



湖南师范大学
HUNAN NORMAL UNIVERSITY

| DataFunSummit

复杂语境下的实 体关系抽取

曾道建 湖南师范大学信息科学与工程学院

zengdj@hunnu.edu.cn



目录

CONTENTS

01 任务简介

03 文档级关系抽取

02 实体关系联合抽取

04 总结与展望

■ 实体关系抽取任务介绍

- 关系定义为两个或多个实体的某种联系
- 实体关系抽取是自动识别出实体间是否存在某种关系

乔布斯和沃兹尼亚克在1976年共同创立了苹果公司。



实体：乔布斯（人），苹果（公司）
关系：创始人

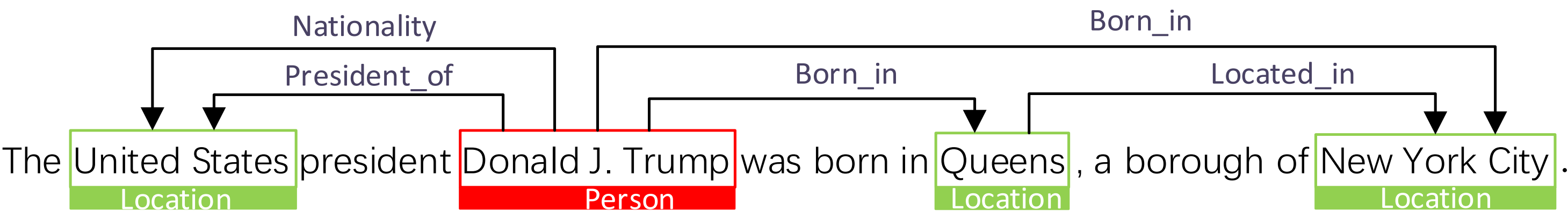


<创始人, 乔布斯, 苹果 >

复杂语境

The [introduction]_{e₁} in the [book]_{e₂} is a summary of what is in the text.

简单语境：针对一个句子中的两个实体之间的提取语义关系特征，忽略其他实体或者关系的影响



复杂语境1：同一个句子中多个三元组之间相互影响

Kungliga Hovkapellet

[1] *Kungliga Hovkapellet* (The *Royal Court Orchestra*) is a *Swedish* orchestra, originally part of the *Royal Court* in *Sweden's* capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at *Riddarhuset* in *Stockholm*. [5] Since *1773*, when the *Royal Swedish Opera* was founded by *Gustav III* of *Sweden*, the *Kungliga Hovkapellet* has been part of the opera's company.

Subject: *Kungliga Hovkapellet; Royal Court Orchestra*

Object: *Royal Swedish Opera*

Relation: **part_of**

Supporting Evidence: 5

Subject: *Riddarhuset*

Object: *Sweden*

Relation: **country**

Supporting Evidence: 1, 4

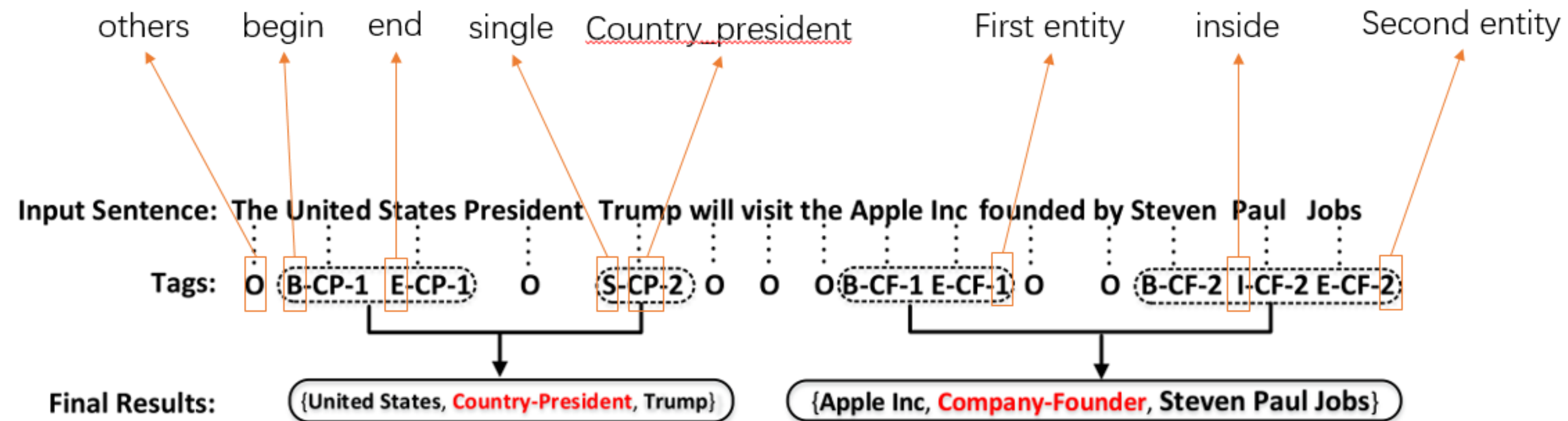
复杂语境2：大量的实体间关系是通过多个句子表达的。涉及多个实体间的跨句关系抽取

■ 实体关系联合抽取

- 序列标注
- 表填充
- 序列到序列
- ...

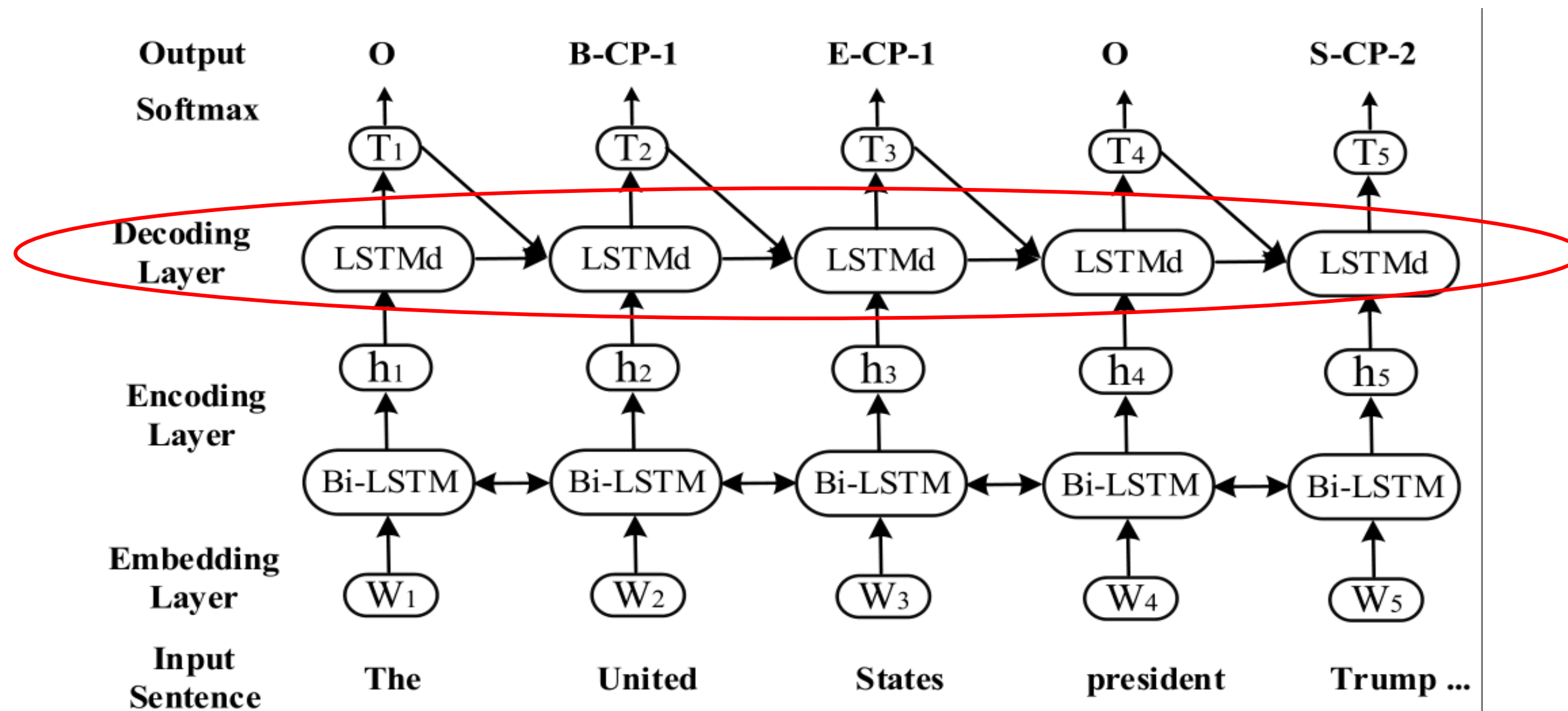
联合抽取：序列标注 (NovelTagging)

- 对每个关系，将其与 (Begin, Inside, End, Single) 以及头实体和尾实体的序号 (1, 2) 组合
- 额外考虑一个Other标签，表示不属于任何一个关系
- 如果总共有 $|R|$ 个关系，那么一共有 $2*4*|R|+1$ 个标签



联合抽取：序列标注 (NovelTagging)

● LSTM+LSTM



- ◆ CRF善于捕捉近距离的标签依赖
- ◆ LSTM可以捕捉长距离的标签依赖

● LSTM+LSTM+bias

$$L = \max \sum_{j=1}^{|\mathbb{D}|} \sum_{t=1}^{L_j} (\log(p_t^{(j)} = y_t^{(j)} | x_j, \Theta) \cdot I(O))$$

$$+ \alpha \cdot \log(p_t^{(j)} = y_t^{(j)} | x_j, \Theta) \cdot (1 - I(O))),$$

Bias weight $I(O) = \begin{cases} 1, & \text{if } tag = 'O' \\ 0, & \text{if } tag \neq 'O'. \end{cases}$

损失因子：对other标签的重要程度进行设置

联合抽取：序列标注（NovelTagging）实验结果

- 数据

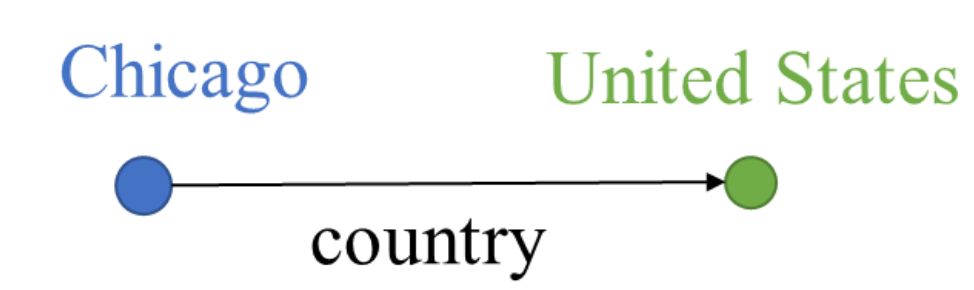
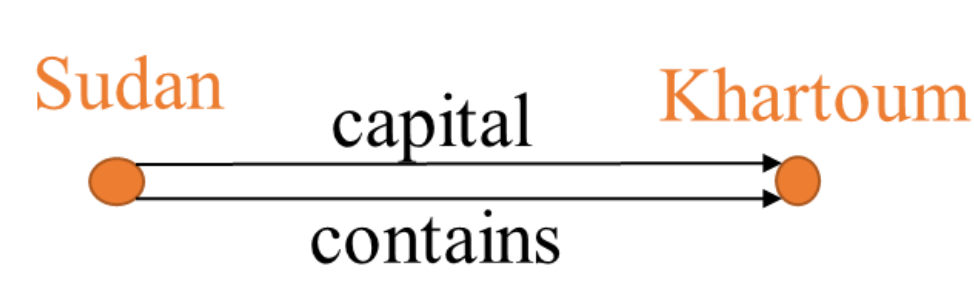
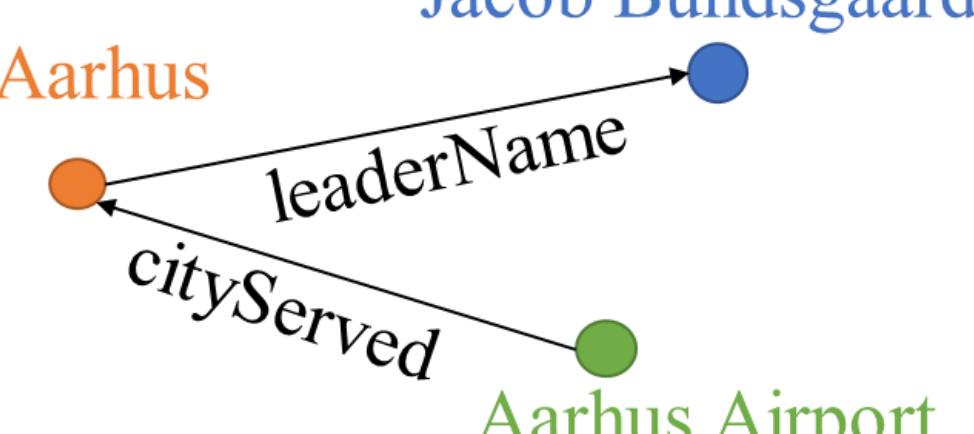
- ◆ 使用弱监督的NYT数据集，看做监督数据。
- ◆ 训练集包括353k个三元组
- ◆ 测试集包括3880个三元组
- ◆ 一共24种关系

- 实验结果

Elements	E1			E2			(E1,E2)		
PRF	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>
LSTM-CRF	0.596	0.325	0.420	0.605	0.325	0.423	0.724	0.341	0.465
LSTM-LSTM	0.593	0.342	0.434	0.619	0.334	0.434	0.705	0.340	0.458
LSTM-LSTM-Bias	0.590	0.479	0.529	0.597	0.451	0.514	0.645	0.437	0.520

联合抽取：NovelTagging缺陷

- 如何处理下面的例子

Normal	S1: Chicago is located in the United States.	
	{<Chicago, country, United States>}	
EPO	S2: News of the list's existence unnerved officials in Khartoum, Sudan's capital.	
	{<Sudan, capital, Khartoum>, <Sudan, contains, Khartoum>}	
SEO	S3: Aarhus airport serves the city of Aarhus who's leader is Jacob Bundsgaard.	
	{<Aarhus, leaderName, Jacob Bundsgaard>, <Aarhus Airport, cityServed, Aarhus>}	

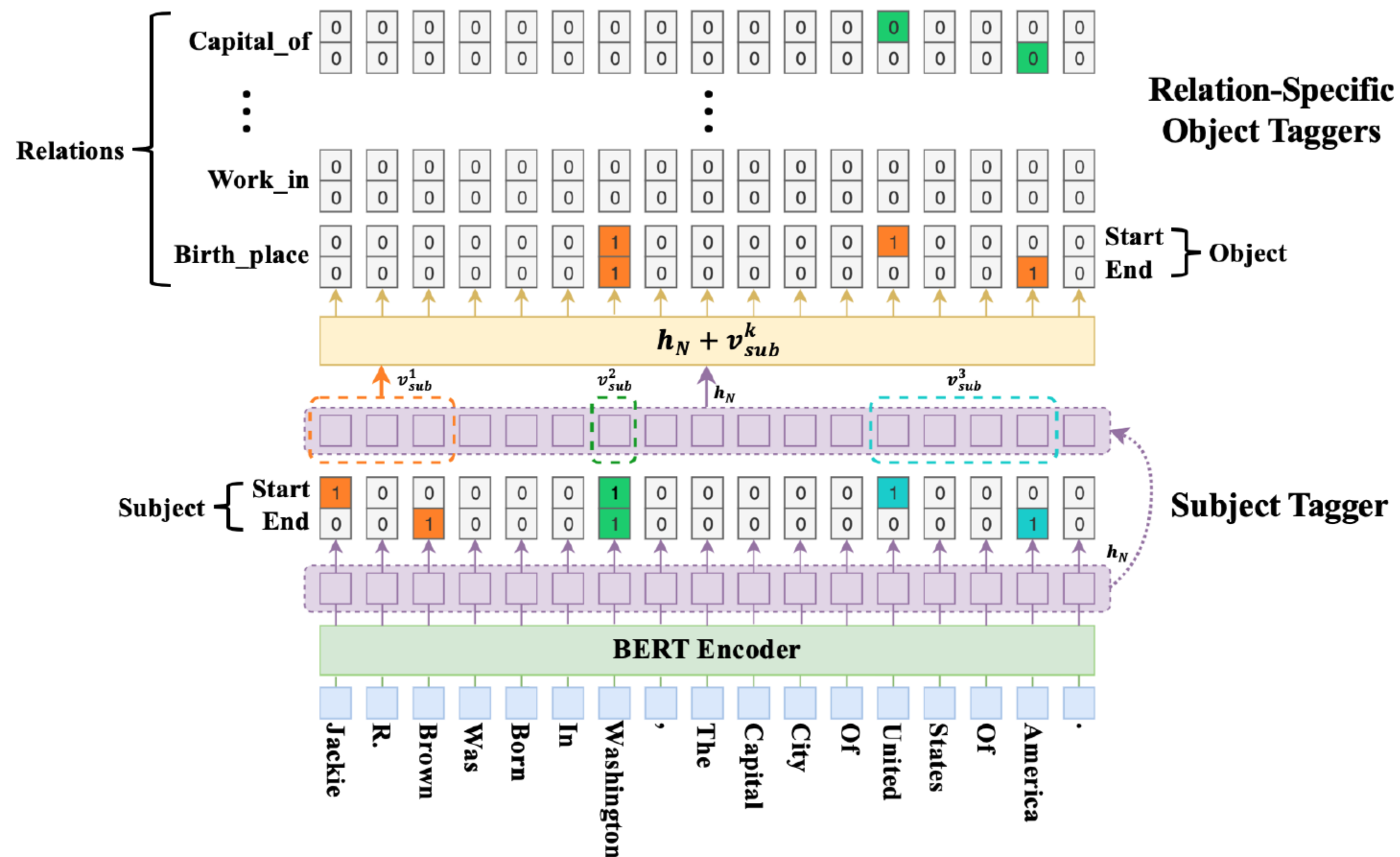
按照基于序列标注的方法，“Suda, Aarhus, Khartoum” 只能打上一个标签，因此只能参与到一个三元组中去，所以这类模型不能同时抽取EPO和SEO三元组。



如何解决
这个问题？

联合抽取：序列标注（HTB）

- 设计了一种 Hierarchical Binary Tagging 的框架
- 将三元组的抽取任务建模为三个级别的问题
- 不再将关系抽取的过程看作实体对的离散标签，而是将其看作两个实体的映射关系



联合抽取：表填充

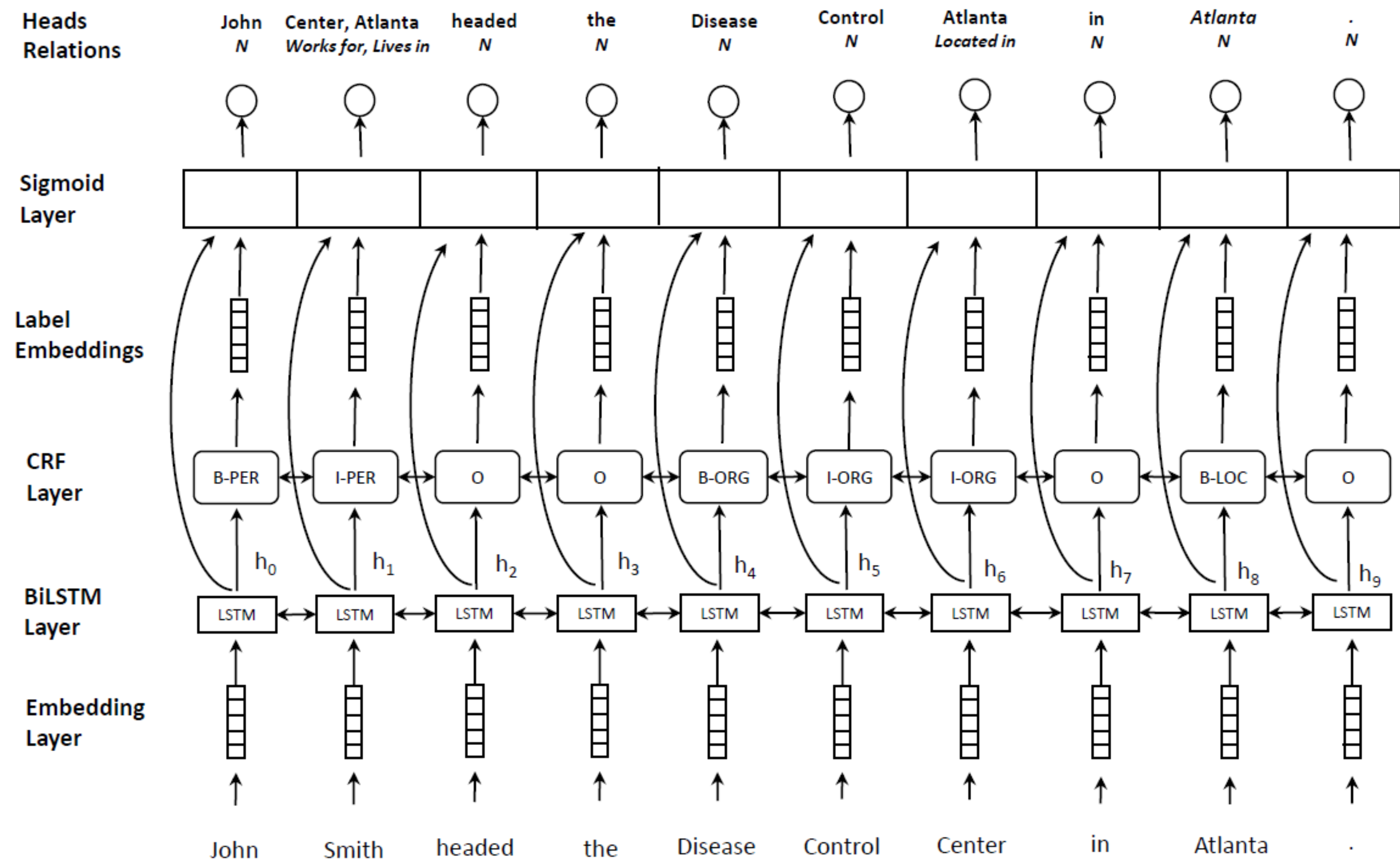
- 对角元填充实体标签
- 下三角填充关系

	Mrs.	Tsutayama	is	from	Kumamoto	Prefecture	in	Japan	.
Mrs.	B-PER								
Tsutayama	⊥	L-PER							
is	⊥	⊥	O						
from	⊥	⊥	⊥	O					
Kumamoto	⊥	⊥	⊥	⊥	B-LOC				
Prefecture	⊥	Live_in→	⊥	⊥	⊥	L-LOC			
in	⊥	⊥	⊥	⊥	⊥	⊥	O		
Japan	⊥	Live_in→	⊥	⊥	⊥	Located_in→	⊥	U-LOC	
.	⊥	⊥	⊥	⊥	⊥	⊥	⊥	⊥	⊥

表格中每个单元格只能有一个标签，不能抽取EPO三元组。

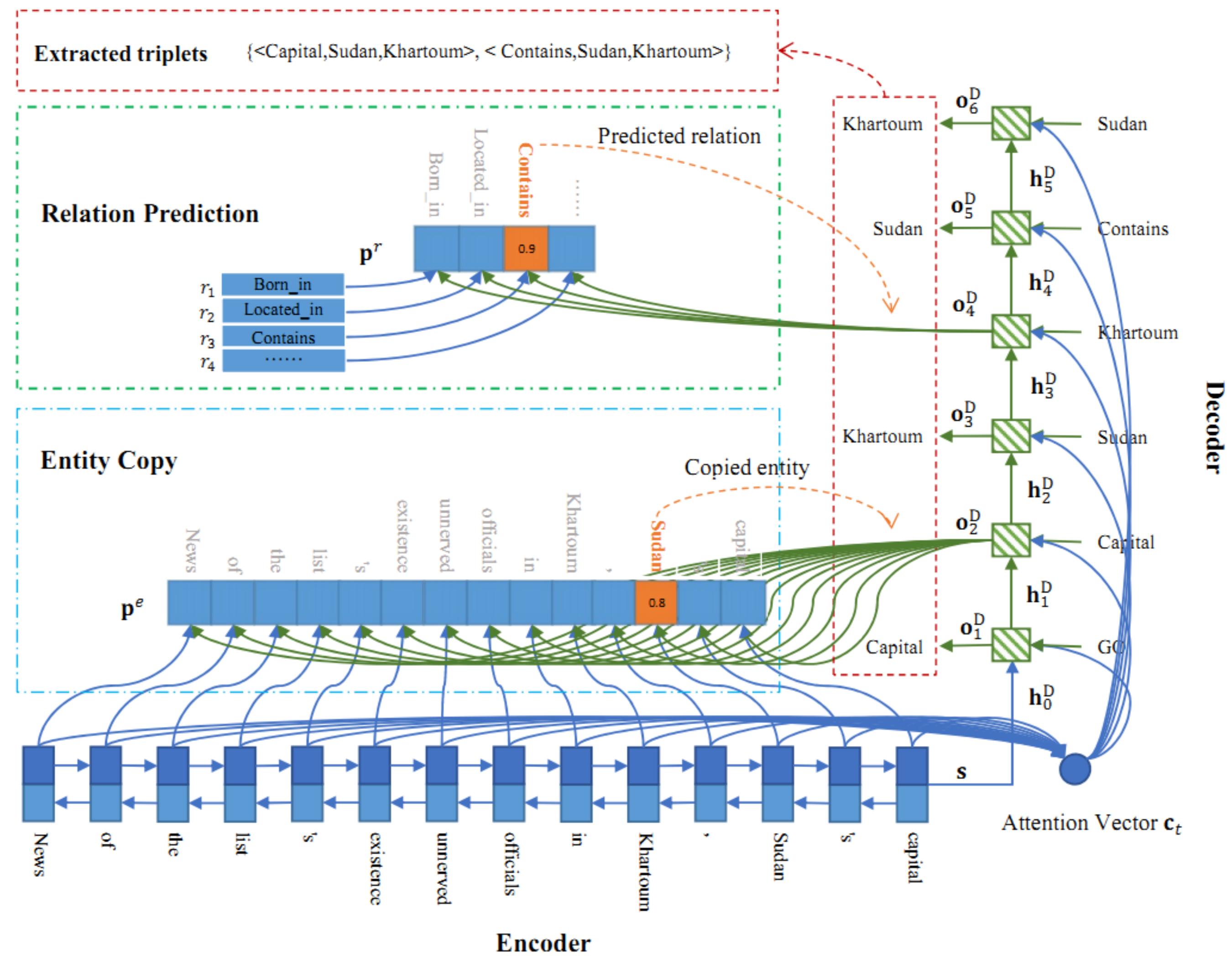
联合抽取：表填充（多头选择）

- 多任务解决实体识别问题
- 表格填充使用Sigmoid函数，允许多标签存在



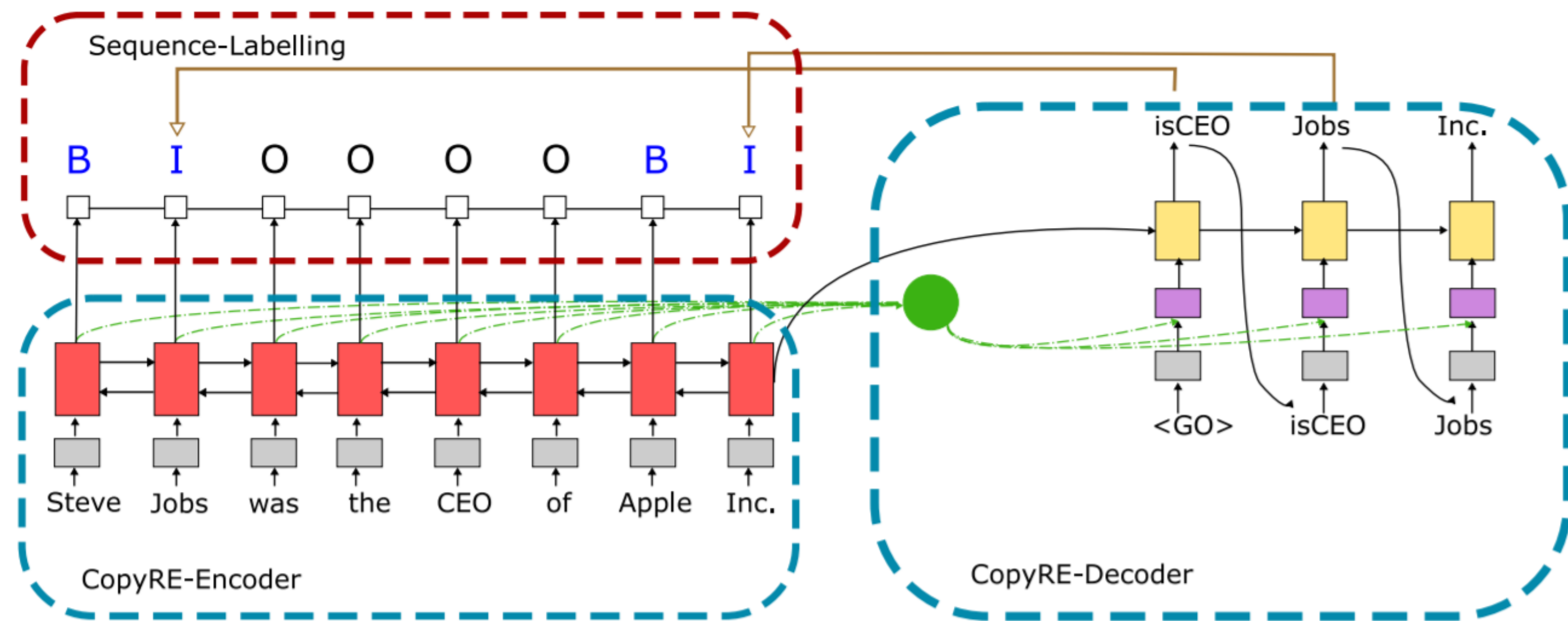
联合抽取：序列到序列 (CopyRE)

- 看作是三元组序列生成问题
- 使用拷贝机制，直接从源句子找到实体



联合抽取：序列到序列 (CopyMTL)

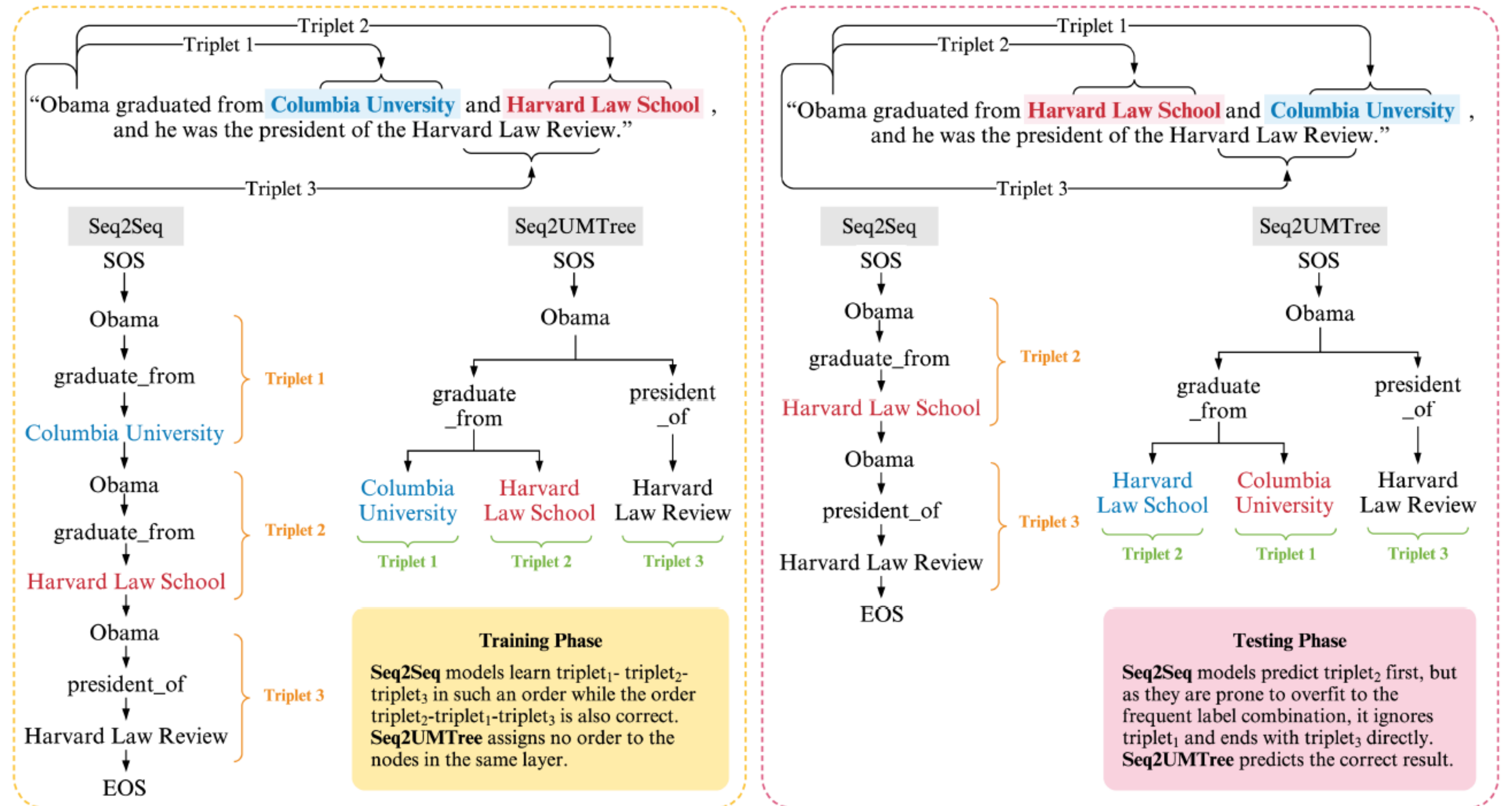
- 多任务学习，解决实体拷贝不全的问题
- 改进了拷贝函数的计算方式，增加非线性变换



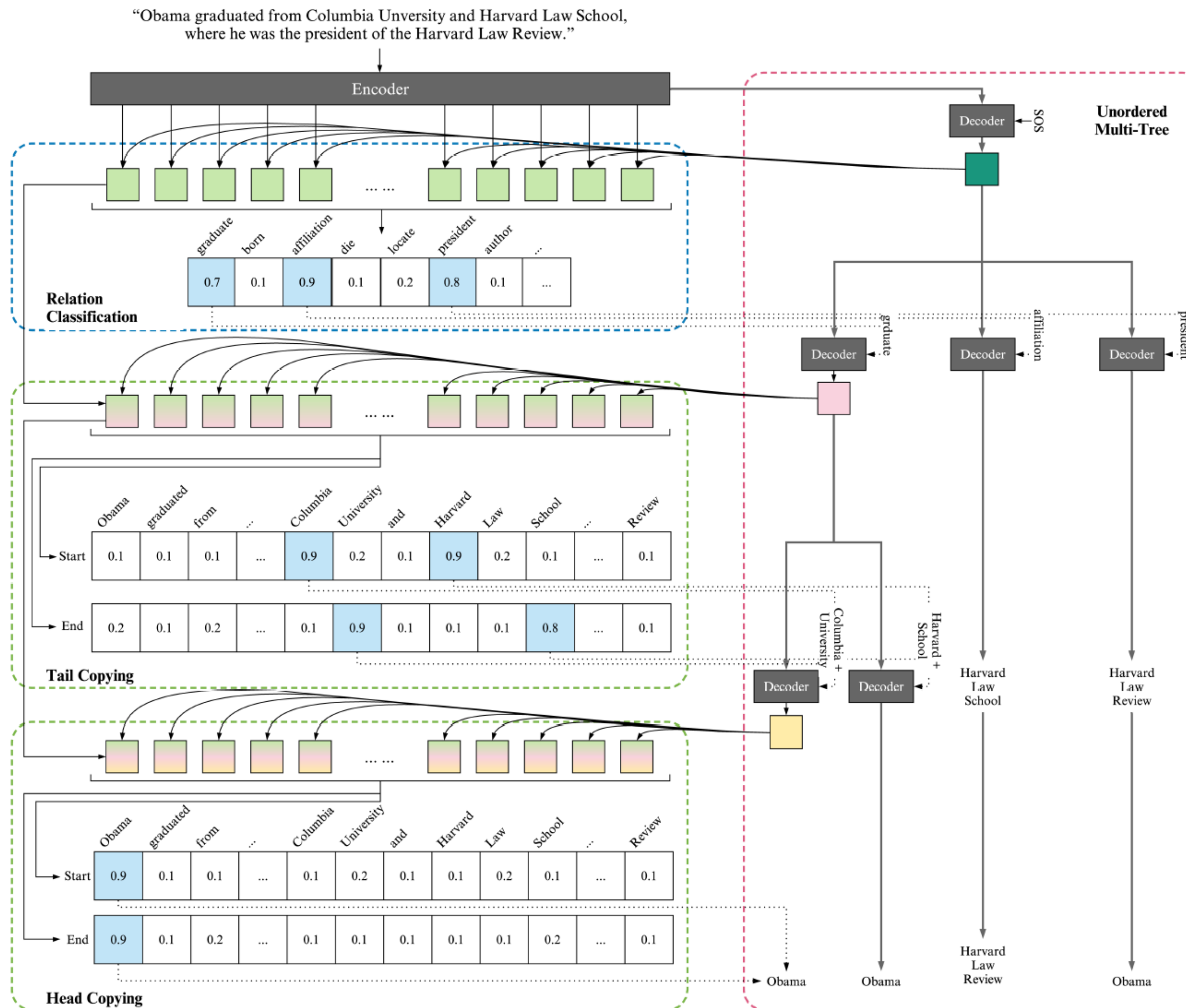
联合抽取：序列到序列 (Seq2UMTree)

● Motivation

- ◆ 自回归解码带来标记偏置问题
- ◆ 强制学习训练数据中三元组的先后顺序



联合抽取：序列到序列 (Seq2UMTree)



文档级关系抽取

- 如何有效的学习实体的多粒度表示?
 - ◆ 实体在多个句子提及
 - ◆ 实体指代
- 如何建模文档内的复杂语义信息?
 - ◆ 逻辑推理、指代推理和常识推理

Kungliga Hovkapellet

[1] *Kungliga Hovkapellet* (The *Royal Court Orchestra*) is a *Swedish* orchestra, originally part of the *Royal Court* in *Sweden*'s capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at *Riddarhuset* in *Stockholm*. [5] Since *1773*, when the *Royal Swedish Opera* was founded by *Gustav III* of *Sweden*, the *Kungliga Hovkapellet* has been part of the opera's company.

Subject: *Kungliga Hovkapellet; Royal Court Orchestra*

Object: *Royal Swedish Opera*

Relation: **part_of**

Supporting Evidence: **5**

Subject: *Riddarhuset*

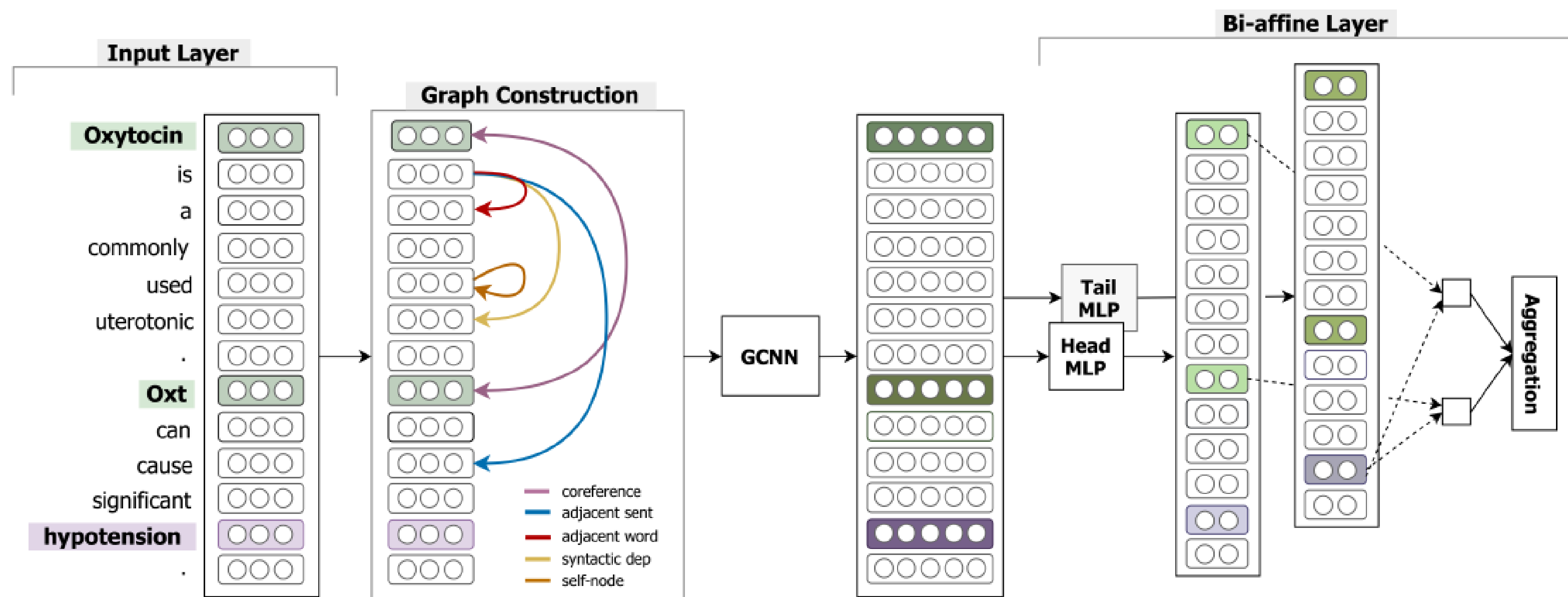
Object: *Sweden*

Relation: **country**

Supporting Evidence: **1, 4**

文档级关系抽取：GCNN

- 使用图神经网络建模文档（Word本身作为节点）
 - ◆ Syntactic dependency edge: 句法依赖，也就是使用每一个sentence中的word之间的句法关系建立edge；
 - ◆ Coreference edge: 指代，对于表示同一个含义的phrase，进行连接；
 - ◆ Adjacent sentence edge: 将sentence的根节点与上下文sentence的根节点进行连接；
 - ◆ Adjacent word edge: 对于同一个sentence，我们去连接当前word的前后节点；
 - ◆ self node edge: word与本身进行连接；
- 在构建好document graph的基础上，使用GCNN来计算得到每一个node的representation
- 多示例学习关系分类
 - ◆ 聚合target entity所有的mention



文档级关系抽取：GCNN实验结果

Data	Model	P (%)	R (%)	F1 (%)
CDR	Xu et al. (2016a)	59.6	44.0	50.7
	Zhou et al. (2016a)	64.8	49.2	56.0
	Gu et al. (2017)	60.9	59.5	60.2
	Li et al. (2018)	55.1	63.6	59.1
	Verga et al. (2018)	49.9	63.8	55.5
	CNN-RE	51.5	65.7	57.7
	RNN-RE	52.6	62.9	57.3
	GCNN	52.8	66.0	58.6
CHR	CNN-RE	81.2	87.3	84.1
	RNN-RE	83.0	90.1	86.4
	GCNN	84.7	90.5	87.5

Model	Overall	Intra	Inter
GCNN (best)	57.19	63.43	36.90
– Adjacent word	55.75	62.53	35.61
– Syntactic dependency	56.12	62.89	34.75
– Coreference	56.44	63.27	35.65
– Self-node	56.85	63.84	33.20
– Adjacent sentence	57.00	63.99	35.20

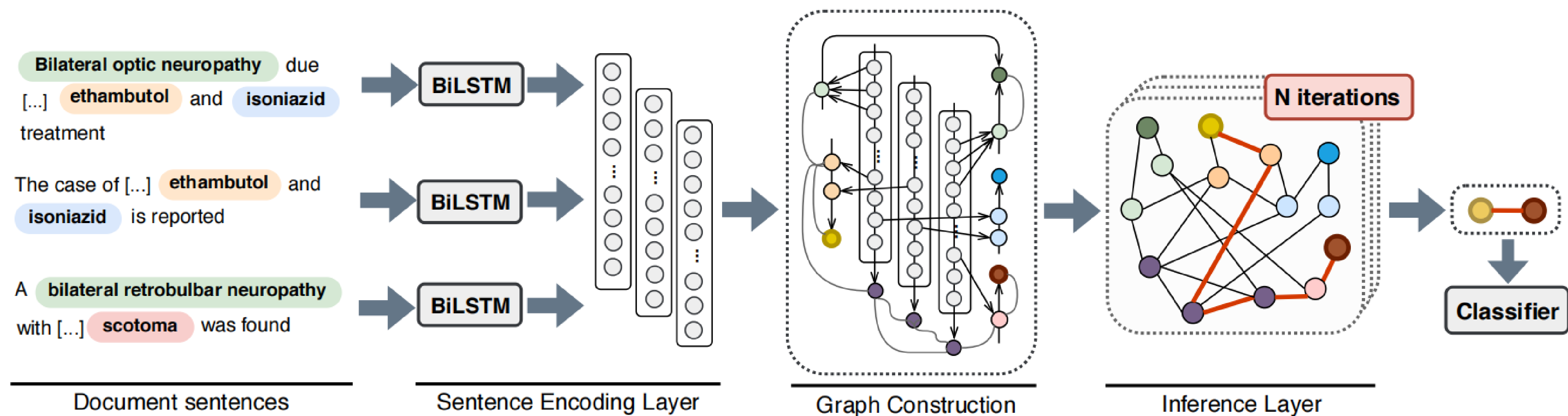
Table 3: Ablation analysis on the CDR development set, in terms of F1-score (%), for intra- (Intra) and inter-sentence (Inter) pairs.

- 消融实验结果
 - ◆ 总体看大部分构建的边对效果都有提升
 - ◆ 指代边不影响句内抽取

文档级关系抽取：EOG

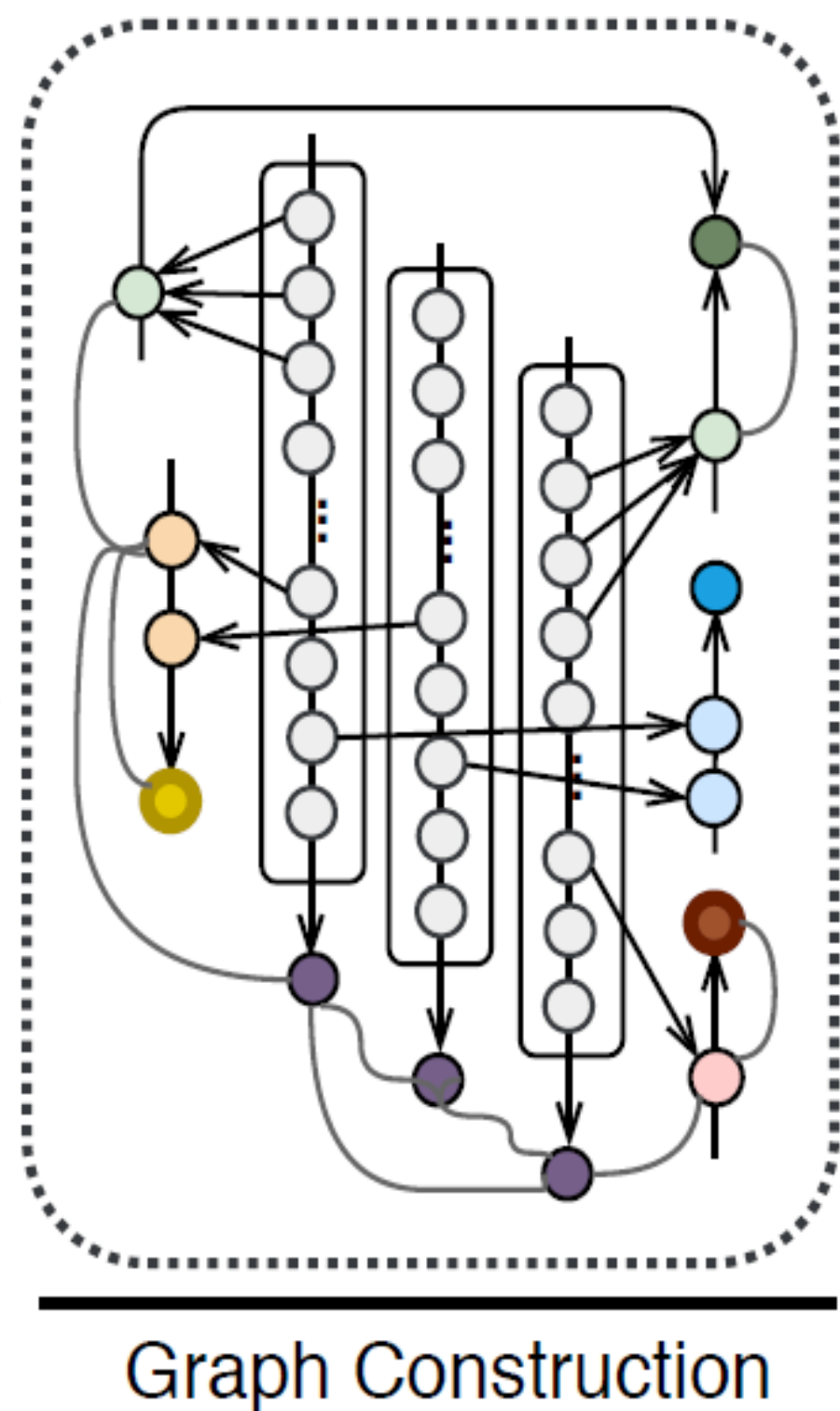
● Motivation

- ◆ 现有的方法使用基于图的模型，以实体作为节点，根据两个目标节点来确定实体间的关系。然而，实体关系可以通过节点间路径形成的唯一的边表示来更好地表达。
- ◆ EoG在不同种类节点之间，建立不同类型的边来决定信息流入节点的多少，可以更好的拟合文档之间异构的交互关系。

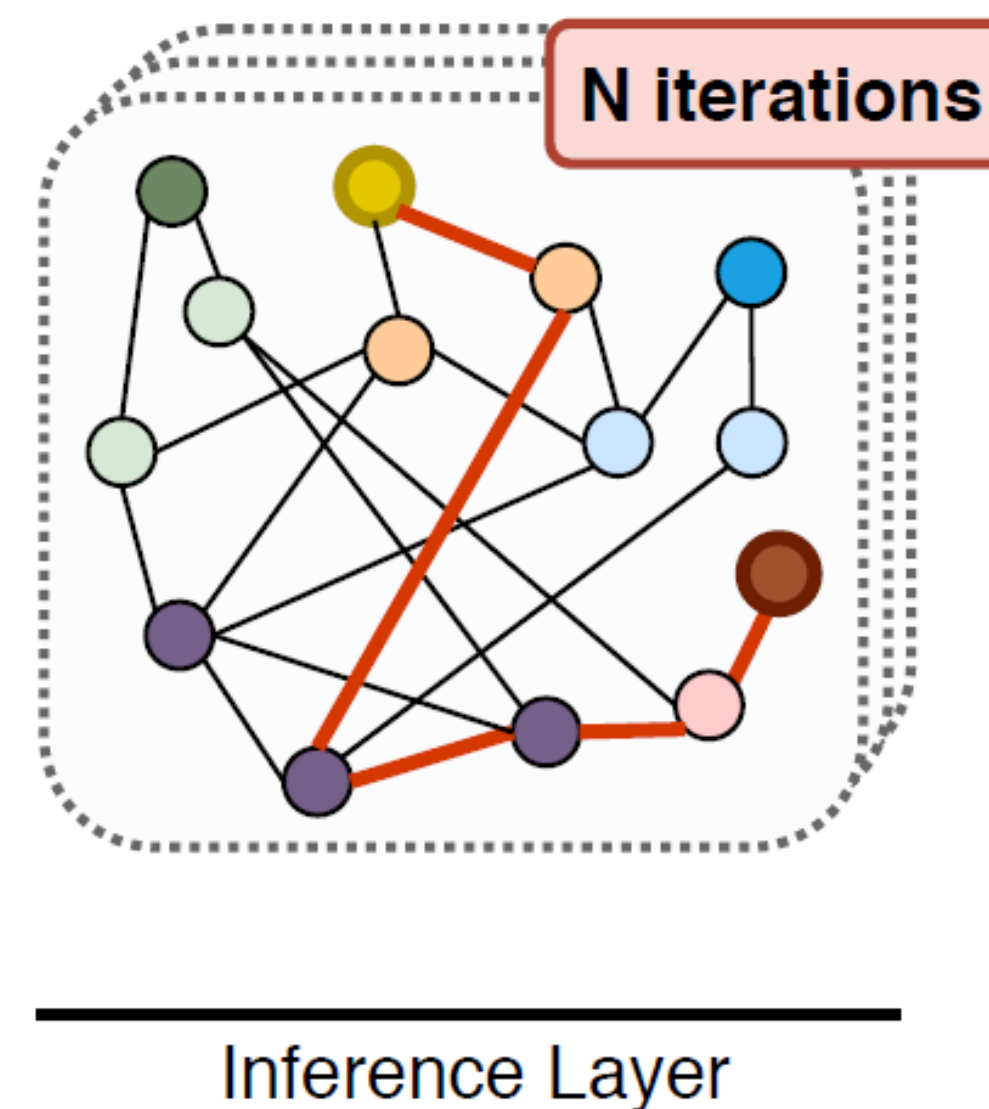


文档级关系抽取：EOG

- 文档图构建：节点构建和边构建
 - ◆ 节点：实体提及 (mention)、实体、句子
 - ◆ 边：mention-mention(MM)、mention-entity(ME)、mention-sentence(MS)、entity-sentence(ES)、sentence-sentence(SS)



- 推理：找到连接两个实体的路径，多次迭代加权路径上的信息。



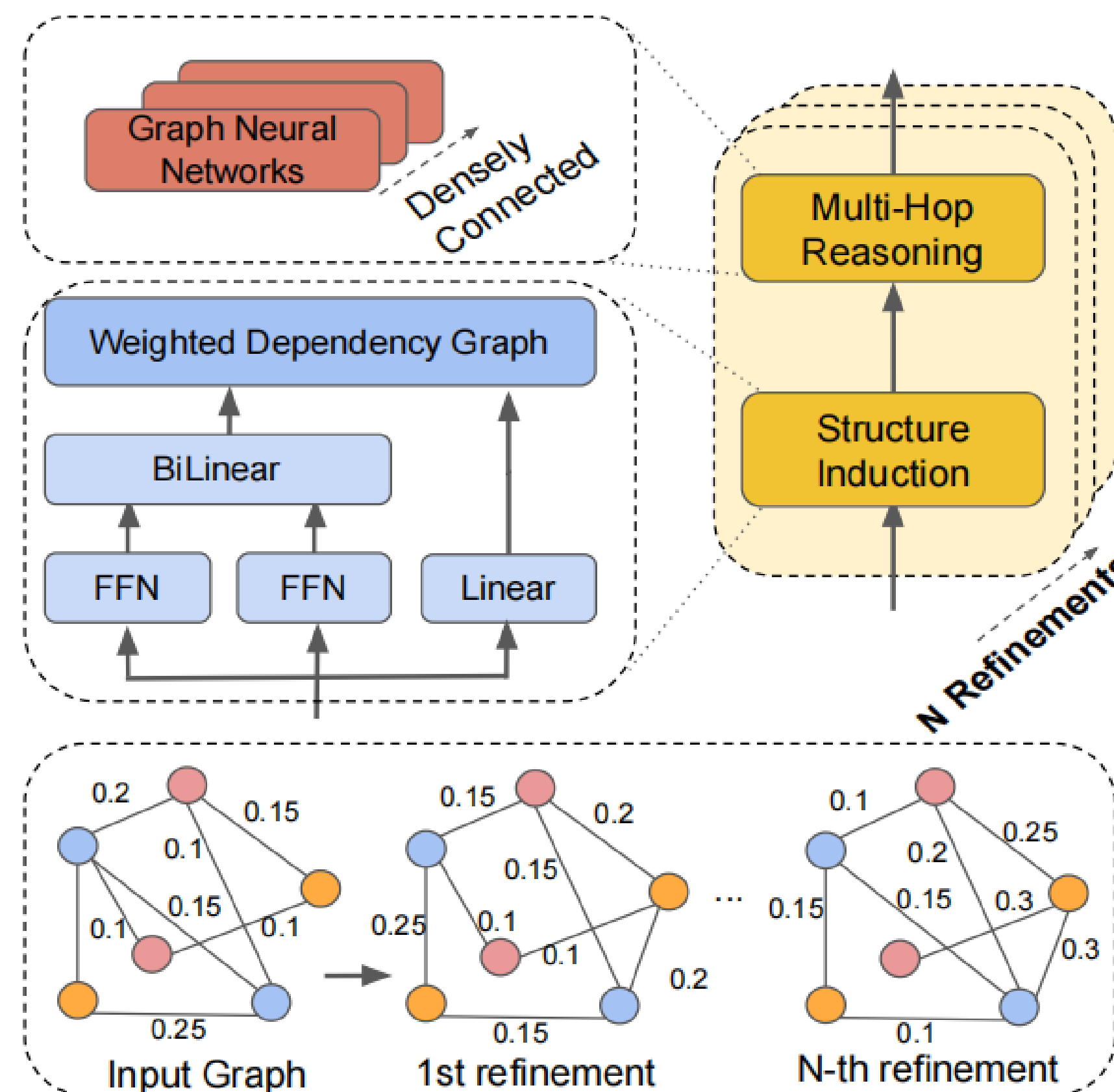
文档级关系抽取：EOG实验结果

Method	Overall (%)			Intra (%)			Inter (%)		
	P	R	F1	P	R	F1	P	R	F1
Gu et al. (2017)	55.7	68.1	61.3	59.7	55.0	57.2	51.9	7.0	11.7
Verga et al. (2018)	55.6	70.8	62.1	-	-	-	-	-	-
Nguyen and Verspoor (2018)	57.0	68.6	62.3	-	-	-	-	-	-
EoG	62.1	65.2	63.6	64.0	73.0	68.2	56.0	46.7	50.9
EoG (Full)	59.1	56.2	57.6	71.2	62.3	66.5	37.1	42.0	39.4
EoG (NoInf)	48.2	50.2	49.2	65.8	55.2	60.2	25.4	38.5	30.6
EoG (Sent)	56.9	53.5	55.2	56.9	76.4	65.2	-	-	-
Zhou et al. (2016)	55.6	68.4	61.3	-	-	-	-	-	-
Peng et al. (2016)	62.1	64.2	63.1	-	-	-	-	-	-
Li et al. (2016b)	60.8	76.4	67.7	67.3	52.4	58.9	-	-	-
Panyam et al. (2018)	53.2	69.7	60.3	54.7	80.6	65.1	47.8	43.8	45.7
Zheng et al. (2018)	56.2	67.9	61.5	-	-	-	-	-	-

文档级关系抽取：LSR

- Motivation

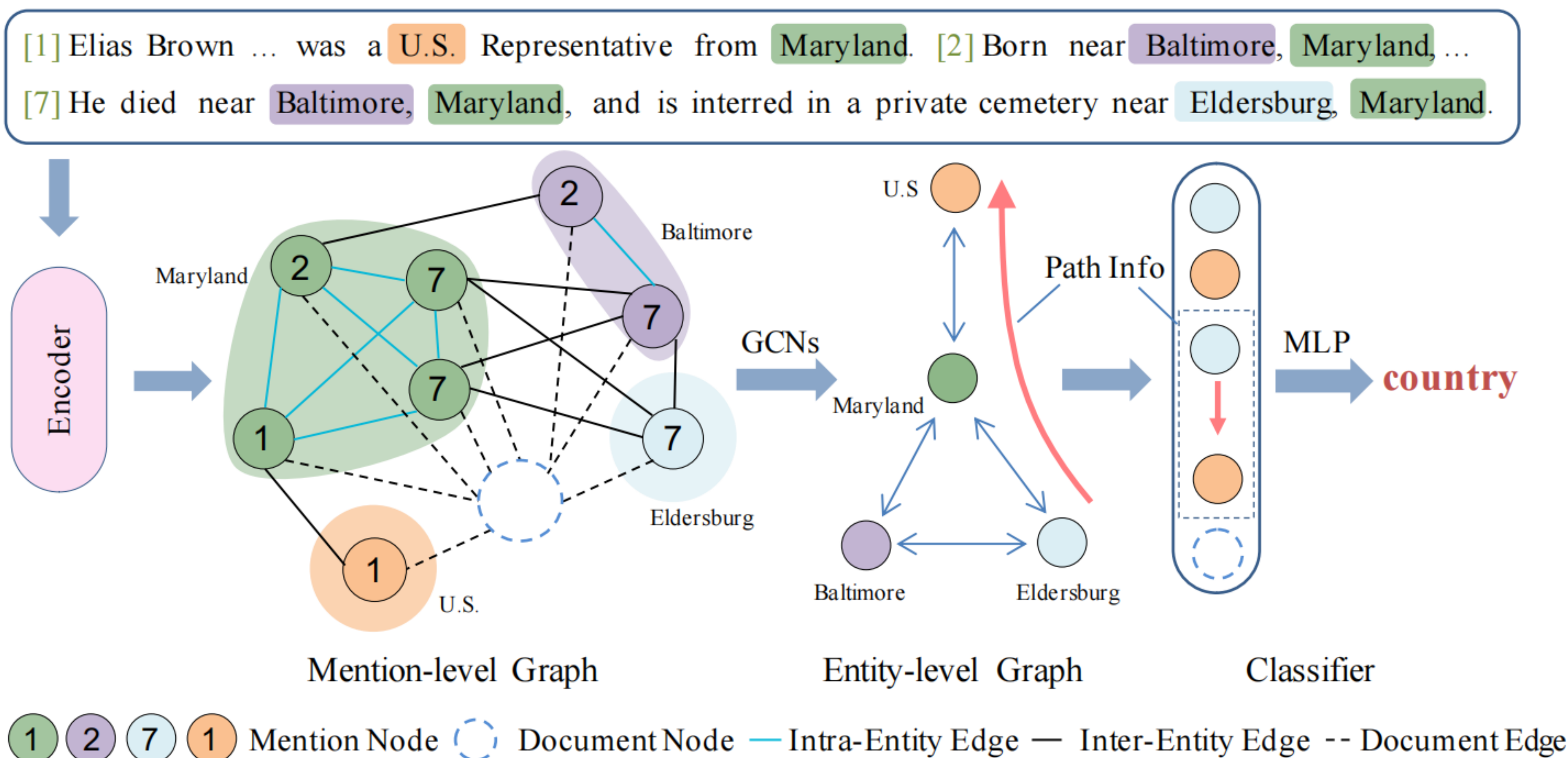
- ◆ 以往的工作大都通过启发式的方法建立文档图。
- ◆ 该文提出的模型将图结构视为一个潜在变量，并以端到端的方式对其进行归纳推理。



文档级关系抽取：LSR实验结果

Model	Dev				Test	
	Ign <i>F1</i>	<i>F1</i>	Intra- <i>F1</i>	Inter- <i>F1</i>	Ign <i>F1</i>	<i>F1</i>
CNN (Yao et al., 2019)	41.58	43.45	51.87*	37.58*	40.33	42.26
LSTM (Yao et al., 2019)	48.44	50.68	56.57*	41.47*	47.71	50.07
BiLSTM (Yao et al., 2019)	48.87	50.94	57.05*	43.49*	48.78	51.06
ContexAware (Yao et al., 2019)	48.94	51.09	56.74*	42.26*	48.40	50.70
GCNN ♣ (Sahu et al., 2019)	46.22	51.52	57.78	44.11	49.59	51.62
EoG ♣ (Christopoulou et al., 2019)	45.94	52.15	58.90	44.60	49.48	51.82
GAT ♣ (Veličković et al., 2018)	45.17	51.44	58.14	43.94	47.36	49.51
AGGCN ♣ (Guo et al., 2019a)	46.29	52.47	58.76	45.45	48.89	51.45
GloVe+LSR	48.82	55.17	60.83	48.35	52.15	54.18
BERT (Wang et al., 2019)	-	54.16	61.61*	47.15*	-	53.20
Two-Phase BERT (Wang et al., 2019)	-	54.42	61.80*	47.28*	-	53.92
BERT+LSR	52.43	59.00	65.26	52.05	56.97	59.05

文档级关系抽取: Double Graph



文档级关系抽取：Double Graph实验结果

Model	Dev				Test	
	Ign F1	Ign AUC	F1	AUC	Ign F1	F1
CNN* (Yao et al., 2019)	41.58	36.85	43.45	39.39	40.33	42.26
LSTM* (Yao et al., 2019)	48.44	46.62	50.68	49.48	47.71	50.07
BiLSTM* (Yao et al., 2019)	48.87	47.61	50.94	50.26	48.78	51.06
Context-Aware* (Yao et al., 2019)	48.94	47.22	51.09	50.17	48.40	50.70
HIN-GloVe* (Tang et al., 2020)	51.06	-	52.95	-	51.15	53.30
GAT [‡] (Velickovic et al., 2017)	45.17	-	51.44	-	47.36	49.51
GCNN [‡] (Sahu et al., 2019)	46.22	-	51.52	-	49.59	51.62
EoG [‡] (Christopoulou et al., 2019)	45.94	-	52.15	-	49.48	51.82
AGGCN [‡] (Guo et al., 2019)	46.29	-	52.47	-	48.89	51.45
LSR-GloVe* (Nan et al., 2020)	48.82	-	55.17	-	52.15	54.18
GAIN-GloVe	53.05	52.57	55.29	55.44	52.66	55.08
BERT-RE _{base} * (Wang et al., 2019a)	-	-	54.16	-	-	53.20
RoBERTa-RE _{base} [†]	53.85	48.27	56.05	51.35	53.52	55.77
BERT-Two-Step _{base} * (Wang et al., 2019a)	-	-	54.42	-	-	53.92
HIN-BERT _{base} * (Tang et al., 2020)	54.29	-	56.31	-	53.70	55.60
CorefBERT-RE _{base} * (Ye et al., 2020)	55.32	-	57.51	-	54.54	56.96
LSR-BERT _{base} * (Nan et al., 2020)	52.43	-	59.00	-	56.97	59.05
GAIN-BERT _{base}	59.14	57.76	61.22	60.96	59.00	61.24
BERT-RE _{large} * (Ye et al., 2020)	56.67	-	58.83	-	56.47	58.69
CorefBERT-RE _{large} * (Ye et al., 2020)	56.73	-	58.88	-	56.48	58.70
RoBERTa-RE _{large} * (Ye et al., 2020)	57.14	-	59.22	-	57.51	59.62
CorefRoBERTa-RE _{large} * (Ye et al., 2020)	57.84	-	59.93	-	57.68	59.91
GAIN-BERT _{large}	60.87	61.79	63.09	64.75	60.31	62.76

总结与展望

- 联合抽取
 - ◆ 序列到序列方法 =》 序列到集合
- 文档级别抽取
 - ◆ Mention、实体、句子级别的信息传递=》 实体对级别信息
 - ◆ GNN过平滑、异构图



湖南師範大學
HUNAN NORMAL UNIVERSITY

| DataFunSummit

THANKS!

今天的分享就到这里...

zengdj@hunnu.edu.cn

Ending

