Datawhale paper

Double Graph Based Reasoning for Document-level Relation Extraction (EMNLP 2020)

报告人：陈海顺
时间：2020.11.21

Background

西安电子科技大学坐落于西安市雁塔区

位于

西安电子科技大学

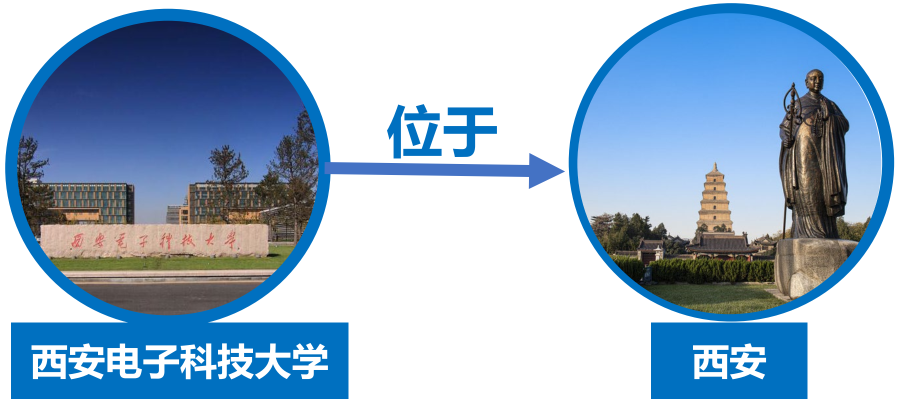
西安

文本中蕴含着大量的事实知识

无结构化 → 结构化

从**文本**中**自动化的抽取**
事实知识具有重大意义

Background



三元组 表示实体间关系 （**西安电子科技大学，位于，西安**）

Background

许多事实知识隐藏在更复杂的上下文中，比如段落和文档当中。

位于西安的西安电子科技大学是中华人民共和国教育部直属高校。其现任校长是杨宗凯教授。

A Case

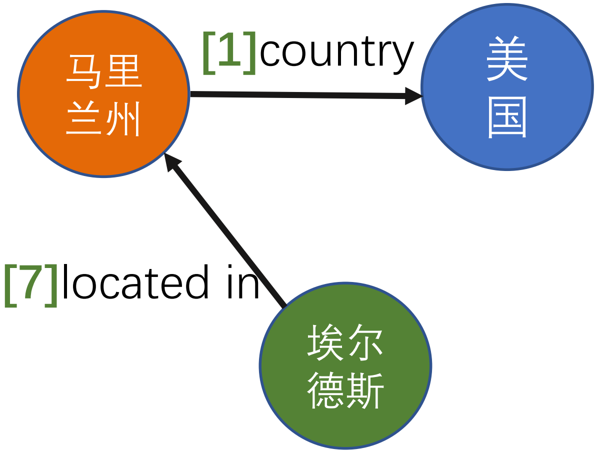
Elias Brown
[1] *Elias Brown* (*May 9, 1793– July 7, 1857*) was a ***U.S.*** Representative from ***Maryland***. [2] Born near ***Baltimore***, ***Maryland***, *Brown* attended the common schools. … [7] He died near ***Baltimore***, ***Maryland***, and is interred in a private cemetery near ***Eldersburg***, ***Maryland***.

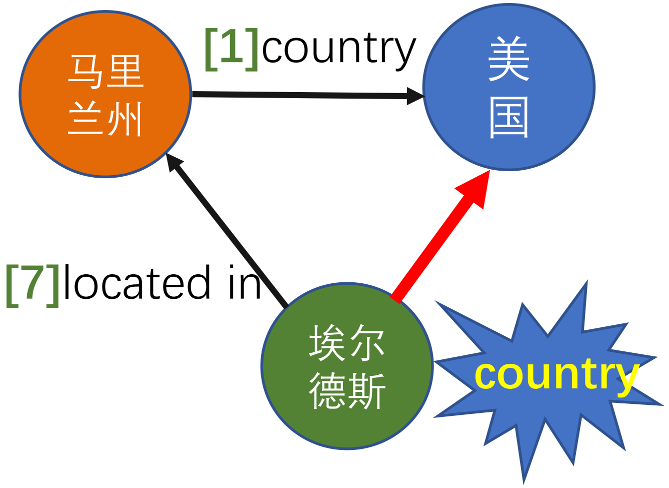
Subject: ***Maryland***
Object: ***U.S.***
relation: **country**

Subject: ***Baltimore；Eldersburg***
Object: ***Maryland***
relation: **located in the administrative territorial entity**

Subject: ***Baltimore；Eldersburg***
Object: ***U.S.***
relation: **country**

[1] 埃利亚斯-布朗 (1793.5.9-1857.7.7)曾是一名来自马里兰州的美国代表。[2] 布朗出生于马里兰州的巴尔的摩市，就读于一所普通学校 ….[5]….[6]….[7] 他于马里兰州的巴尔的摩市逝世,安葬在马里兰州的埃尔德斯堡附近的私人墓地中。





| Background

Dataset	# Doc.	# Word	# Sent.	# Ent.	# Rel.	# Inst.	# Fact
SemEval-2010 Task 8	-	205k	10,717	21,434	9	8,853	8,383
ACE 2003-2004	-	297k	12,783	46,108	24	16,771	16,536
TACRED	-	1,823k	53,791	152,527	41	21,773	5,976
FewRel	-	1,397k	56,109	72,124	100	70,000	55,803
BC5CDR	1,500	282k	11,089	29,271	1	3,116	2,434
DocRED (Human-annotated)	5,053	1,002k	40,276	132,375	96	63,427	56,354
DocRED (Distantly Supervised)	101,873	21,368k	828,115	2,558,350	96	1,508,320	881,298

https://github.com/thunlp/DocRED

Baseline

Given a document : $\mathcal{D} = \{w_i\}_{i=1}^{n}$

Word feature:

$$\{h_i\}_{i=1}^{n} = encoding\{embedding_{word_i}\}_{i=1}^{n}$$

Named entity mention representation :

$$e_i = \frac{1}{K}\sum_{i=1}^{K} m_k$$
$$m_k = \frac{1}{t-s+1}\sum_{j=t}^{s} h_j$$

Prediction :

$$\tilde{e}_i = [e_i; E(d_{ij})], \tilde{e}_j = [e_j; E(d_{ji})]$$
$$P(r \mid e_i, e_j) = \mathrm{sigmoid}\left(\tilde{e}_i^T W_r \tilde{e}_j + b_r\right)$$

Baseline

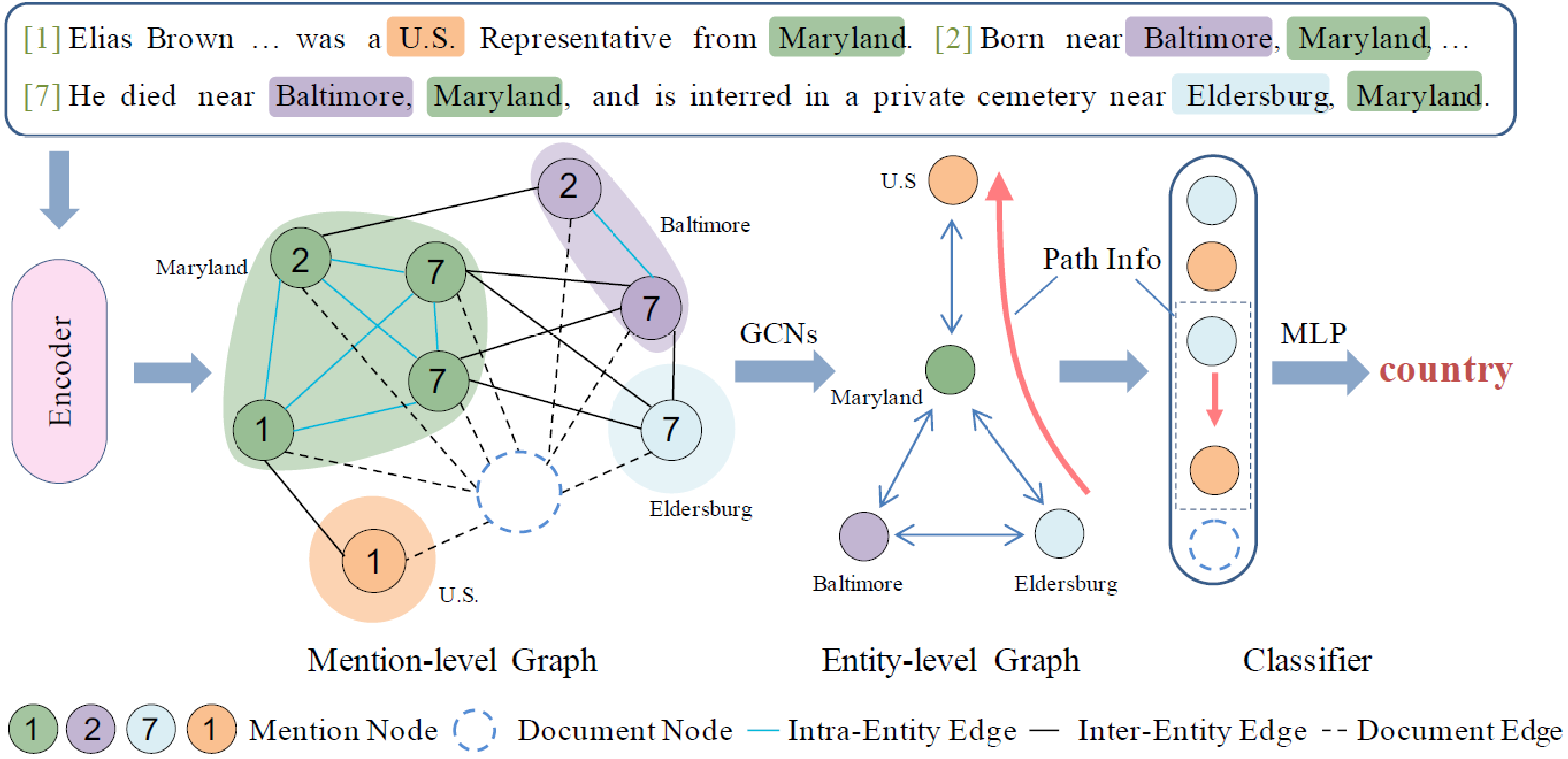
Model	Dev Ign F1	Dev Ign AUC	Dev F1	Dev AUC	Test Ign F1	Test Ign AUC	Test F1	Test AUC
Supervised Setting								
CNN	37.99	31.47	43.45	39.41	36.44	30.44	42.33	38.98
LSTM	44.41	39.78	50.66	49.48	43.60	39.02	50.12	49.31
BiLSTM	**45.12**	**40.93**	50.95	**50.27**	**44.73**	**40.40**	**51.06**	**50.43**
Context-Aware	44.84	40.42	**51.10**	50.20	43.93	39.30	50.64	49.70
Weakly Supervised Setting								
CNN	26.35	14.18	42.75	38.01	25.40	13.46	42.02	36.86
LSTM	30.86	15.62	49.91	42.78	29.75	14.97	49.91	42.78
BiLSTM	32.05	**16.50**	**51.72**	**44.42**	29.96	**15.50**	49.82	**42.90**
Context-Aware	**32.43**	15.86	51.39	43.02	**30.27**	15.11	**50.14**	41.52

Table 4: Performance of different RE models on DocRED (%).

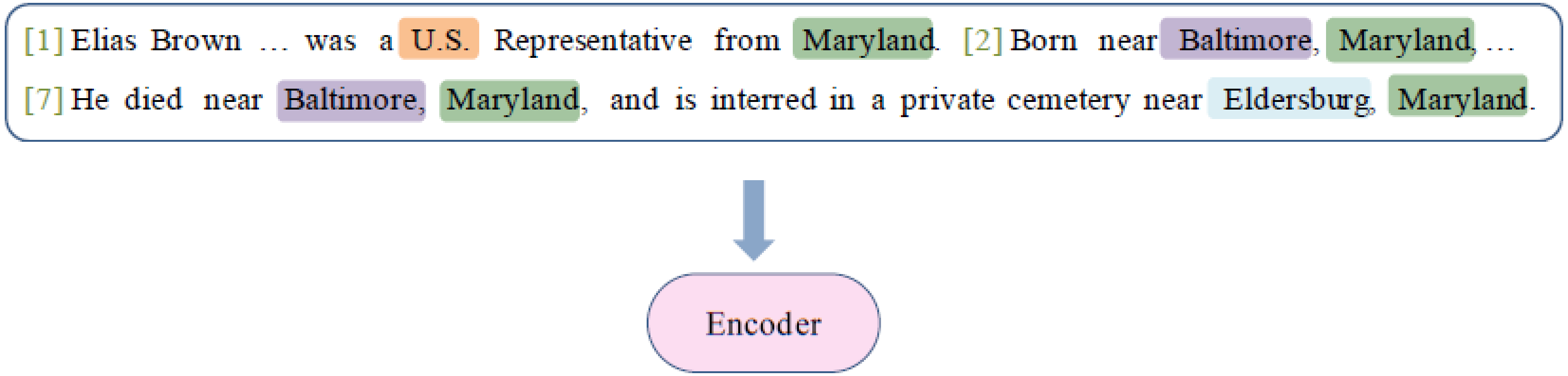
Method	RE P	RE R	RE F1	RE+Sup P	RE+Sup R	RE+Sup F1
Model	55.6	52.6	54.1	46.4	43.1	44.7
Human	**89.7**	**86.3**	**88.0**	**71.2**	**75.8**	**73.4**

Table 5: Human performance (%).

| Model


[1] Elias Brown … was a U.S. Representative from Maryland. [2] Born near Baltimore, Maryland, …
[7] He died near Baltimore, Maryland, and is interred in a private cemetery near Eldersburg, Maryland.
Encoder
Maryland
Baltimore
Eldersburg
U.S.
GCNs
U.S
Maryland
Baltimore
Eldersburg
Path Info
MLP
country
Mention-level Graph
Entity-level Graph
Classifier
Mention Node
Document Node
Intra-Entity Edge
Inter-Entity Edge
Document Edge

Encoding Module

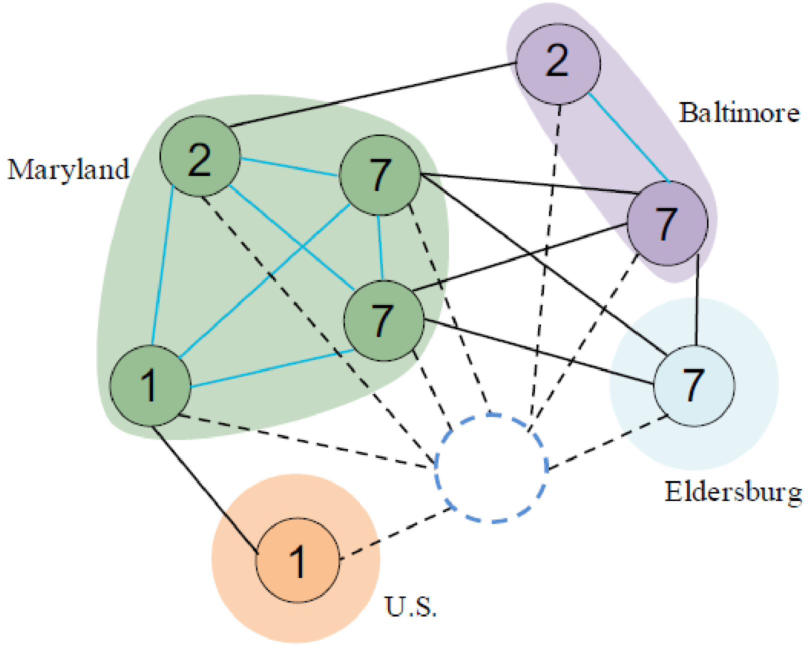


$$x_i = [E_w(w_i); E_t(t_i); E_c(c_i)] \qquad [g_1, g_2, \ldots, g_n] = Encoder([x_1, x_2, \ldots, x_n]) \tag{2}$$

[Word embedding; Entity type embedding; Coreference embedding]

Mention-level Graph Aggregation Module

[1] Elias Brown … was a U.S. Representative from Maryland. [2] Born near Baltimore, Maryland, …
[7] He died near Baltimore, Maryland, and is interred in a private cemetery near Eldersburg, Maryland.



Mention-level Graph

两种节点：提及节点，文档节点

三种类型边：
Intra-Entity Edge: 连接同一实体提及；
Inter-Entity Edge：同一句话下的不同实体提及；
Document Edge：所有的提及都连接到文档节点

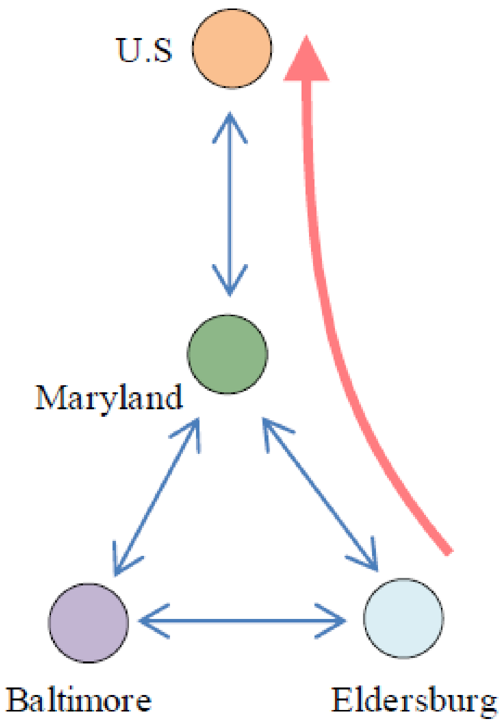
GCN：

$$h_u^{(l+1)} = \sigma\left(\sum_{k\in\mathcal{K}}\sum_{v\in\mathcal{N}_k(u)} W_k^{(l)} h_v^{(l)} + b_k^{(l)}\right)$$

$$\mathbf{m}_u = [h_u^{(0)}; h_u^{(1)}; \ldots; h_u^{(N)}]$$

1 2 7 1 Mention Node Document Node — Intra-Entity Edge — Inter-Entity Edge -- Document Edge

Entity-level Graph Inference Module



Entity-level Graph

Entity representation: $$\mathbf{e}_i = \frac{1}{N}\sum_n \mathbf{m}_n$$

Edge representation: $$\mathbf{e}_{ij} = \sigma\left(W_q[\mathbf{e}_i; \mathbf{e}_j] + b_q\right)$$

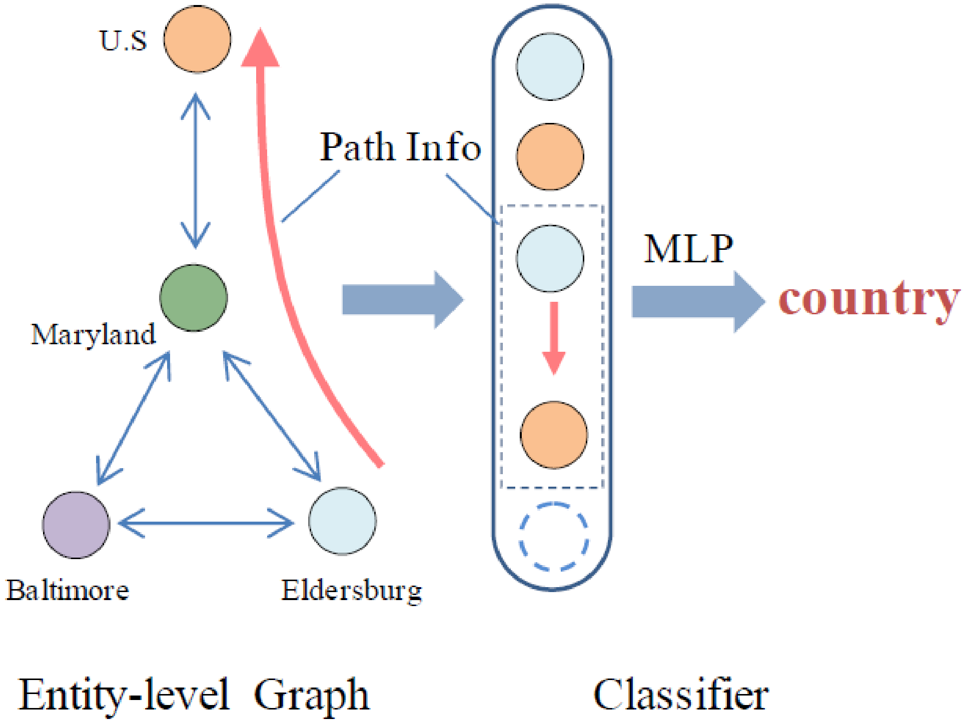
Two-hop path： $$\mathbf{p}^i_{h,t} = [\mathbf{e}_{ho}; \mathbf{e}_{ot}; \mathbf{e}_{to}; \mathbf{e}_{oh}]$$

Attention mechanism $$s_i = \sigma([\mathbf{e}_h; \mathbf{e}_t] \cdot W_l \cdot \mathbf{p}^i_{h,t})$$

$$\alpha_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$\mathbf{p}_{h,t} = \sum \alpha_i \mathbf{p}^i_{h,t}$$

Classification Module



关系向量的表示:

$$I_{h,t} = [\mathbf{e}_h; \mathbf{e}_t; |\mathbf{e}_h - \mathbf{e}_t|; \mathbf{e}_h \odot \mathbf{e}_t; \mathbf{m}_{doc}; \mathbf{p}_{h,t}]$$

$$P(r|\mathbf{e}_h, \mathbf{e}_t) = sigmoid\left(W_b \sigma(W_a I_{h,t} + b_a) + b_b\right) \tag{12}$$

Experiment

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$^{\dagger}_{base}$	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**

Experiment

Model	Dev				Test	
	Ign F1	Ign AUC	F1	AUC	Ign F1	F1
GAIN-GloVe	**53.05**	**52.57**	**55.29**	**55.44**	**52.66**	**55.08**
- *hMG*	50.97 ↓	48.84	53.10	51.73	50.76	53.06
- *Inference Module*	50.84 ↓	48.68	53.02	51.58	50.32	52.66
- *Document Node*	50.86 ↓	48.68	53.01	52.46	50.32	52.67
GAIN-BERT$_{base}$	**59.14**	**57.76**	**61.22**	**60.96**	**59.00**	**61.24**
- *hMG*	57.12	51.54	59.17	54.61	57.31	59.56
- *Inference Module*	56.97	54.29	59.28	57.25	57.01	59.34
- *Document Node*	57.26	52.07	59.62	55.51	57.01	59.63

Table 3: Performance of GAIN with different embeddings and submodules.

Thanks!