

Year	Market cap
2022	81.2
2021	76.1
2020	63.8
...	...
1961	12.1
1960	14.1

(a) Diversity Enhancement via Type Control
The cheese market cap has **risen by** 17.4B USD **between 2022 and 2020**
The cheese market cap had passed a value of 60B USD in **only 3 years**
The **average cheese market cap** between 1980 to 2000 was 51.3B USD

(b) Diversity Enhancement via Decoding Techniques
2022 is the year with the **highest** cheese market cap with 81.2B USD
2022 is the year with the **largest** cheese market cap at 81.2B USD
In 2022, the **largest** cheese market cap was 81.2B USD

Figure 1: T2T generation of 3-statement sets for the table on the left; (a) logic-type controlled: each statement delivers a unique piece of information, yielded by the control employed: compare, count, and aggregation; (b) decoding-based diversity: all are focused on one fact, hence demonstrating a weak diversity.

Table-to-Text任务的关键特征:

- 多样性
  - (a) 基于逻辑类型控制
  - (b) 基于解码技术
- 可控性

七种逻辑类型:

- Count
- Comparative
- Superlative
- Unique
- Ordinal: 第n个最大值或最小值
- Aggregation: 列上的和或平均值
- Majority: 列上的大多数值

多样性增强的方法

Diversity Enhanced Table-to-Text Generation via Type Control(2022)

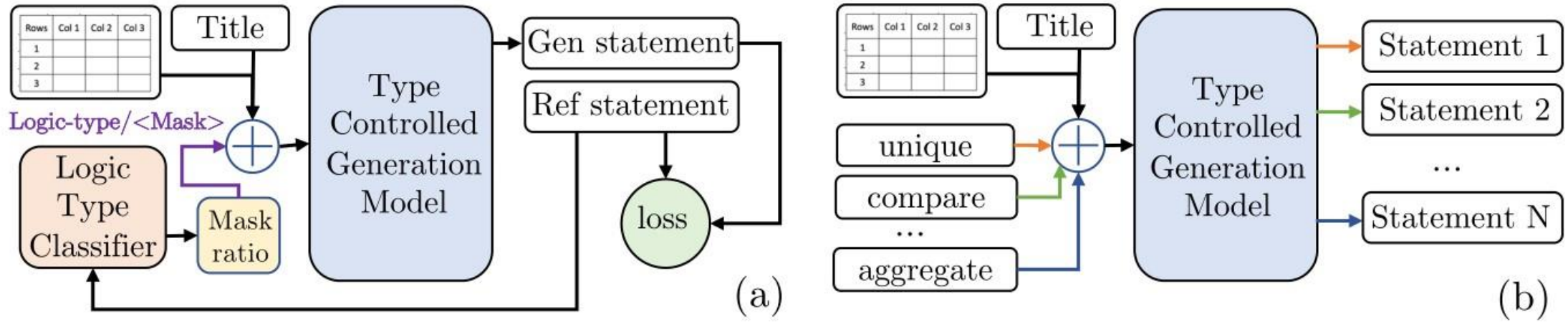


Figure 2: DevTC architecture; (a) DEVTC is trained to generate a reference statement given the statement logic-type as it is predicted by our type classifier; (b) at inference time, DEVTC can receive a table and multiple logic-types, enabling the generation of a diverse set of statements for a single table.

- （a）对DEVTC进行训练，使其根据类型分类器预测的逻辑类型生成语句
- （b）DEVTC可以接收一个表和多个逻辑类型，从而支持为单个表生成不同的语句集

目录

基于分解的方法

Exploring Decomposition for Table-based Fact Verification(2021)

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation
Sentence: Canada obtained 1 more gold medal than Mexico. **Sentence**: Canada obtained the most gold medals in the game.

表面事实/简单事实要如何得到呢?

可以由最终带有逻辑的句子进行分解

基于分解的方法

Exploring Decomposition for Table-based Fact Verification(2021)

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

两种推理形式： Ⅰ.语义推理 Ⅱ.符号推理

基于分解的方法

[Exploring Decomposition for Table-based Fact Verification(2021)]

Statement

The firhill venue had the highest attendance.

Table

Date	Venue	Attendance
march 2009	east end park	2736
april 2009	firhill	4909
april 2009	mcdiarmid park	2830
april 2009	cappielow	3323

Intermediate Evidence

e1: The highest attendance is **4909**.

e2: The attendance of firhill venue is **4909**.

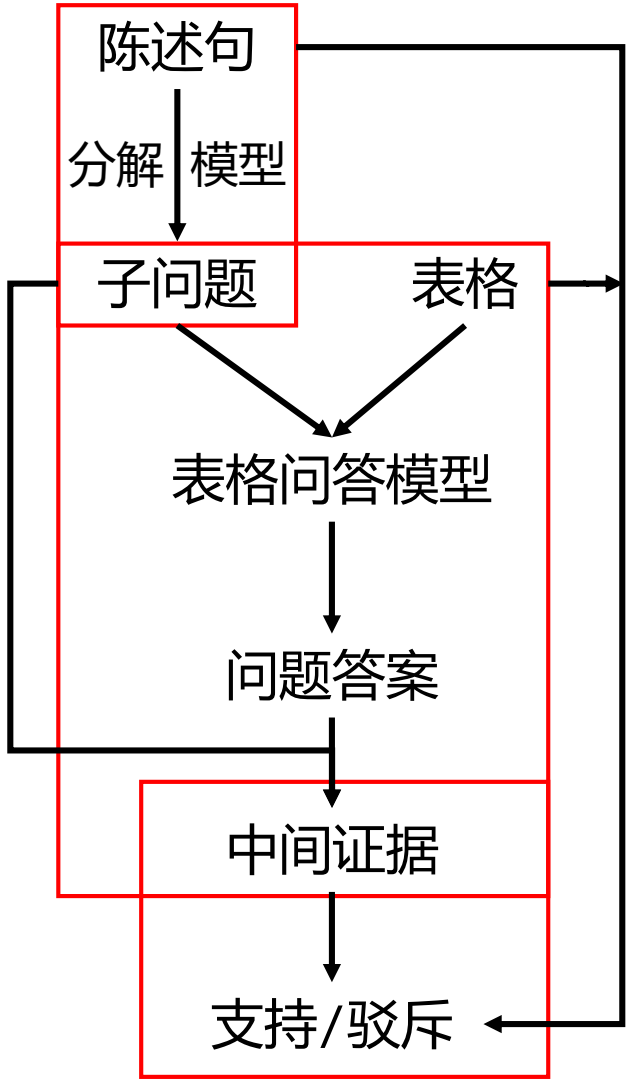
Entailed

Decomposition Model

d1: What is the highest attendance ?

d2: What is the attendance of firhill venue?

Subproblem Solver



基于分解的方法——语句分解

Exploring Decomposition for Table-based Fact Verification(2021)

1．使用LPA算法将陈述句解析为候选程序

2．针对程序分解为四种类型：

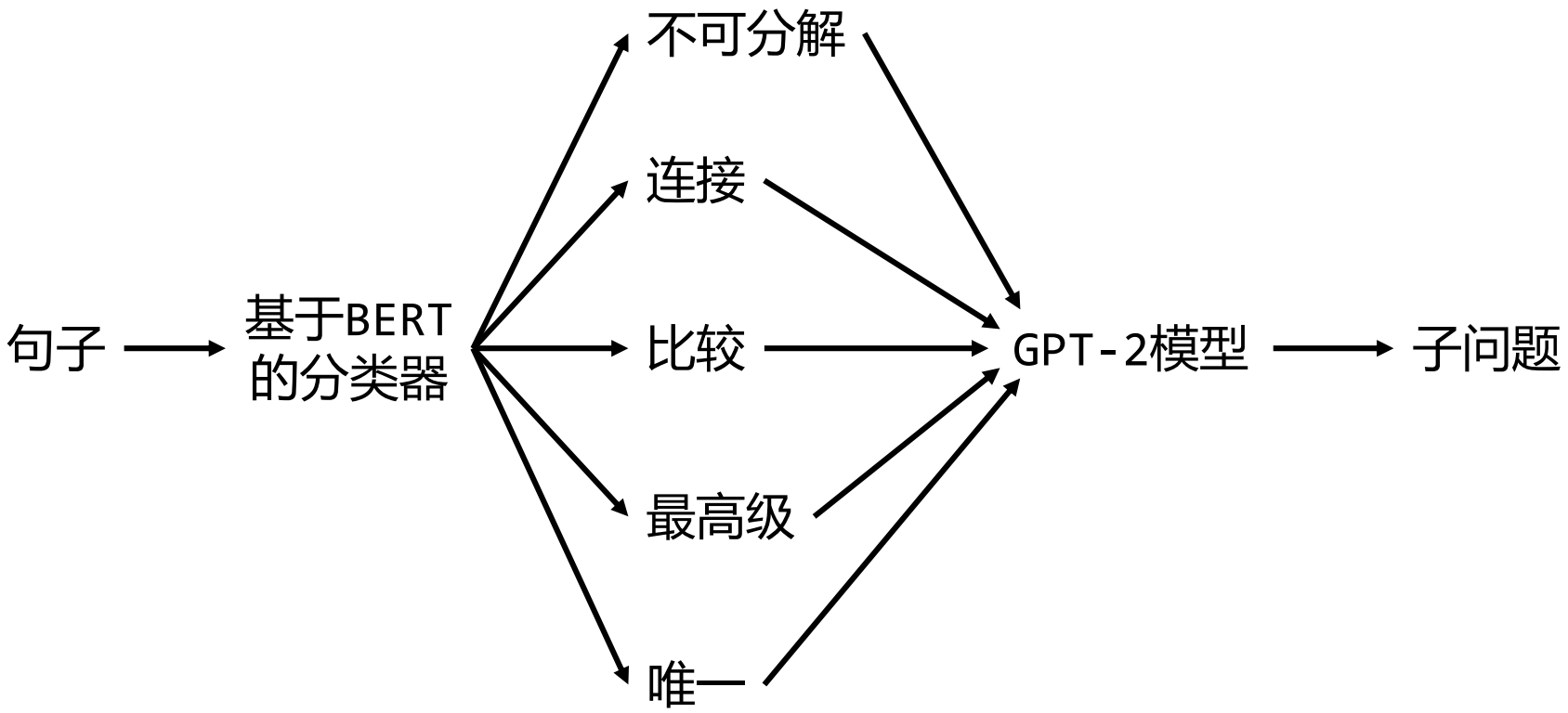
- 连接
- 比较
- 最高级
- 唯一

Conjunction	
S	rayo earns 36 points and ferrol earns 41 points
z	*and { eq { hop { filter_eq { all_rows ; club ; rayo } ; points } ; 36 } ; eq { hop { filter_eq { all_rows ; club ; ferrol } ; points } ; 41 } }*
d1	rayo earns 36 points .
d2	ferrol earns 41 points .
Superlative	
S	princes park venue recorded the highest crowd participation
z	*eq { hop { argmax { all_rows ; crowd } ; crowd } ; hop { filter_eq { all_rows ; venue ; princes park } ; crowd } }*
d1	what is the highest crowd ?
d2	what is the crowd of princes park ?
Comparative	
S	daniel had a longer react than felix
z	*greater { hop { filter_eq { all_rows ; athlete ; daniel } ; react } ; hop { filter_eq { all_rows ; athlete ; felix} ; react } }*
d1	what is the react of daniel ?
d2	what is the react of felix ?
Uniqueness	
S	itf 25k was only the tier on may 8th
z	*and { only { filter_eq { all_rows ; date ; may 8th} } ; eq { hop { filter_eq { all_rows ; date ; may 8th} ; tier } ; itf 25k } }*
d1	how many tier on may 8th ?
d2	itf 25k was the tier on may 8th .

Figure 2: Decomposition templates.

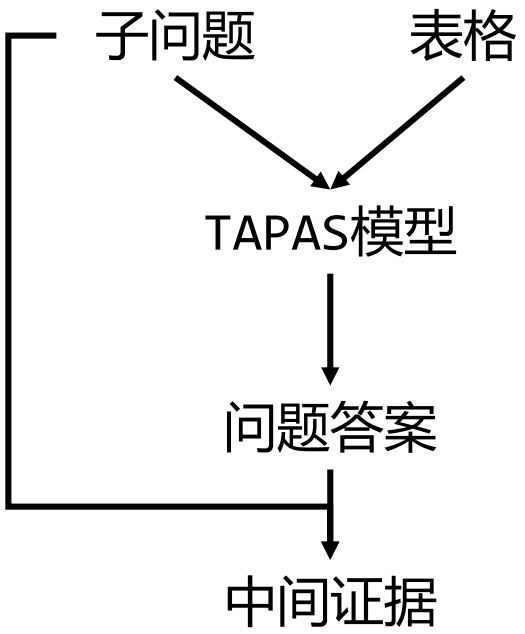
基于分解的方法——语句分解

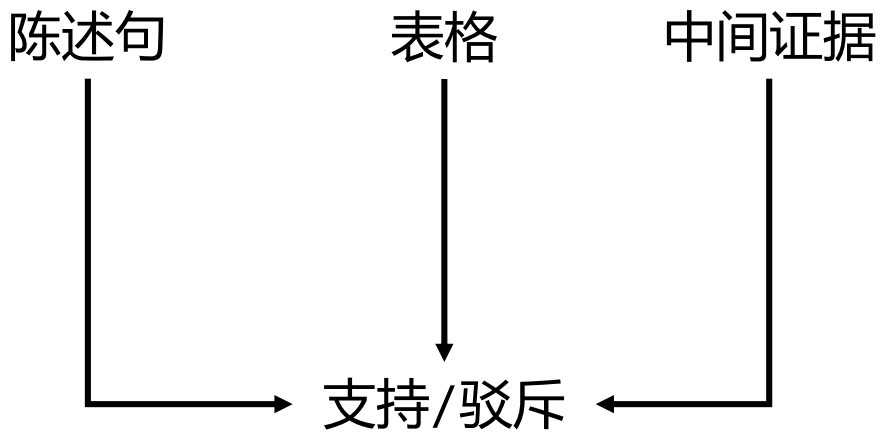
Exploring Decomposition for Table-based Fact Verification(2021)



基于分解的方法——解决子问题和重组中间证据

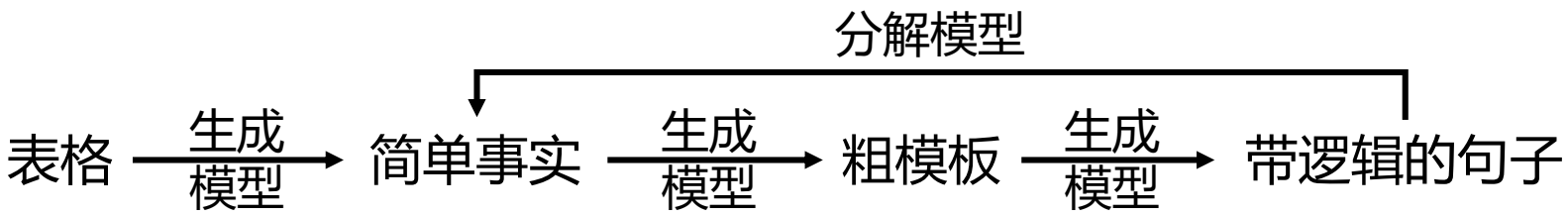
Exploring Decomposition for Table-based Fact Verification(2021)





目录

我的论文



目录

总结

1.表格生成文本：

- 任务内容
- 数据集：RotoWire、WikiBio、WeatherGOV、ToTTo、LogicNLG、Logic2Text
- 评估方法：BLEU、基于解析、基于NLI、对抗性评估

2.方法

- 由粗到细：更好地捕获逻辑依赖关系
- 基于关系实体图：用实体图捕获表间关系（ 不适用于LogicNLG数据集）
- 多样性增强：基于逻辑类型控制的句子生成（ 一个研究方向）
- 基于分解：先生成简单事实，再生成带有逻辑推理的句子（ 另一个研究方向）

谢谢