IJCAI 2019

Beyond Word Attention: Using Segment Attention in Neural Relation Extraction

Bowen Yu[1,2], **Zhenyu Zhang**[1,2], **Tingwen Liu**[1 *], **Bin Wang**[3], **Sujian Li**[4] and **Quangang Li**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
[4]Key Laboratory of Computational Linguistics, Peking University, MOE, China
{yubowen, zhangzhenyu1996, liutingwen, liquangang}@iie.ac.cn, wangbin11@xiaomi.com,
lisujian@pku.edu.cn

Task and Motivation

- Sentence-level Relation Extraction on TACRED datasets.
- Attention mechanisms are often used in this task to alleviate the inner-sentence noise by performing soft selections of words independently
- Observation: information pertinent to relations is usually contained within segments (continuous words in a sentence)

Task and Motivation

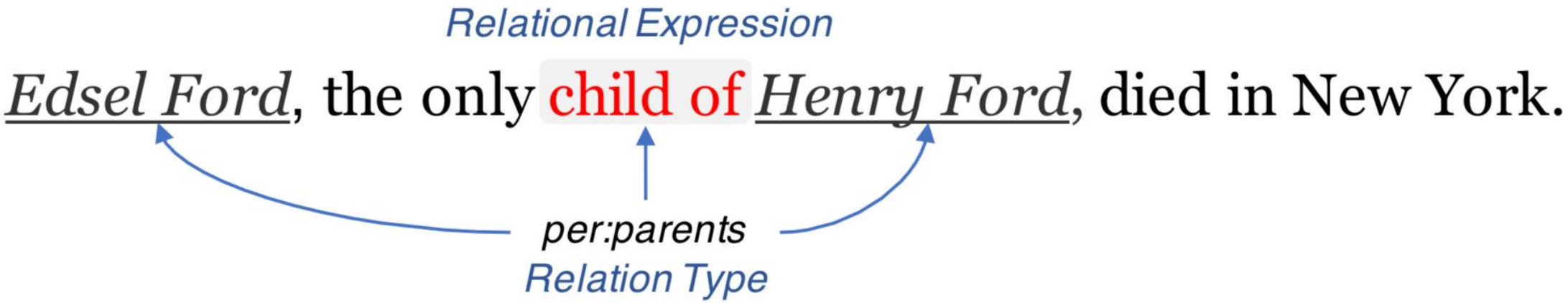
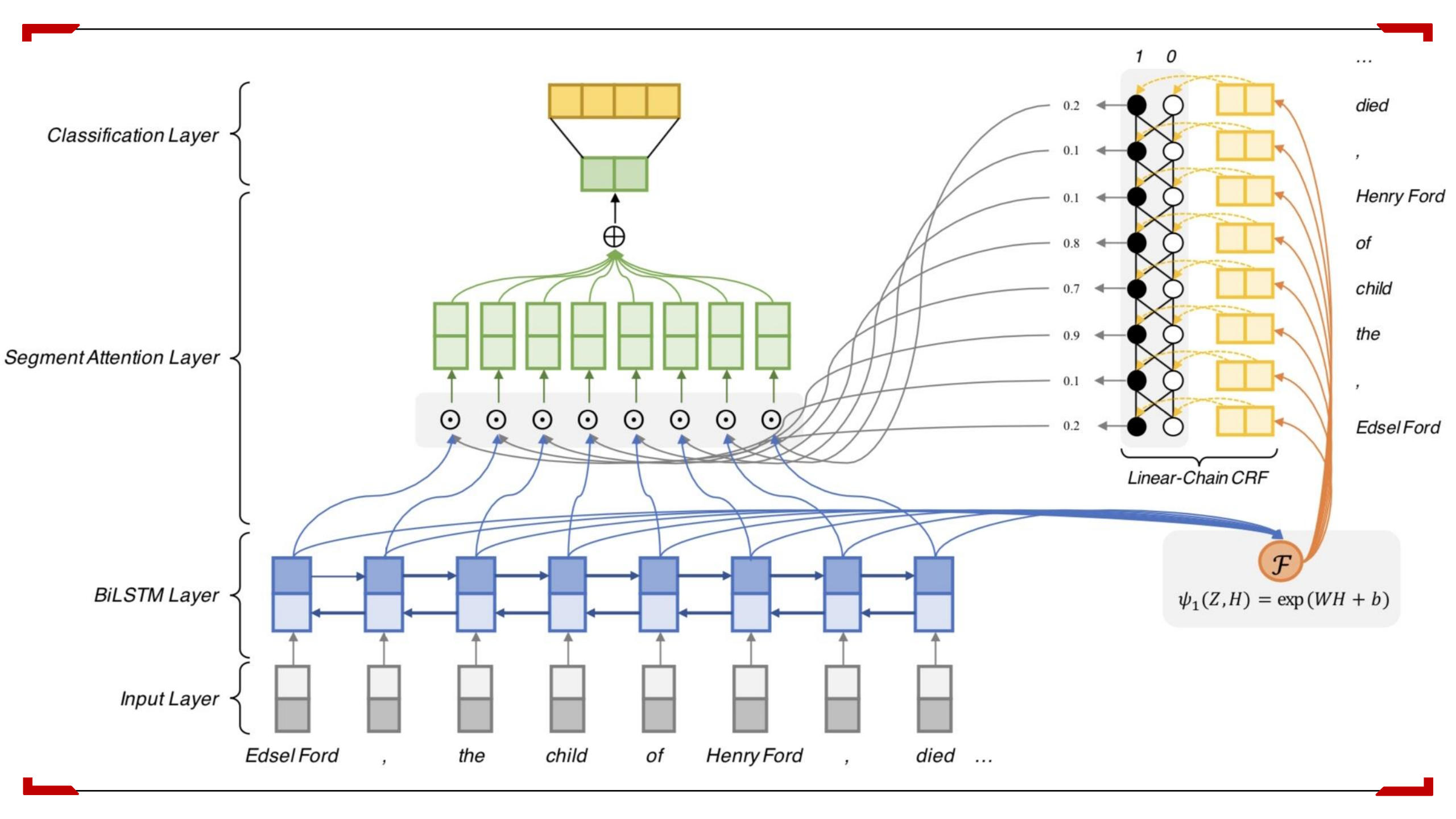


Figure 1: An example modified from the TACRED dataset. The relational expression is highlighted with a box and points to its corresponding entity pair.

Task and Motivation

- we sample out 200 examples from the standard dataset and annotate the relational expression of each sentence manually.
- We find that half of the relational expressions are in the form of segment and longer than 2 words, which means that accurately extracting and modeling such segment information can be extremely crucial.


Classification Layer
Segment Attention Layer
BiLSTM Layer
Input Layer
Edsel Ford , the child of Henry Ford , died …
1 0
0.2
0.1
0.1
0.8
0.7
0.9
0.1
0.2
Linear-Chain CRF
… died , Henry Ford of child the , Edsel Ford
$\mathcal{F}$
$\psi_1(Z, H) = \exp(WH + b)$

Input Layer and BiLSTM layer

- $\boldsymbol{x}_i = [\boldsymbol{w}_i; \boldsymbol{p}_{ih}; \boldsymbol{p}_{it}] \in \mathbb{R}^{d_w + d_p \times 2}$

$$\mathbf{h}_i = [\overrightarrow{\mathrm{LSTM}}(\mathbf{x}_i); \overleftarrow{\mathrm{LSTM}}(\mathbf{x}_i)], i \in [1, n]$$

Segment Attention Layer

- Similar to the standard attention used in RE, our segment attention is also a linear weighted combination of the input representations
- We introduce a discrete latent binary variable $\mathbf{z} \in \{0, 1\}$ for each word, which indicates whether its corresponding word is part of a relational expression or not.
- The representation of the given sequence $\mathbf{m}$ is defined as

$$\mathbf{m} = \sum_i p(z_i = 1|\mathbf{H})\mathbf{h}_i$$

Linear-chain CRF

- $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$
- $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$
- $\mathcal{Z}$: set of possible label sequences z

$$p(\mathbf{z}|\mathbf{H}) = \frac{1}{Z(\mathbf{H})} \prod_{c \in C} \psi(\mathbf{z_c}, \mathbf{H})$$

$$Z(\mathbf{H}) = \sum_{\mathbf{z}' \in \mathcal{Z}} \prod_{c \in C} \psi(\mathbf{z'_c}, \mathbf{H})$$

$$\prod_{c \in C} \psi(\mathbf{z_c}, \mathbf{H}) = \prod_{i=1}^{n} \psi_1(z_i, \mathbf{h}_i) \prod_{i=1}^{n-1} \psi_2(z_i, z_{i+1})$$

Linear-chain CRF

- Vertex feature $\psi_1(z_i, \boldsymbol{h}_i)$ represents the mapping from the input hi to output zi through a single layer neural network

$$\psi_1(z_i, \mathbf{H}) = \exp(\mathbf{W}^v_{z_i} \cdot \mathbf{h}_i + b) \qquad \mathbf{W}^v \in \mathbb{R}^{2 \times 2d_h}$$

- Edge Feature $\psi_2(z_i, z_{i+1})$ models the transition (position independent) from i-th state to $i + 1$-th for a pair of consecutive time steps.

$$\psi_2(z_i, z_{i+1}) = \exp(\mathbf{W}^t_{z_i, z_{i+1}}) \qquad \mathbf{W}^t \in \mathbb{R}^{2 \times 2}$$

Linear-chain CRF

$$\alpha_{i+1}(z|\mathbf{H}) = \sum_{z' \in \{0,1\}} \alpha_i(z'|\mathbf{H})\psi_1(z, \mathbf{h}_{i+1})\psi_2(z', z)$$

$$p(z_i = 1|\mathbf{H}) = \frac{\alpha_i(1|\mathbf{H}) * \beta_i(1|\mathbf{H})}{Z(\mathbf{H})}$$

Classifier Layer

$$p(r|\mathbf{m}) = \mathrm{softmax}(\mathbf{W}_\mathrm{r} \cdot \mathbf{m} + \mathrm{b}_\mathrm{r})$$

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} -y_i \log p(y_i)$$

$$\Omega_s = \sum_{i=1}^{n} p(z_i = 1|\mathbf{H})$$

$$\Omega_t = \max(0, \mathbf{W}^t_{1,0} - \mathbf{W}^t_{1,1}) + \max(0, \mathbf{W}^t_{0,1} - \mathbf{W}^t_{0,0})$$

$$L(\theta) = J(\theta) + \lambda_1 \Omega_t + \lambda_2 \Omega_s$$

System	P	R	F_1
Pattern† [Angeli *et al.*, 2015]	**85.3**	23.4	36.8
LR† [Zhang *et al.*, 2017]	72.0	47.8	57.5
CNN-PE‡ [Zeng *et al.*, 2014]	68.2	55.4	61.1
PCNN‡ [Zeng *et al.*, 2015]	67.4	57.3	62.0
SDP-LSTM† [Xu *et al.*, 2015]	66.3	52.7	58.7
Tree-LSTM† [Tai *et al.*, 2015]	66.0	59.2	62.4
PA-LSTM† [Zhang *et al.*, 2017]	65.7	64.5	65.1
PA-LSTM+D‡	67.2	65.0	66.0
C-GCN† [Zhang *et al.*, 2018]	69.9	63.3	66.4
SA-LSTM	68.1	65.7*	66.9*
SA-LSTM+D	69.0	**66.2***	**67.6***

	Example	Predicted relation	True relation
PA-LSTM	SUBJ-PER SUBJ-PER, the son of Israel's first astronaut, OBJ-PER OBJ-PER, died in his home yesterday.	children	parents
SA-LSTM	SUBJ-PER SUBJ-PER, the son of Israel's first astronaut, OBJ-PER OBJ-PER, died in his home yesterday.	parents	
PA-LSTM	Prosecutors had accused SUBJ-PER, 22, then a student at OBJ-ORG OBJ-ORG, and her boyfriend Raffaele.	employee of	schools attended
SA-LSTM	Prosecutors had accused SUBJ-PER, 22, then a student at OBJ-ORG OBJ-ORG, and her boyfriend Raffaele.	schools attended	

1. *OBJ-PER OBJ-PER*, the president of the *SUBJ-ORG*, was sued by the SEC.

2. Founded in *OBJ-DATE*, *SUBJ-ORG* is a non-profit membership association.

3. *SUBJ-PER*, who served as bureau chief, was convicted of accepting bribes, *OBJ-CRIMINAL*.

4. Defendants are brought in together with *SUBJ-PER* including his wife Zhou Xiao and *OBJ-PER*.

Thanks!