Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention

Yue Zhao, Xiaolong Jin,
Yuanzhuo Wang, Xueqi Cheng

University of Chinese Academy of Sciences
CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology,
Chinese Academy of Sciences

☰ Content

Introduction

- Event Detection
  - subtask of **event extraction**
  - given a **document**, extract **event triggers** from **individual sentences** and further identifies the (pre-defined) **type of events**
- Event Trigger
  - words in sentences that most clearly expresses occurrence of events

… They have been ***married*** for three years. …

❖ Event Trigger is "married", which represents a marry event

Motivation

… I knew it was time to ***leave***. …

? Transport event ? End-Position event

- A single sentence may cause ambiguous

… I knew it was time to ***leave***.
Is not that a great argument for term limits? …

√ End-Position event

- The contextual information of a individual sentence offers more confident for classifying

Motivation

Some shortcomings of existing works

- Manually designed document-level feature

  Ji and Grishman, ACL, 2008

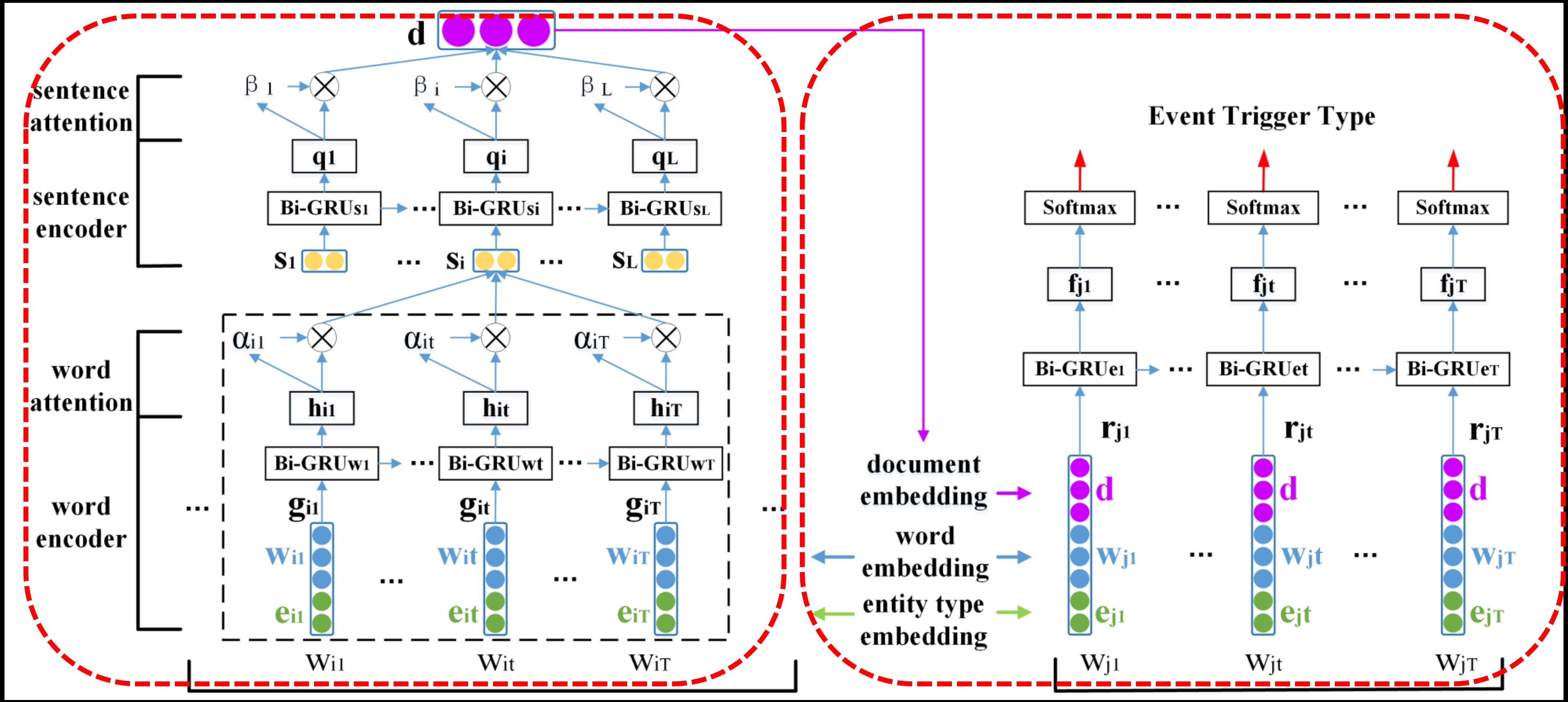
  Liao and Grishman, ACL, 2010

  Huang and Riloff, AAAI, 2012

- Learning document embedding without supervision, cannot specifically capture event-related information

  Duan et al., IJCNLP, 2017

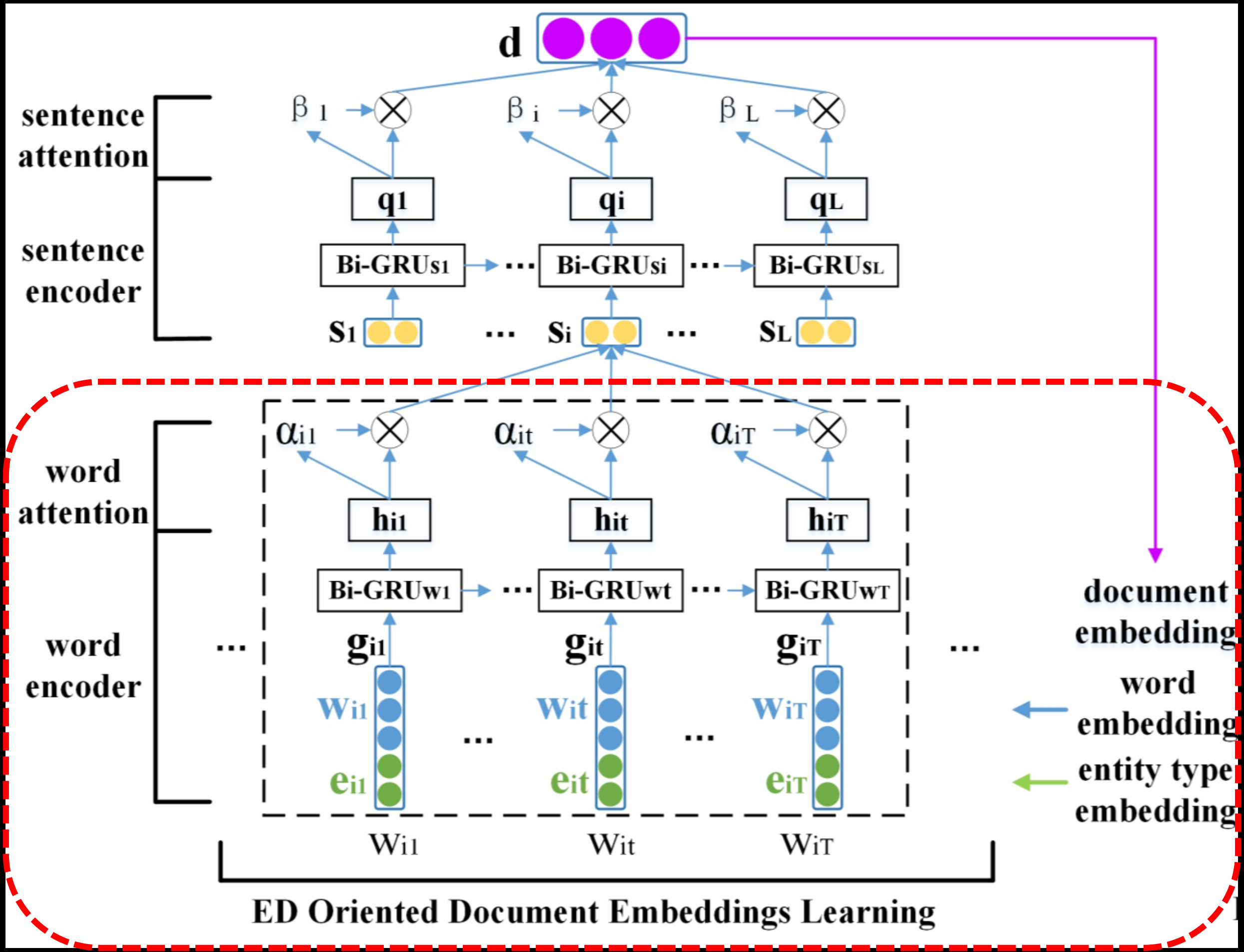
DEEB-RNN : The Proposed Model



ED Oriented Document Embedding Learning

Document-level Enhanced Event Detector

Model - ED Oriented Document Embedding Learning



Word-level embeddings

- Word encoder

$$h_{it} = \text{Bi-GRU}_w([w_{it}, e_{it}])$$

- Word attention

$$u_{it} = \tanh(W_w h_{it})$$

$$\alpha_{it} = u_{it}^T c_w$$

- Sentence representation

$$s_i = \sum_{t=1}^{T} \alpha_{it} h_{it}$$

Model - ED Oriented Document Embedding Learning

➢ Gold word-level attention signal:

Joy	Fenter	was	indicted	by	the	grand	Jury	.	
0	0	0	1	0	0	0	0	0	α_i^*

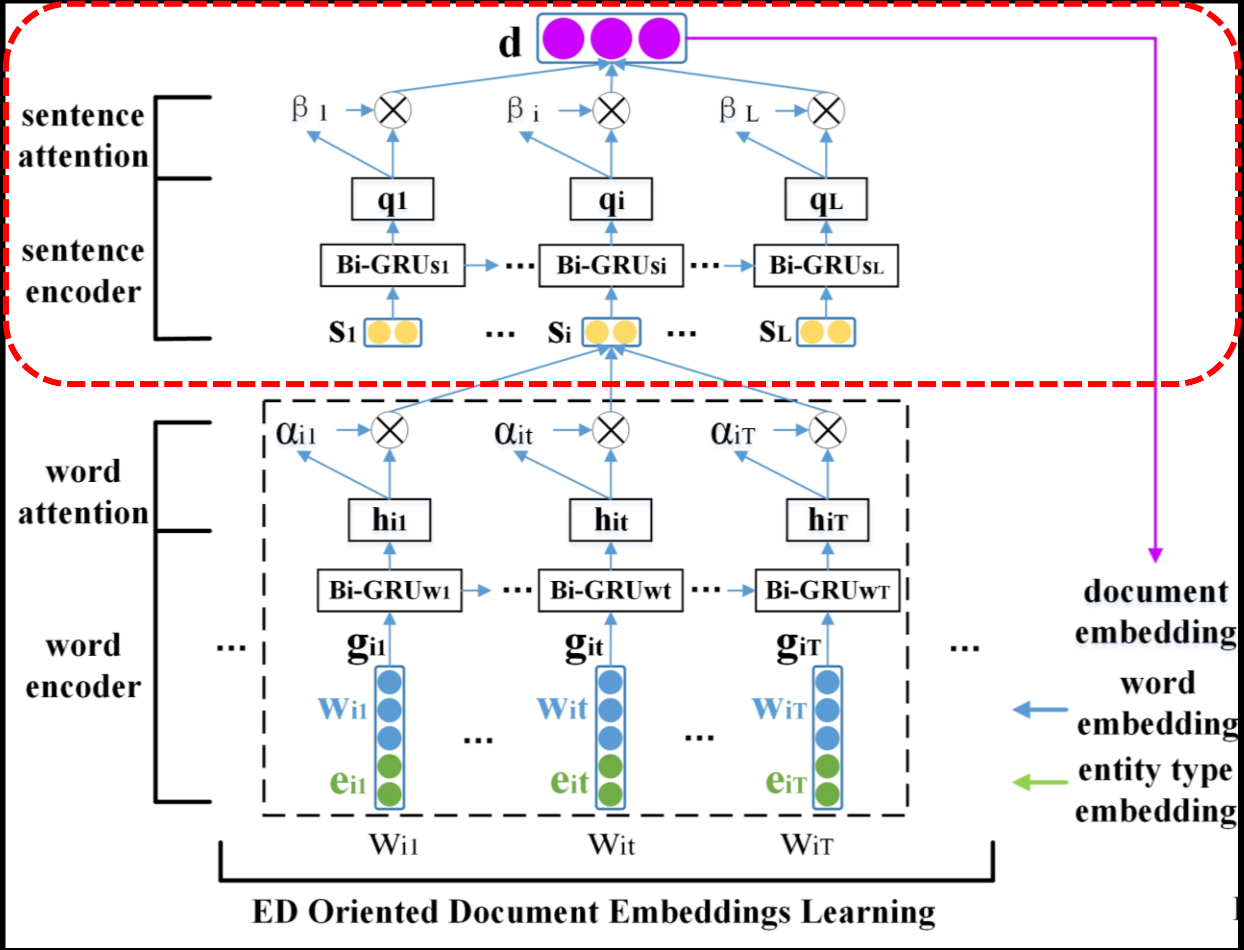
❖ "Indicated" is a event trigger and is setted as 1, other words are setted as 0.

➢ Loss function:

$$E_w(\alpha^*, \alpha) = \sum_{i=1}^{L} \sum_{t=1}^{T} (\alpha_{it}^* - \alpha_{it})^2$$

❖ The square error as the general loss of the attention at word level to supervise the learning process.

Model - ED Oriented Document Embedding Learning



Sentence-level embeddings

- Sentence encoder

$$q_i = \text{Bi-GRU}_s(s_i)$$

- Sentence attention

$$t_i = \tanh(W_s q_i)$$

$$\beta_i = t_i^T c_s$$

- Document representation

$$d = \sum_{i=1}^{L} \beta_i s_i$$

Model - ED Oriented Document Embedding Learning

- Gold sentence-level attention signal:

S_1	S_2	S_3	···	S_{L-2}	S_{L-1}	S_L	
1	0	1	···	0	0	1	β^*

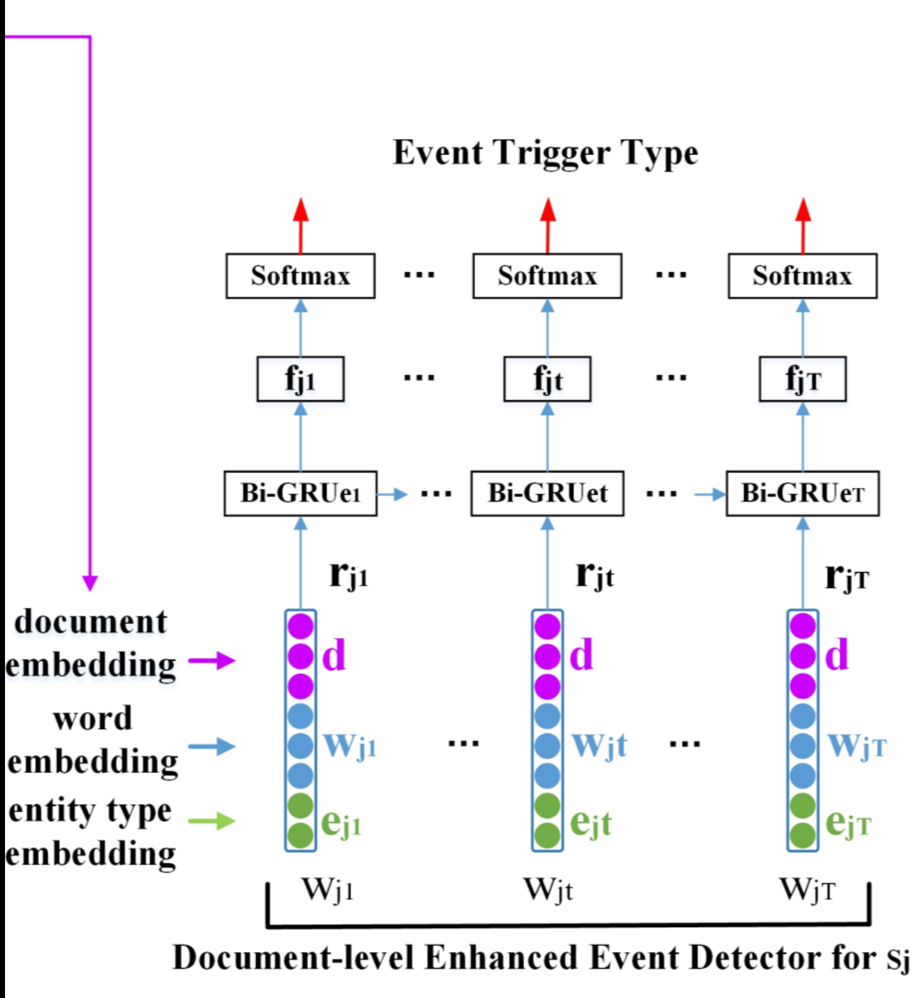
  - S1, S3 and SL are sentences with event triggers and is setted as 1, other sentences are setted as 0.

- Loss function:

$$E_s(\beta^*, \beta) = \sum_{i=1}^{L} (\beta_i^* - \beta_i)^2$$

  - The square error as the general loss of the attention at sentence level to supervise the learning process.

Model - Document-level Enhanced Event Detector



- **Event Detector:**

$$f_{jt} = \text{Bi-GRU}_e([d, w_{jt}, e_{jt}])$$

- softmax output layer to get the predicted probability for each word

- **Loss function:**

$$J(y,o) = -\sum_{j=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{I}(y_{jt} = k)\log o_{jt}^{(k)}$$

- cross-entropy error

Model - Joint Training

Joint Loss Function:

$$J(\theta) = \sum_{\forall d \in \phi} (J(y,o) + \lambda E_w(\alpha^*, \alpha) + \mu E_s(\beta^*, \beta))$$

- θ denotes all parameters used in DEEB-RNN
- ϕ is the training document set
- λ and μ are hyper-parameters for striking a balance

Experiments

ACE 2005 Corpus

- 33 categories
- 6 sources
- 599 documents
- 5349 labeled events

English								
words					files			
1P	DUAL	ADJ	NORM	1P	DUAL	ADJ	NORM	
NW	60658	57807	33459	48399	128	124	81	106
BN	59239	58144	52444	55967	239	234	217	226
BC	46612	46110	33874	40415	68	67	52	60
WL	45210	43648	35529	37897	127	122	114	119
UN	45161	44473	26371	37366	58	57	37	49
CTS	47003	47003	34868	39845	46	46	34	39
Total	303833	297185	216545	259889	666	650	535	599

Experiments - Configuration

Partitions	#Documents
Training set	529
Validation set	30
Test set	40

Parameters	Setting
$\mathrm{GRU}_w, \mathrm{GRU}_s, \mathrm{GRU}_e$	300, 200, 300
W_w, W_s	600, 400
entity type embeddings	50 (randomly initialized)
word embeddings	300 (Google pre-trained)
dropout rate	0.5
training	SGD

Experiments – Model analysis

Model Variants:

- DEEB-RNN computes attentions without supervision
- DEEB-RNN1 uses only the gold word-level attention signal
- DEEB-RNN2 uses only the gold sentence-level attention signal
- DEEB-RNN3 employs the gold attention signals at both word and sentence levels

Methods	λ	μ	P	R	F_1
Bi-GRU	-	-	66.2	72.3	69.1
DEEB-RNN	0	0	69.3	75.2	72.1
DEEB-RNN1	1	0	70.9	76.7	73.7
DEEB-RNN2	0	1	72.3	74.5	73.4
DEEB-RNN3	1	1	72.3	75.8	74.0

- Models with document embeddings outperform the pure Bi-GRU method.
- The model with both gold attention signals at word and sentence levels performs best.

Experiments - Baselines

- Feature-based methods without document-level information :
  - Sentence-level(2011), Joint Local(2013)
- Representation-based methods without document-level information :
  - JRNN(2016), Skip-CNN(2016), ANN-S2(2017)
- Feature-based methods using document level information :
  - Cross-event(2010), PSL(2016)
- Representation-based methods using document-level information :
  - DLRNN(2017)

Experiments – Main Results

Traditional Event Detection Models
- Feature-based without Document-level
- Representation-based without Document-level
- Using Document-level

DEEB Models

Methods	P	R	F_1
Sentence-level (2011)	67.6	53.5	59.7
Joint Local (2013)	73.7	59.3	65.7
JRNN (2016)	66.0	73.0	69.3
Skip-CNN (2016)	N/A	N/A	71.3
ANN-S2 (2017)	78.0	66.3	71.7
Cross-event (2010)†	68.7	68.9	68.8
PSL (2016)†	75.3	64.4	69.4
DLRNN (2017)†	77.2	64.9	70.5
DEEB-RNN1†	70.9	76.7	73.7
DEEB-RNN2†	72.3	74.5	73.4
DEEB-RNN3†	72.3	75.8	74.0

- Our models consistently out-perform the existing state-of-the-art methods in terms of both recall and F1-measure.

Summary

Conclusions

- We proposed a hierarchical and supervised attention based and document embedding enhanced Bi-RNN method.
- We explored different strategies to construct gold word- and sentence-level attentions to focus on event information.
- We also showed this method achieves best performance in terms of both recall and F1-measure.

Future work

- Automatically determine the weights of sentence and document embeddings.
- Use the architecture for another text task.

Thank you for your attention！

Q&A

Name： Yue Zhao
Email： zhaoyue@software.ict.ac.cn