What the role is vs. What plays the role: Semi-supervised Event Argument Extraction via Dual Question Answering

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, JinLong Li
Background

Event Mention: He claimed Iraqi troops had destroyed five tanks

Event Detection:
• Event type: Attack
• Trigger: destroyed

Event Argument Extraction:
• Attacker: Iraqi troops
• Target: five tanks
Motivation

• Event Argument Extraction become the bottleneck:
  Event detection has gained great popularity and reached a fairly high performance (Wang et al. 2019), event argument extraction becomes the key to event extraction.

• Data sparse:
  According to our statistics, about 60% event types in ACE 2005 English corpus (Doddington et al. 2004) have less than 100 labeled samples and only 1.11% events in ACE 2005 have all roles that the type should contain.


Motivation

• Model
  • Insufficient parameter sharing
    • Previous studies always model different roles separately
  • Insufficient utilizing semantics of the roles
    • Previous studies treat the roles as labels, without allowing the model to understand the meaning of labels.

• Data
  • Rely heavily on external resources
Method

• Model
  • We formulate the EAE as Machine Reading Comprehension (MRC)
  • Define EAR (Primal Task) and ERR (Dual Task)

• Data
  • Design a dual training process
Method

• **Question Generation**
  
  • **EAR:** What plays the role $x_r$ in $x_{ts}$? ($x_d^1, \ldots, x_d^n$)
  
  • **ERR:** What is the role of $x_a$ in $x_{ts}$?

• **Example**
  
  • *Destination: destination* is a type of goal; terminus is a translation of *destination*.
Method

- **Instance Encode**
  - BERT
  - Parameter-sharing

- **Flow Attention**
  - Parameter-sharing

- **Classifier**
  - MLP
  - CNN
Method

• Semi-supervised Dual Training Strategy
  • Joint Train
  • Optimize alternative
  • Mutual
  \[
  O(\theta) = \max(\mathbb{E}_{(c,r,a) \in S_A} [\log(p(a|c,r,\theta))]) \\
  = \min(-\mathcal{L}_A(S_A, \theta)) \\
  = -\min \sum_{k=1}^{|S_A|} \log(p(a^k_s|c^k,r^k,\theta)) \\
  + \log(p(a^k_c|c^k,r^k,\theta)),
  \]

  \[
  O(\phi) = \max(\mathbb{E}_{(c,r,a) \in S_R} [\log(p(r|c,a,\phi))]) \\
  = \min(-\mathcal{L}_R(S_R, \phi)) \\
  = -\min \sum_{k=1}^{|S_R|} \log(p(r^k|c^k,a^k,\phi)),
  \]

• Label Data
  • Verify each other

Algorithm 1 DualQA Learning Algorithm

Input: Labeled data $S_A = \{(c_i, a_i, r_i)\}_{i=1}^{|S_A|}$ and $S_R = \{(c_i, a_i, r_i)\}_{i=1}^{|S_R|}$, unlabeled data $S_U = \{(c_j)\}_{j=1}^{|S_U|}$

1: while $S_U \neq \emptyset$ and not converge do
2: $M^a_{\theta}, M^r_{\phi} \leftarrow$ Initialize
3: $M^a_{\theta}, M^r_{\phi} \leftarrow$ Joint train using $S_A$ and $S_R$ (Eq. 10)
4: for all $c_j$ in $S_U$ do
5: for all $r$ in event schema of $c_j$ do
6: $\hat{a} \leftarrow M^a_{\theta}(c_j, r)$
7: $\hat{r} \leftarrow M^r_{\phi}(c_j, \hat{a})$
8: if $\hat{a}$ not neg and $\hat{r}$ not neg and $\hat{r} = r$ then
9: $\text{Append}(c_j, \hat{a}, r)$ to $S_A$ and $S_R$
10: end if
11: end for
12: for all $a$ in argument candidate of $c_j$ do
13: $\hat{\rho} \leftarrow M^a_{\theta}(c_j, a)$
14: $\hat{a} \leftarrow M^r_{\phi}(c_j, \hat{\rho})$
15: if $\hat{a}$ not neg and $\hat{r}$ not neg and $\hat{a} = a$ then
16: $\text{Append}(c_j, a, \hat{\rho})$ to $S_A$ and $S_R$
17: end if
18: end for
19: if all role of $c_j$ and all argument related to $c_j$ has credible answer then
20: Remove $(c_j)$ from $S_U$
21: end if
22: end for
23: end while

Output: Enhanced $M^a_{\theta}$
Experiments

• Experimental Settings
  • Dataset
    • We choose two public event extraction datasets from completely different fields to validate the effectiveness and annotation ability of our method.
  • Data Settings
    • labeled set, unlabeled set form training set.
  • Baselines
    • BERT-based baselines and their enhanced versions
Experiments

Comparisons with SOTA methods

- In ACE 2005 English corpus, we sample 10% training data as labeled set and 60% training data as unlabeled set.
- In FewFC, we sample 1% training data as labeled set and 60% training data as unlabeled set.
- Under low-resource settings, DualQA can outperform SOTA methods.
- We have high precision.

<table>
<thead>
<tr>
<th>Method/Dataset</th>
<th>ACE P</th>
<th>R</th>
<th>F1</th>
<th>FewFC P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-EE(Devlin et al.)</td>
<td>26.7</td>
<td>38.2</td>
<td>31.4</td>
<td>18.9</td>
<td>35.9</td>
<td>24.8</td>
</tr>
<tr>
<td>BERT-EE*</td>
<td>28.3</td>
<td>41.9</td>
<td>33.8</td>
<td>19.4</td>
<td>37.6</td>
<td>25.6</td>
</tr>
<tr>
<td>PLMEE(Yang et al.)</td>
<td>36.3</td>
<td>46.8</td>
<td>40.9</td>
<td>52.0</td>
<td>30.9</td>
<td>38.8</td>
</tr>
<tr>
<td>PLMEE*</td>
<td>37.6</td>
<td>46.6</td>
<td>41.6</td>
<td>54.1</td>
<td>31.9</td>
<td>40.2</td>
</tr>
<tr>
<td>DualQA</td>
<td>49.1</td>
<td>42.3</td>
<td><strong>45.4</strong></td>
<td>57.4</td>
<td>34.4</td>
<td><strong>43.1</strong></td>
</tr>
</tbody>
</table>
Ablation Study

- The effectiveness of MRC framework
  - MRC-based methods make significant improvements compare with the sequence labeling model.

- The effectiveness of dual learning
  - Our approach is more efficient in benefiting from unlabeled data.
  - Dual learning leads to high precision.
  - Best model will get in the middle of training epoch
Ablation Study

• The effectiveness under different amounts of labeled data.
  • Our approach is more robust than the baseline under extremely low resource situations.

• The quality of annotations.
  • The annotations quality of our method outperforms other methods.
THANKS

2020.12.19