Beyond Word Attention: Using Segment Attention in Neural Relation Extraction

Bowen Yu[†], Zhenyu Zhang[†], Tingwen Liu[†], Bin Wang[‡],
Sujian Li[∓], Quangang Li[†]

[†] Institute of Information Engineering,
Chinese Academy of Sciences, Beijing, China
[‡]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
[∓]Peking University, MOE, China

IJCAI 2019, Macao, China, August 2019

Outline

- Background
- Our work
- Experiments
- Conclusion & Outlook

Relation Extraction

Relation Extraction(RE), which is also called Relation Classification (RC), is the task of extracting semantic relationships between two target (given) entities from plain text. This task is an important and challenging stage in the construction of knowledge graph.

Relation Type with Labeled Dataset

↓

Plain Text Corpus (Unstructured Info) → **Classifier** → **Entity-Relation Triple (Structured Info)**

Main Challenge

Regarding relation extraction as a simple text classification problem is undesirable because of the inner-sentence noise.

Free Text *Edsel Ford*, the only child of *Henry Ford*, died in New York.

per:children

Relation Type

noise

Existing Solutions

- **Sentence Pruning Strategy**
  - empirically removes irrelevant content according to the distance of each word to the target entity in the sentence or in the dependency tree of the sentence
  - works on the input layer, can be combined with other methods
- **Attention Mechanism**
  - computes the attention score for each word to indicate how well the word can express the relation between the two entities.
  - can be viewed as the process of performing soft selections of individual words independently
  - neglects the rich dependencies among the words that describe the relation.

Motivation

- The relational expression may be in the form of a segment structure

Relational Expression

Free Text *Edsel Ford*, the only child of *Henry Ford*, died in New York.

per:children

Relation Type

- Half of the relational expressions in TACRED are in the form of segment and longer than 2 words.

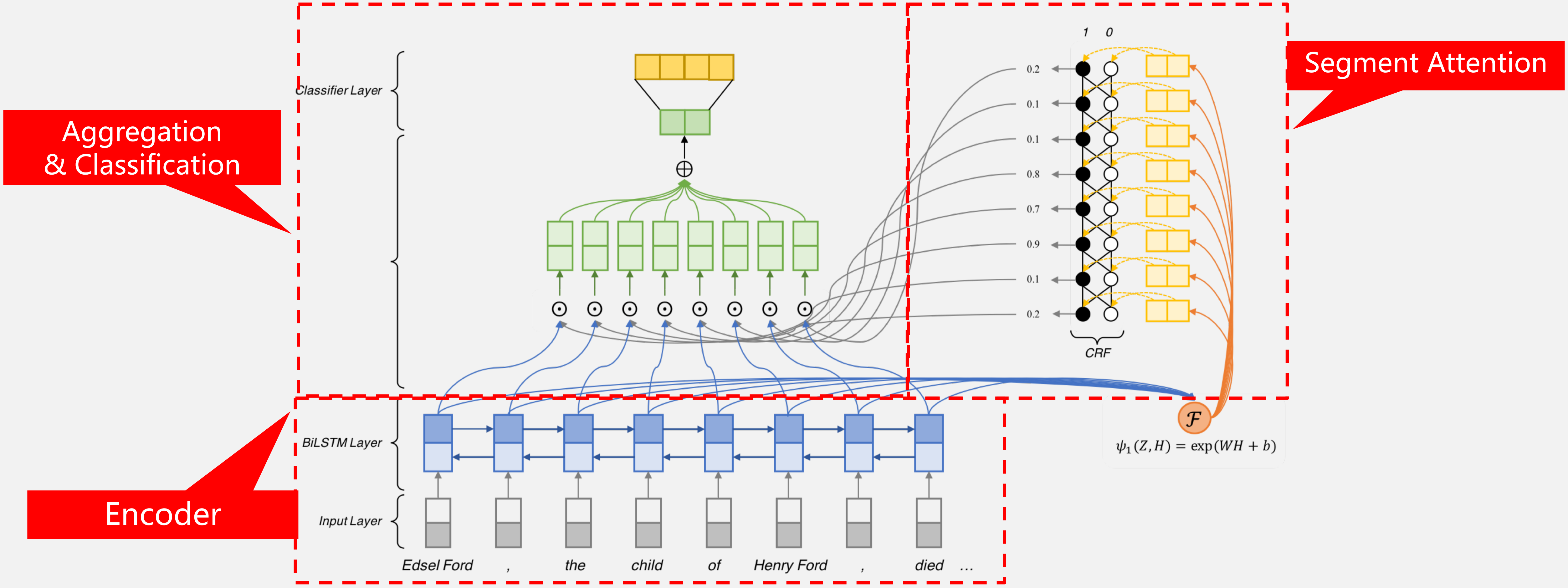
Accurately modeling such segment information can be extremely crucial.

Outline

- Background
- Our work
- Experiments
- Conclusion & Outlook

Our Model

- Our approach views the attention mechanism as a linear-chain CRF over a set of latent variables whose edges encode the desired structure.

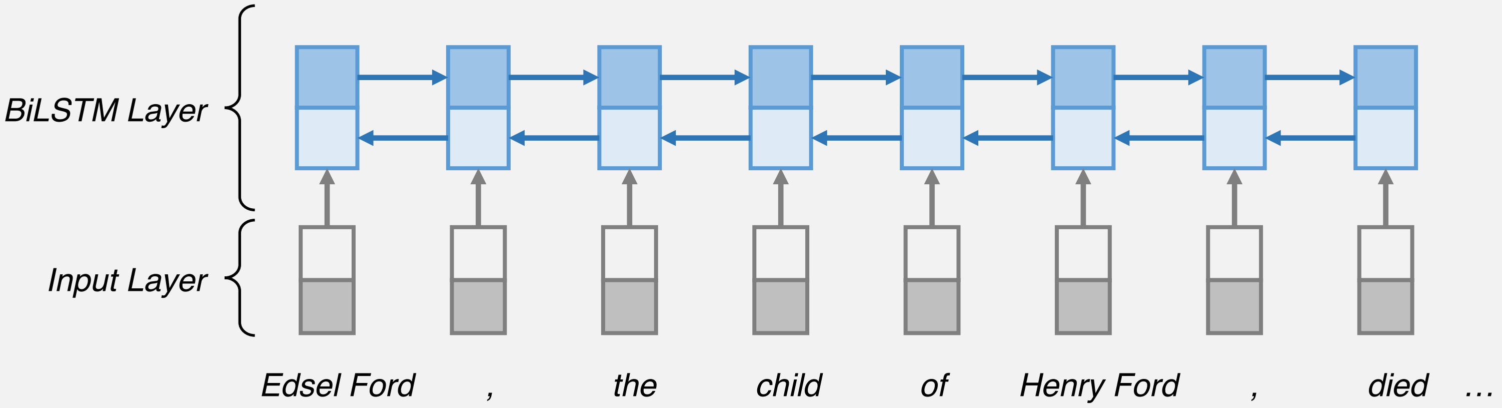


Encoder

- A BiLSTM layer is adopted to capture the contextual information for each word.

$$\mathbf{h}_i = [\overrightarrow{\mathrm{LSTM}}(\mathbf{x}_i); \overleftarrow{\mathrm{LSTM}}(\mathbf{x}_i)$$

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_1, \cdots, \mathbf{h}_n\}$$



Segment Attention

- Segment attention is incorporated to perform soft selections of a sequence of words.

$$z \in \{0,1\}$$

indicates whether its corresponding word is part of a relational expression or not

$$weight(i) = p(z_i = 1|\mathbf{H})$$

represents a generic sequence of labels for $\mathbf{H}$

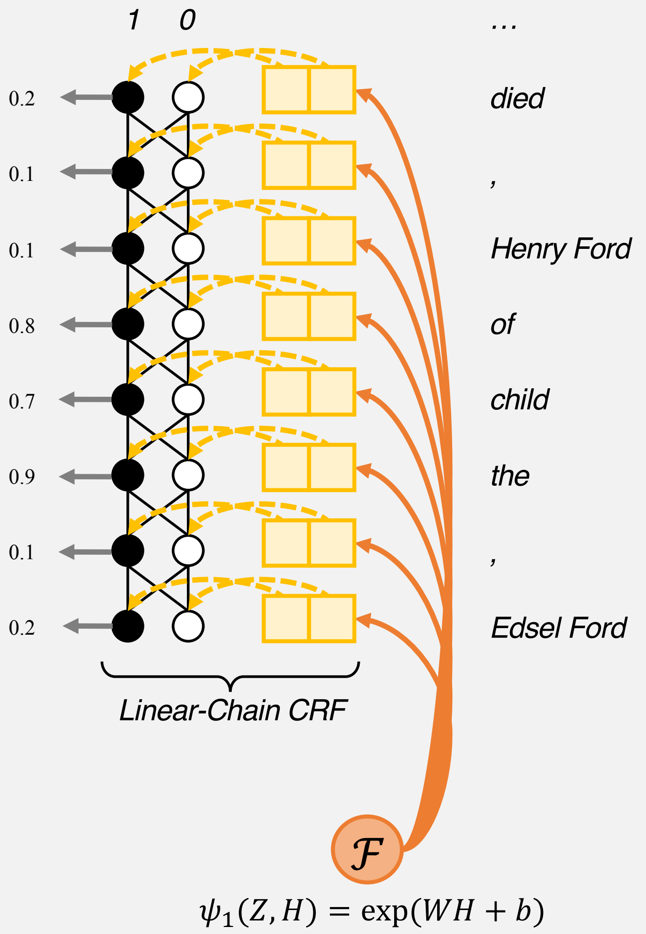
$$\mathbf{z} = [z_i, z_i, \cdots, z_i]$$

calculates $p(\mathbf{z}|\mathbf{H})$ over all possible label sequences $\mathbf{z}$, $\mathbf{z}_{\mathrm{c}}$ indicates the subset of $\mathbf{z}$ given by individual clique c

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

is the normalization constant that makes the probability of all sequences sum to one. $\mathcal{Z}$ denotes the set of possible label sequences $\mathbf{z}$

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



Segment Attention

- Segment attention is incorporated to perform soft selections of a sequence of words.

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

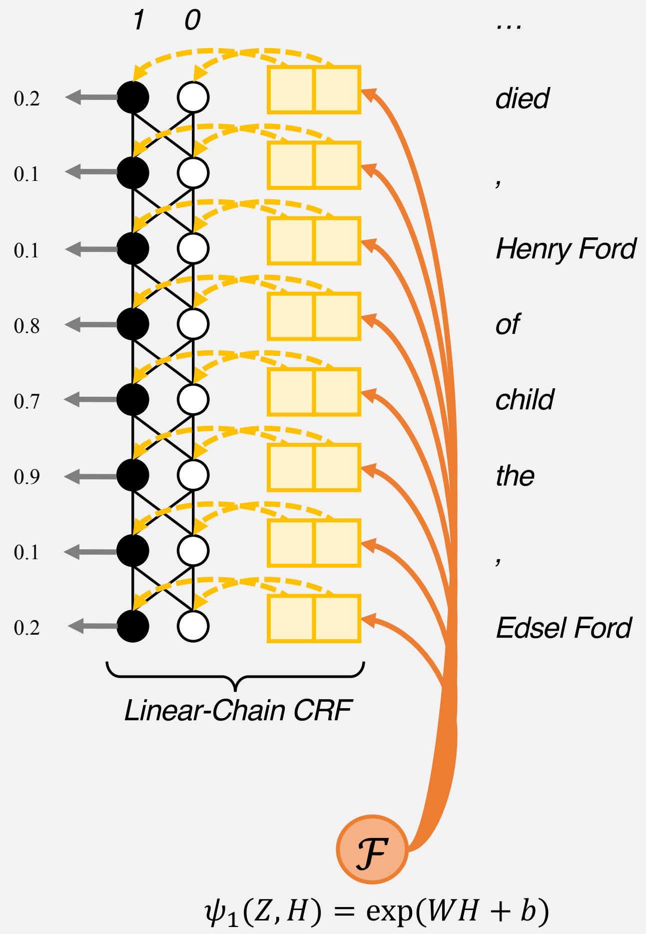
Two types features

Vertex feature $\psi_1(z_i, \mathbf{h}_i)$ represents the mapping from $\mathbf{h}_i$ to z_i

$$\psi_1(z_i, \mathbf{h}_i) = \exp(\mathbf{W}_{z_i}^{\mathrm{v}} \cdot \mathbf{h}_i + b)$$

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

$$\psi_2(z_i, z_{i+1}) = \exp(\mathbf{W}_{z_i, z_{i+1}}^{\mathrm{t}})$$



Segment Attention

- Segment attention is incorporated to perform soft selections of a sequence of words.

$p(z_i = 1|\mathbf{H})$ can be computed by a dynamic programming inference procedure similar to the forward-backward procedure.

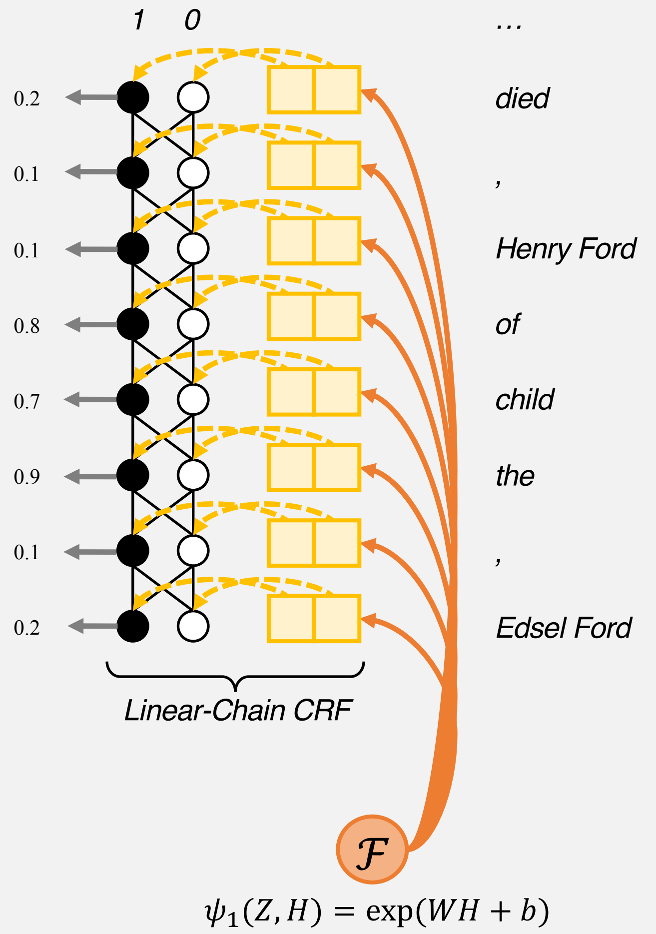
$$p(z_i = 1|\mathrm{H}) = \frac{\alpha_i(1|\mathbf{H}) * \beta_i(1|\mathbf{H})}{z(H)}$$

forward values

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

backward values

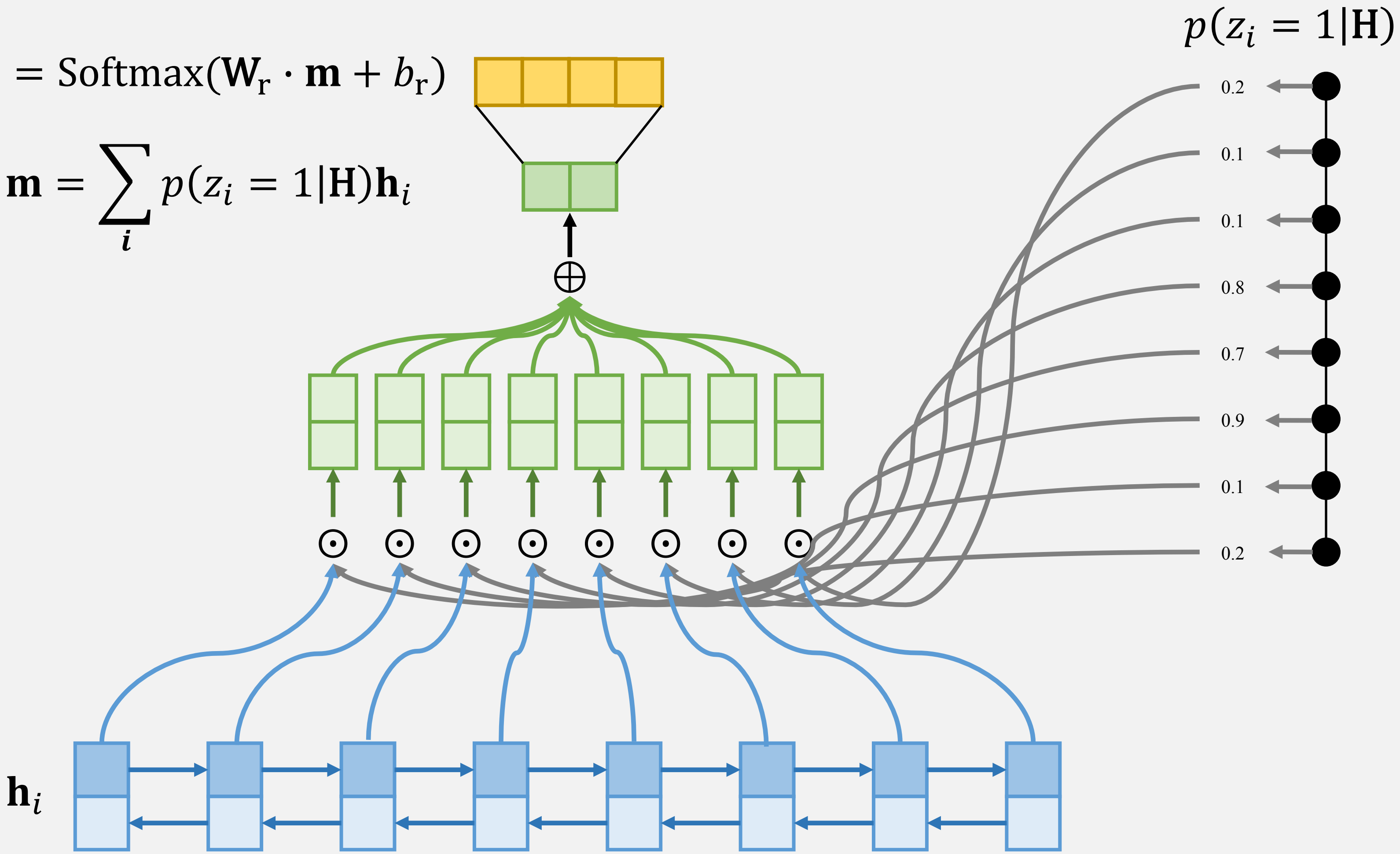
$$\beta_{i-1}(z|\mathrm{H}) = \sum_{z' \in \{0,1\}} \alpha_i(z'|\mathrm{H})\psi_1(z, \mathbf{h}_{i-1})\psi_2(z', z)$$



Aggregation & Classification

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

$$\mathbf{m} = \sum_{\boldsymbol{i}} p(z_i = 1|\mathrm{H})\mathbf{h}_i$$



Objective Function

the cross entropy loss of relation extraction

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

the transition regularizer to encourage the state to stay the same and discourage frequent transitions between different states

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

the sparse regularizer to enforce the model to attend to few words that really matter

$$\Omega_s = \sum_i p(z_i = 1|\mathrm{H})$$

final objective function

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

Outline

- Background
- Our work
- Experiments
- Conclusion & Outlook

Overall Performance

- **TACRED ---** 106k entity pairs

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***

augmented with the shortened sentences used in C-GCN

Case Study

	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	

Case Study

- We sample out some instances and use Viterbi decoding algorithm to extract the relation expressions explicitly

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*.

Outline

- Background
- Our work
- Experiments
- Conclusion & Outlook

Conclusion & Outlook

- We propose a novel model that learns the latent relational expressions based on the segment attention layer for relation extraction.

- By incorporating a linear-chain CRF into the attention layer, our model is capable of capturing the dependencies between target entities and their relations.

- In the future, we will conduct research on how to design more sophisticated attention mechanism to alleviate the inter-sentence noise.

Thanks!

Questions and Advices?