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
- Motivation
- Model
- Experiments
- Summary
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
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.

  ![Attention signal example]

  - “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.
Sentence-level embeddings

- Sentence encoder
  \[ q_i = \text{Bi-GRU}_s(s_i) \]

- Sentence attention
  \[ t_i = \tanh(W_s q_i) \]

- Document representation
  \[ \beta_i = t_i^T c_s \]

\[ d = \sum_{i=1}^{L} \beta_i s_i \]
Model - ED Oriented Document Embedding Learning

- Gold **sentence-level** attention signal:

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>...</th>
<th>SL-2</th>
<th>SL-1</th>
<th>SL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- 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:**

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

  - Cross-entropy error
Model - Joint Training

Joint Loss Function:

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

- $\theta$ denotes all parameters used in DEEB-RNN
- $\phi$ is the training document set
- $\lambda$ and $\mu$ are hyper-parameters for striking a balance
Experiments

ACE 2005 Corpus

- 33 categories
- 6 sources
- 599 documents
- 5349 labeled events
## Experiments - Configuration

<table>
<thead>
<tr>
<th>Partitions</th>
<th>#Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>529</td>
</tr>
<tr>
<td>Validation set</td>
<td>30</td>
</tr>
<tr>
<td>Test set</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{GRU}_w, \text{GRU}_s, \text{GRU}_e$</td>
<td>300, 200, 300</td>
</tr>
<tr>
<td>$W_w, W_s$</td>
<td>600, 400</td>
</tr>
<tr>
<td>entity type embeddings</td>
<td>50 (randomly initialized)</td>
</tr>
<tr>
<td>word embeddings</td>
<td>300 (Google pre-trained)</td>
</tr>
<tr>
<td>dropout rate</td>
<td>0.5</td>
</tr>
<tr>
<td>training</td>
<td>SGD</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$P$</th>
<th>$R$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-GRU</td>
<td>-</td>
<td>-</td>
<td>66.2</td>
<td>72.3</td>
<td>69.1</td>
</tr>
<tr>
<td>DEEB-RNN</td>
<td>0</td>
<td>0</td>
<td>69.3</td>
<td>75.2</td>
<td>72.1</td>
</tr>
<tr>
<td>DEEB-RNN1</td>
<td>1</td>
<td>0</td>
<td>70.9</td>
<td>76.7</td>
<td>73.7</td>
</tr>
<tr>
<td>DEEB-RNN2</td>
<td>0</td>
<td>1</td>
<td>72.3</td>
<td>74.5</td>
<td>73.4</td>
</tr>
<tr>
<td>DEEB-RNN3</td>
<td>1</td>
<td>1</td>
<td>72.3</td>
<td>75.8</td>
<td>74.0</td>
</tr>
</tbody>
</table>

- 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
- Using Document-level Representation

DEEB Models

<table>
<thead>
<tr>
<th>Methods</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-level (2011)</td>
<td>67.6</td>
<td>53.5</td>
<td>59.7</td>
</tr>
<tr>
<td>Joint Local (2013)</td>
<td>73.7</td>
<td>59.3</td>
<td>65.7</td>
</tr>
<tr>
<td>JRNN (2016)</td>
<td>66.0</td>
<td>73.0</td>
<td>69.3</td>
</tr>
<tr>
<td>Skip-CNN (2016)</td>
<td>N/A</td>
<td>N/A</td>
<td>71.3</td>
</tr>
<tr>
<td>ANN-S2 (2017)</td>
<td>78.0</td>
<td>66.3</td>
<td>71.7</td>
</tr>
<tr>
<td>Cross-event (2010)†</td>
<td>68.7</td>
<td>68.9</td>
<td>68.8</td>
</tr>
<tr>
<td>PSL (2016)†</td>
<td>75.3</td>
<td>64.4</td>
<td>69.4</td>
</tr>
<tr>
<td>DLRNN (2017)†</td>
<td>77.2</td>
<td>64.9</td>
<td>70.5</td>
</tr>
<tr>
<td>DEEB-RNN1†</td>
<td>70.9</td>
<td>76.7</td>
<td>73.7</td>
</tr>
<tr>
<td>DEEB-RNN2†</td>
<td>72.3</td>
<td>74.5</td>
<td>73.4</td>
</tr>
<tr>
<td>DEEB-RNN3†</td>
<td>72.3</td>
<td>75.8</td>
<td>74.0</td>
</tr>
</tbody>
</table>

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