组会

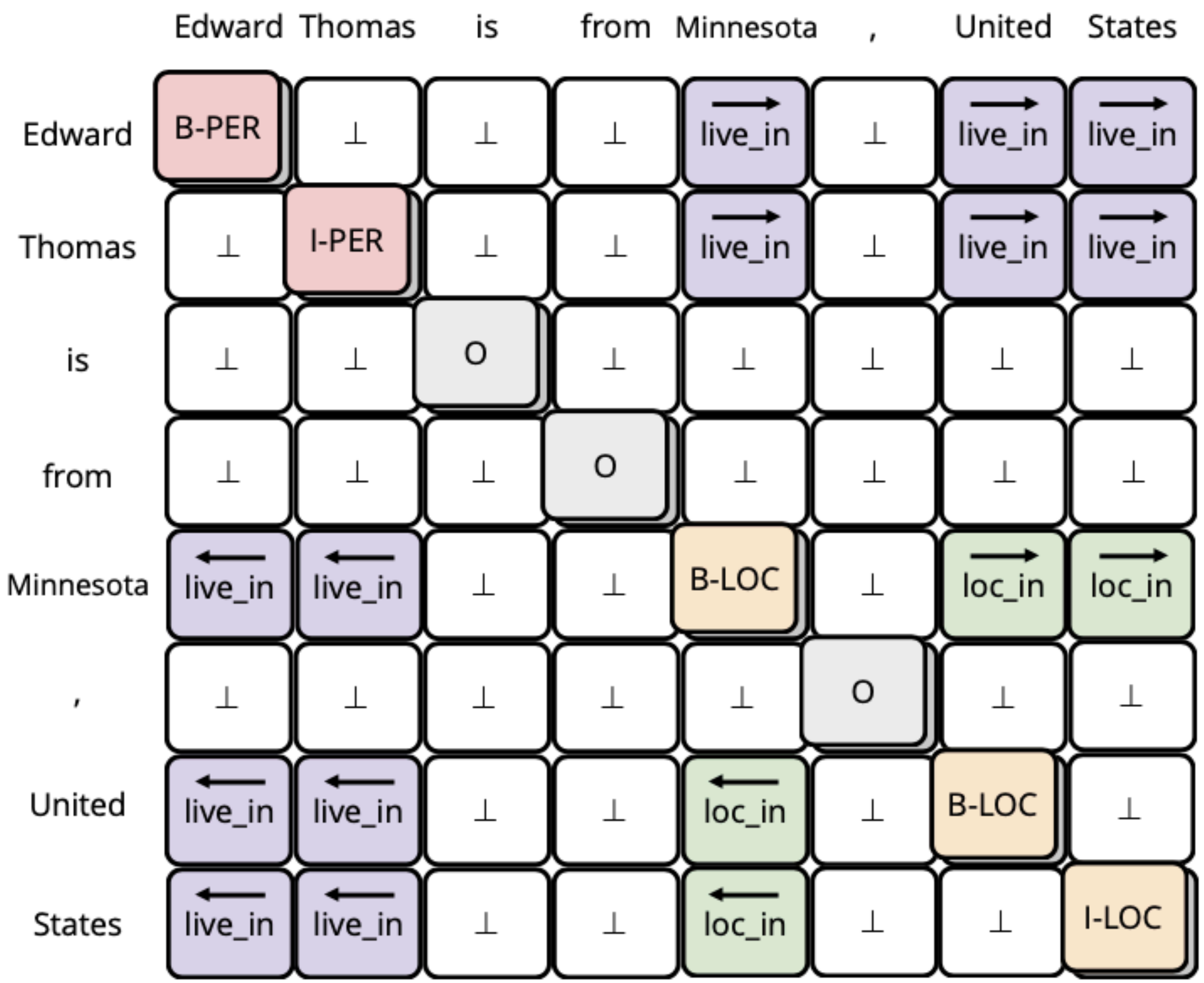
林佳兴

2020-12-12

EMNLP 2020

Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders

Jue Wang[1] and **Wei Lu**[2]
[1]College of Computer Science and Technology, Zhejiang University
[2]StatNLP Research Group, Singapore University of Technology and Design
zjuwangjue@zju.edu.cn, luwei@sutd.edu.sg



Edward Thomas is from Minnesota , United States
Edward B-PER ⊥ ⊥ ⊥ live_in ⊥ live_in live_in
Thomas ⊥ I-PER ⊥ ⊥ live_in ⊥ live_in live_in
is ⊥ ⊥ O ⊥ ⊥ ⊥ ⊥ ⊥
from ⊥ ⊥ ⊥ O ⊥ ⊥ ⊥ ⊥
Minnesota live_in live_in ⊥ ⊥ B-LOC ⊥ loc_in loc_in
, ⊥ ⊥ ⊥ ⊥ ⊥ O ⊥ ⊥
United live_in live_in ⊥ ⊥ loc_in ⊥ B-LOC ⊥
States live_in live_in ⊥ ⊥ loc_in ⊥ ⊥ I-LOC

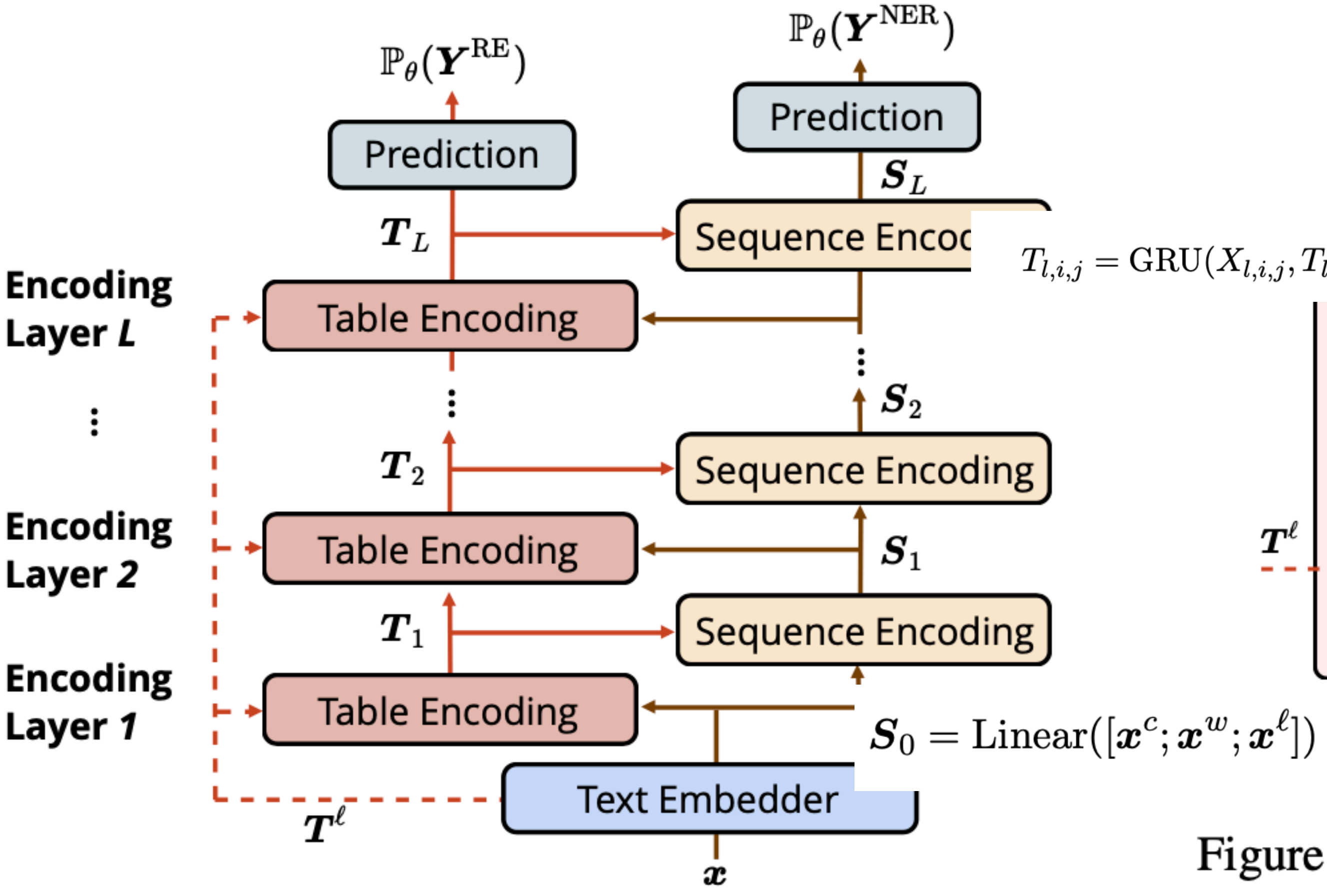
Motivation

- Suffer from feature confusion as they use a single representation for the two tasks – NER and RE.
- Underutilize the table structure as they usually convert it to a sequence.

Contribution

- We propose to learn two separate encoders – a table encoder and a sequence encoder. They interact with each other, and can capture task-specific information for the NER and RE tasks.
- We propose to use multidimensional recurrent neural networks to better exploit the structural information of the table representation.
- We effectively leverage the word-word interaction information carried in the attention weights from BERT, which further improves the performance.

Architecture



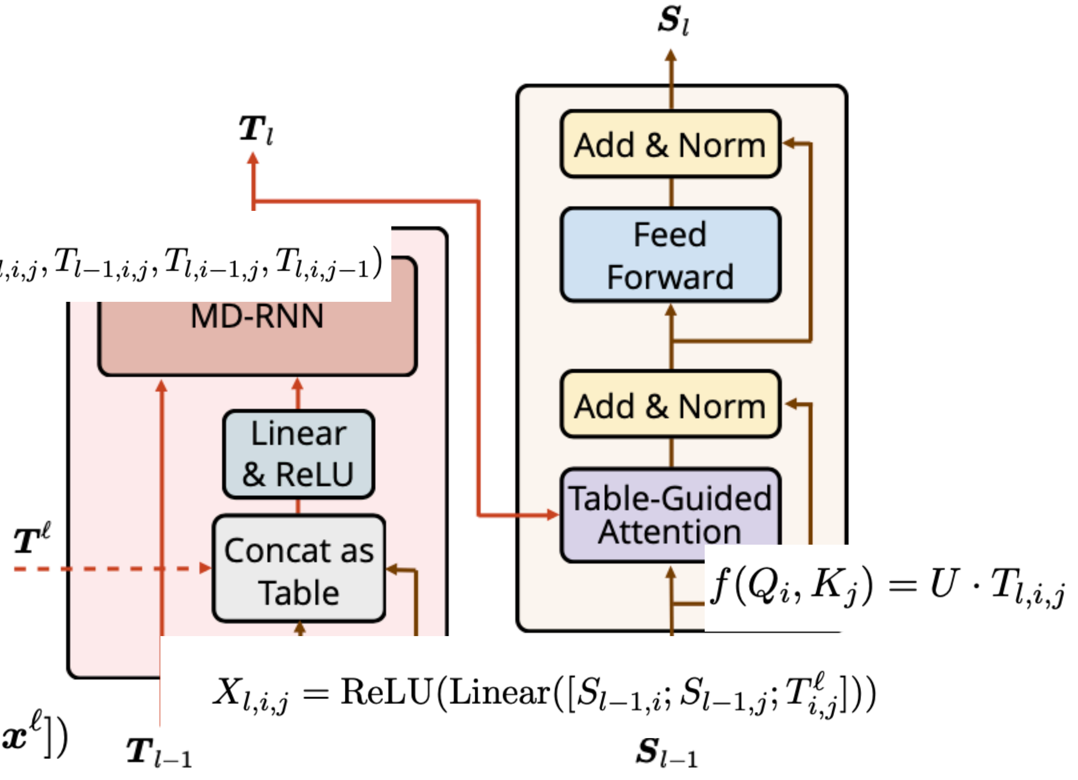


Figure 3: A layer in the table-sequence encoders.

MD-RNN

$$T_{l,i,j} = \mathrm{GRU}(X_{l,i,j}, T_{l-1,i,j}, T_{l,i-1,j}, T_{l,i,j-1}) \quad (3)$$

$$T^{prev}_{l,i,j} = [T_{l-1,i,j}; T_{l,i-1,j}; T_{l,i,j-1}], \in \mathbb{R}^{3H} \quad (18)$$

$$r_{l,i,j} = \sigma([X_{l,i,j}; T^{prev}_{l,i,j}]W^r + b^r)), \in \mathbb{R}^H \quad (19)$$

$$z_{l,i,j} = \sigma([X_{l,i,j}; T^{prev}_{l,i,j}]W^z + b^z)), \in \mathbb{R}^H \quad (20)$$

$$\tilde{\lambda}_{l,i,j,m} = [X_{l,i,j}; T^{prev}_{l,i,j}]W^\lambda_m + b^\lambda_m, \in \mathbb{R}^H \quad (21)$$

$$\begin{aligned}\lambda_{l,i,j,0}, \lambda_{l,i,j,1}, \lambda_{l,i,j,2} = \\ \mathrm{softmax}(\tilde{\lambda}_{l,i,j,0}, \tilde{\lambda}_{l,i,j,1}, \tilde{\lambda}_{l,i,j,2})\end{aligned} \quad (22)$$

$$\begin{aligned}\tilde{T}_{l,i,j} &= \tanh(X_{l,i,j}W^x \\ &+ r_{l,i,j} \odot (T^{prev}_{l,i,j}W^p) + b^h), \in \mathbb{R}^H\end{aligned} \quad (23)$$

$$\begin{aligned}\tilde{T}^{prev}_{l,i,j} &= \lambda_{l,i,j,0} \odot T_{l-1,i,j} \\ &+ \lambda_{l,i,j,1} \odot T_{l,i-1,j} \\ &+ \lambda_{l,i,j,2} \odot T_{l,i,j-1}, \in \mathbb{R}^H\end{aligned} \quad (24)$$

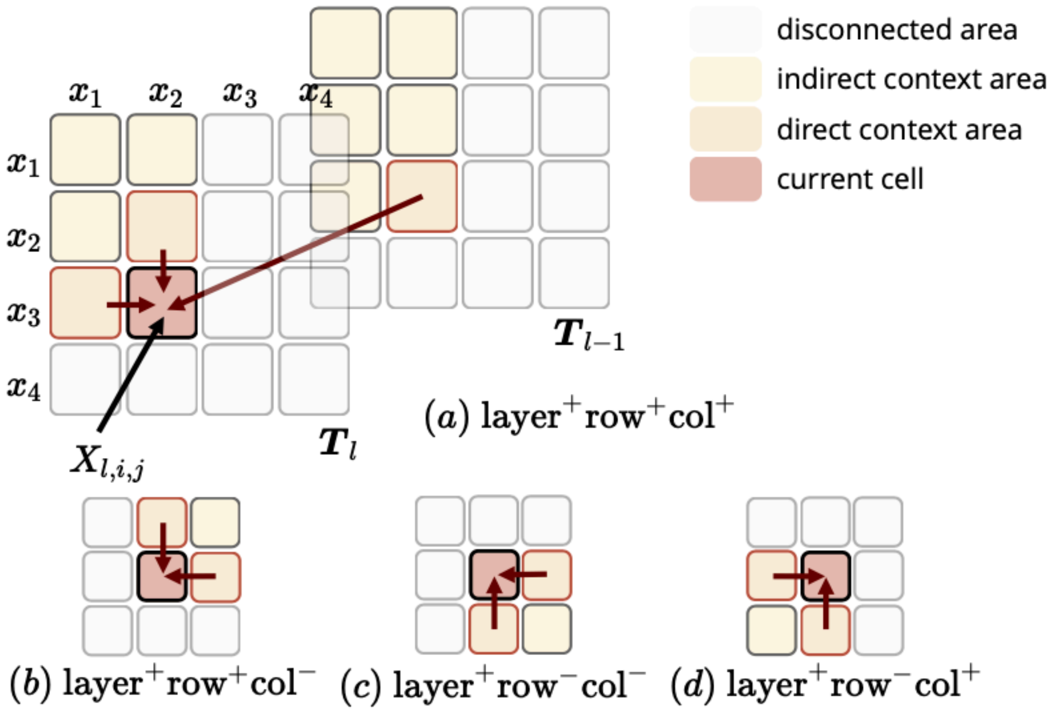
$$\begin{aligned}T_{l,i,j} &= z_{l,i,j} \odot \tilde{T}_{l,i,j} \\ &+ (1 - z_{l,i,j}) \odot \tilde{T}^{prev}_{l,i,j}, \in \mathbb{R}^H\end{aligned} \quad (25)$$

MD-RNN

$$T^{(a)}_{l,i,j} = \mathrm{GRU}^{(a)}(X_{l,i,j}, T^{(a)}_{l-1,i,j}, T^{(a)}_{l,i-1,j}, T^{(a)}_{l,i,j-1})$$

$$T^{(c)}_{l,i,j} = \mathrm{GRU}^{(c)}(X_{l,i,j}, T^{(c)}_{l-1,i,j}, T^{(c)}_{l,i+1,j}, T^{(c)}_{l,i,j+1})$$

$$T_{l,i,j} = [T^{(a)}_{l,i,j}; T^{(c)}_{l,i,j}]$$



Setting	NER	RE
Unidirectional	89.6	66.9
<u>Bidirectional</u>	<u>89.5</u>	<u>67.6</u>
Quaddirectional	89.7	67.6
Layer-wise only	89.3	63.9
Bidirectional w/o column	89.5	67.2
Bidirectional w/o row	89.3	67.4
Bidirectional w/o layer	89.3	66.7

Train

$$P_\theta(\boldsymbol{Y}^{\text{NER}}) = \text{softmax}(\text{Linear}(\boldsymbol{S}_L))$$
$$P_\theta(\boldsymbol{Y}^{\text{RE}}) = \text{softmax}(\text{Linear}(\boldsymbol{T}_L))$$

$$\mathcal{L}^{\text{NER}} = \sum_{i \in [1,N]} -\log P_\theta(Y_i^{\text{NER}} = y_i^{\text{NER}})$$
$$\mathcal{L}^{\text{RE}} = \sum_{i,j \in [1,N]; i \neq j} -\log P_\theta(Y_{i,j}^{\text{RE}} = y_{i,j}^{\text{RE}})$$

Experiment

Data	Model	NER	RE	RE+
ACE04	Li and Ji (2014) ▽	79.7	48.3	45.3
	Katiyar and Cardie (2017) ▽	79.6	49.3	45.7
	Bekoulis et al. (2018b) ▽	81.2	-	47.1
	Bekoulis et al. (2018a) ▽	81.6	-	47.5
	Miwa and Bansal (2016) ▽	81.8	-	48.4
	Li et al. (2019) ▽	83.6	-	49.4
	Luan et al. (2019) ▽	87.4	59.7	-
	Ours ▽	**88.6**	**63.3**	**59.6**
ACE05	Li and Ji (2014) ▽	80.8	52.1	49.5
	Miwa and Bansal (2016) ▽	83.4	-	55.6
	Katiyar and Cardie (2017) ▽	82.6	55.9	53.6
	Zhang et al. (2017) ▽	83.6	-	57.5
	Sun et al. (2018) ▽	83.6	-	59.6
	Li et al. (2019) ▽	84.8	-	60.2
	Dixit and Al (2019) ▽	86.0	62.8	-
	Luan et al. (2019) ▽	88.4	63.2	-
	Wadden et al. (2019) ▽	88.6	63.4	-
	Ours ▽	**89.5**	**67.6**	**64.3**

CoNLL04	Miwa and Sasaki (2014)▽	80.7	-	61.0
	Bekoulis et al. (2018a)▲	83.6	-	62.0
	Bekoulis et al. (2018b)▲	83.9	-	62.0
	Tran and Kavuluru (2019)▲	84.2	-	62.3
	Nguyen and Verspoor (2019)▲	86.2	-	64.4
	Zhang et al. (2017)▽	85.6	-	67.8
	Li et al. (2019)▽	87.8	-	68.9
	Eberts and Ulges (2019)▽	88.9	-	71.5
	Eberts and Ulges (2019)▲	86.3	-	72.9
	Ours▽	**90.1**	**73.8**	**73.6**
	Ours▲	**86.9**	**75.8**	**75.4**
ADE	Li et al. (2016) ▲	79.5	-	63.4
	Li et al. (2017) ▲	84.6	-	71.4
	Bekoulis et al. (2018b) ▲	86.4	-	74.6
	Bekoulis et al. (2018a) ▲	86.7	-	75.5
	Tran and Kavuluru (2019) ▲	87.1	-	77.3
	Eberts and Ulges (2019) ▲	89.3	-	79.2
	Ours ▲	**89.7**	**80.1**	**80.1**

Experiment

LM	$+\boldsymbol{x}^\ell$		$+\boldsymbol{x}^\ell$ $+\boldsymbol{T}^\ell$	
	NER	RE	NER	RE
ELMo	86.4	64.3	-	-
BERT	87.8	64.8	88.2	67.4
RoBERTa	88.9	66.2	89.3	67.6
ALBERT	89.4	66.0	89.5	67.6

Table 2: Using different pre-trained language models on ACE05. $+\boldsymbol{x}^\ell$ uses the contextualized word embeddings; $+\boldsymbol{T}^\ell$ uses the attention weights.

Experiment

Setting	NER	RE	RE (gold)
Default	89.5	67.6	70.4
w/o Relation Loss	89.4	-	-
w/o Table Encoder	88.4	-	-
w/o Entity Loss	-	-	69.8
w/o Sequence Encoder	-	-	69.2
w/o Bi-Interaction	88.2	66.3	69.2
NER on diagonal	89.4	67.1	70.2
w/o Sequence Encoder	88.6	67.0	70.2

Table 3: Ablation of the two encoders on ACE05. Gold entity spans are given in RE (gold).

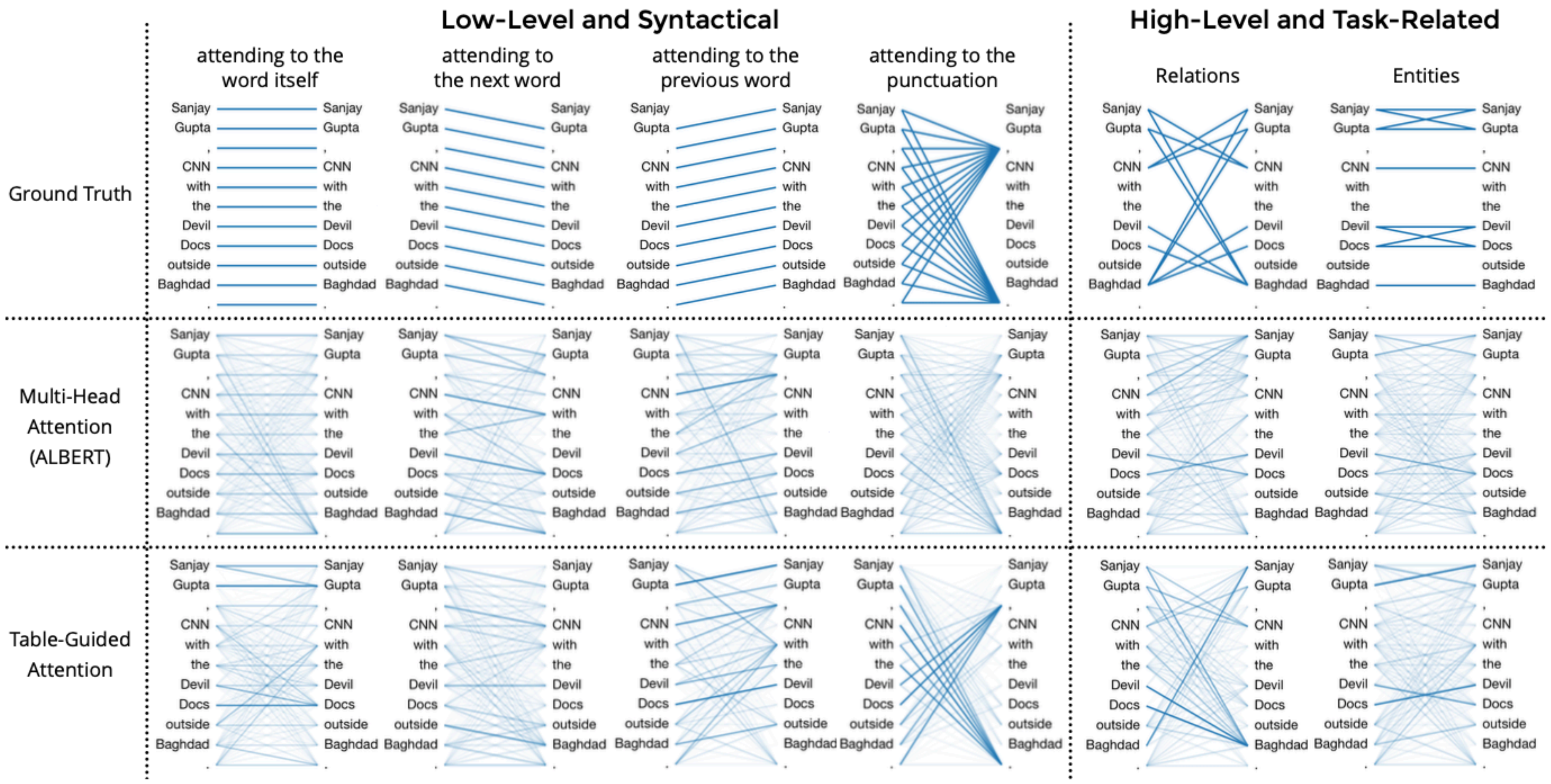


Figure 6: Comparison between ground truth and selected heads of ALBERT and table-guided attention. The sentence is randomly selected from the development set of ACE05.

Experiment

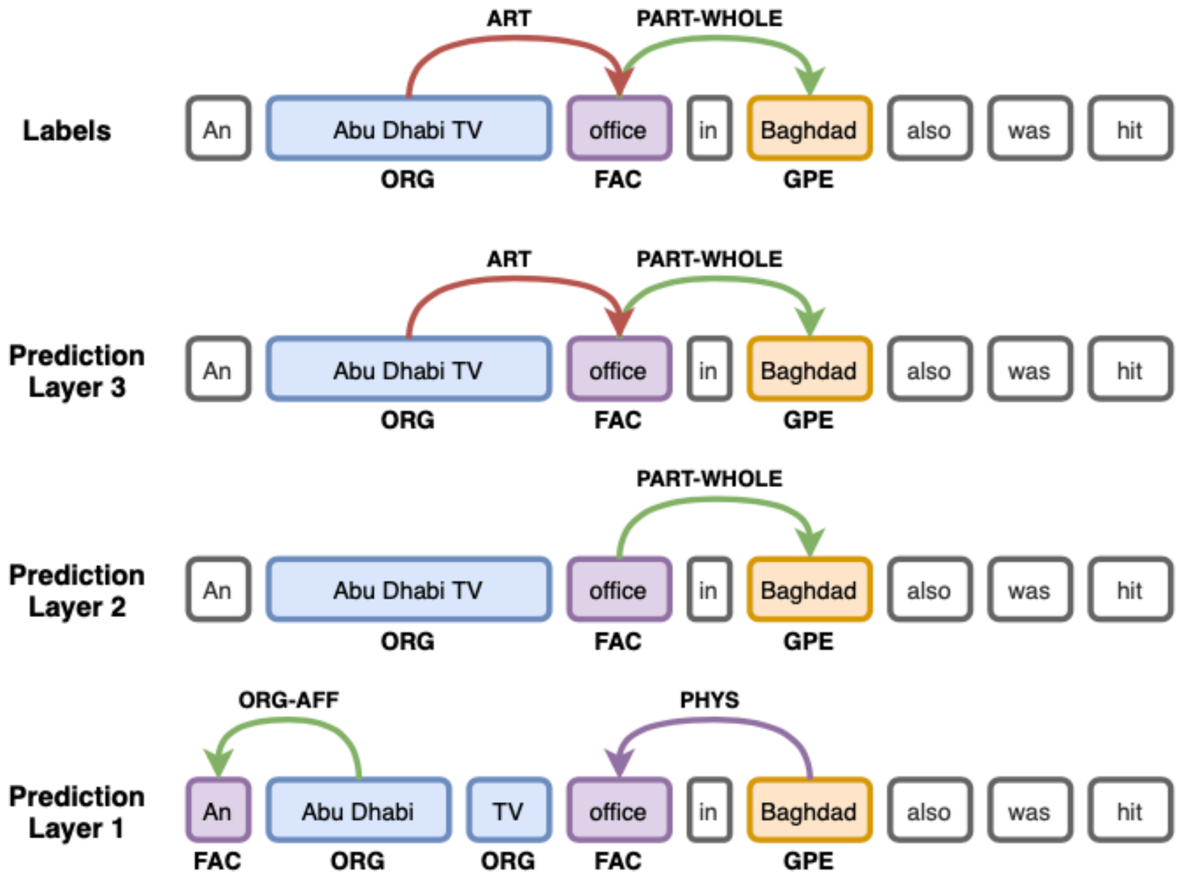


Figure 7: Probing intermediate states

COLING 2020

Span-based Joint Entity and Relation Extraction with Attention-based Span-specific and Contextual Semantic Representations

Bin Ji[†], Jie Yu[†], Shasha Li[*], Jun Ma, Qingbo Wu, Yusong Tan, Huijun Liu[*]
College of Computer,
National University of Defense Technology, Changsha, China
{jibin, yj, shashali, majun}@nudt.edu.cn
{qingbowu, yusongtan, liuhuijun}@nudt.edu.cn

Problems

Sentence 1: The army said troops fired, and hit a boy, after [**a Palestinian youth**]PER threw a stone,⋯

Sentence 2: The weak score is primarily the result of [**Starbucks**]ORG current strategy of increasing equity **ownership** of [**several foreign subsidiaries**]ORG,⋯
Relation: <several foreign subsidiaries, Starbucks, **PART-WHOLE**>

Sentence 3: [**Palestinians**]GPE claim [**all of the West Bank and Gaza**]LOC for a **state**,⋯
Relation: <all of the West Bank and Gaza, Palestinians, **PART-WHOLE**>

Encoder

$$\mathcal{B}_{\mathcal{S}} = [X_0, X_1, X_2, X_3, ..., X_n]$$

$$\mathcal{B}_{\mathbf{s}} = [X_i, X_{i+1}, X_{i+2}, ..., X_{i+j}]$$

NER

$$\mathcal{V}_{\mathbf{k}} = \mathbf{MLP}_k(X_{\mathbf{k}}) \qquad s.t. \quad \mathbf{k} \in [i, i+j]$$

$$\alpha_{\mathbf{k}} = \frac{\mathbf{exp}(\mathcal{V}_{\mathbf{k}})}{\sum_{m=i}^{i+j} \mathbf{exp}(\mathcal{V}_{\mathbf{m}})}$$

$$\mathcal{H}_{\mathbf{s}} = [X_i; X_{i+j}]$$

$$\mathcal{F}_{\mathbf{s}} = \sum_{m=i}^{i+j} \alpha_{\mathbf{m}} X_{\mathbf{m}}$$

$$\mathcal{T}_{\mathbf{s}} = \mathbf{Attention}(\mathcal{F}_{\mathbf{s}}, \mathcal{B}_{\mathcal{S}}, \mathcal{B}_{\mathcal{S}})$$

$$\mathcal{R}_{\mathbf{s}} = [\mathcal{T}_{\mathbf{s}};\ \mathcal{F}_{\mathbf{s}};\ \mathcal{H}_{\mathbf{s}};\ \mathcal{W}_{j+1}]$$

$$y_{\mathbf{s}} = \mathbf{Softmax}(\mathbf{FFNN}(\mathcal{R}_{\mathbf{s}}))$$

RE

$$\mathcal{H}_{\mathbf{r}} = [\mathbf{FFNN}(\mathcal{R}_{s_1}); \mathbf{FFNN}(\mathcal{R}_{s_2})]$$

$$\mathcal{B}_{\mathbf{c}} = (X_m, X_{m+1}, X_{m+2}, \ldots, X_{m+n})$$

$$\mathcal{F}_{\mathbf{r}} = \mathbf{Attention}(\mathcal{H}_{\mathbf{r}}, \mathcal{B}_{\mathbf{c}}, \mathcal{B}_{\mathbf{c}})$$

$$\mathcal{T}_{\mathbf{r}} = \mathbf{Attention}(\mathcal{H}_{\mathbf{r}}, \mathcal{B}_{\mathcal{S}}, \mathcal{B}_{\mathcal{S}})$$

$$\mathcal{R}_{\mathbf{r}} = [\mathcal{H}_{\mathbf{r}}; \mathbf{FFNN}_{\mathcal{F}}(\mathcal{F}_{\mathbf{r}}); \mathbf{FFNN}_{\mathcal{T}}(\mathcal{T}_{\mathbf{r}})]$$

$$y_{\mathbf{r}} = \mathbf{Softmax}(\mathbf{FFNN}(\mathcal{R}_{\mathbf{r}}))$$

$$\mathcal{L} = 0.4\mathcal{L}^{\mathbf{s}} \ + \ 0.6\mathcal{L}^{\mathbf{r}}$$

Experiment

Dataset	Method	Entity			Relation		
		P	R	F1	P	R	F1
ACE05	Multi-turn QA †	84.7	84.9	84.8	64.8	56.2	60.2
	DyGIE ‡	-	-	88.4	-	-	63.2
	SPAN$_{Multi-Head}$	**89.32**	**89.86**	**89.59**	**71.22**	**60.19**	**65.24**
CoNLL04	Multi-turn QA †	89.0	86.6	87.8	69.2	68.2	68.9
	SpERT ‡	88.25	89.64	88.94	73.04	70.00	71.47
	SPAN$_{Multi-Head}$	**90.11**	**90.36**	**90.23**	**76.96**	**71.88**	**74.33**
ADE	Relation-Metric †	86.16	88.08	87.11	77.36	77.25	77.29
	SpERT ‡	88.99	89.59	89.28	77.77	79.96	78.84
	SPAN$_{Multi-Head}$	**89.88**	**91.32**	**90.59**	**79.56**	**81.93**	**80.73**

Experiment

$$Multi-Head\ attention:\ \mathbf{score} = \frac{\mathcal{Q} \odot \mathcal{K}}{\sqrt{d_{\mathcal{K}}}}$$

$$Additive\ attention:\ \mathbf{score} = \mathcal{W}_1 \cdot \mathcal{Q} + \mathcal{W}_2 \cdot \mathcal{K}$$

$$Dot-Product\ attention:\ \mathbf{score} = \mathcal{W} \cdot (\mathcal{Q} \odot \mathcal{K})$$

$$General\ attention:\ \mathbf{score} = \mathcal{Q} \cdot \mathcal{W} \cdot \mathcal{K}$$

Method	ACE05		CoNLL04		ADE	
	Entity (F1)	Relation (F1)	Entity (F1)	Relation (F1)	Entity (F1)	Relation (F1)
$\mathbf{SPAN}_{Multi-Head}$	**89.59**	**65.24**	**90.23**	**74.33**	**90.59**	**80.73**
$\mathbf{SPAN}_{Dot-Product}$	87.94	62.88	88.23	70.89	88.15	77.31
$\mathbf{SPAN}_{General}$	88.66	63.56	88.96	73.48	89.93	80.14
$\mathbf{SPAN}_{Additive}$	89.07	64.53	89.17	71.36	89.68	79.75

Experiment

Method	Entity (F1)	Relation (F1)
$\mathbf{SPAN}_{Multi-Head}$	88.10	62.13
`-SpanSpecific`	86.78	60.21
`-SentenceLevel`	87.57	61.12
`base`	85.80	59.00

Method	Entity (F1)	Relation (F1)
$\mathbf{SPAN}_{Multi-Head}$	88.10	62.13
`-local`	87.96	60.56
`-SentenceLevel`	88.21	61.77
`base`	87.91	59.66

Thank you