国际人工智能会议
AAAI 2021 论文北京预讲会

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.

Wang, X.; Wang, Z.; Han, X.; Liu, Z.; Li, J.; Li, P.; Sun, M.; Zhou, J.; and Ren, X. 2019. HMEAE: Hierarchical Modular Event Argument Extraction. In Proceedings of the EMNLP- IJCNLP 2019, 5776–5782. Association for Computational Linguistics
Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In Lrec, volume 2, 837–840. Lisbon.

Motivation

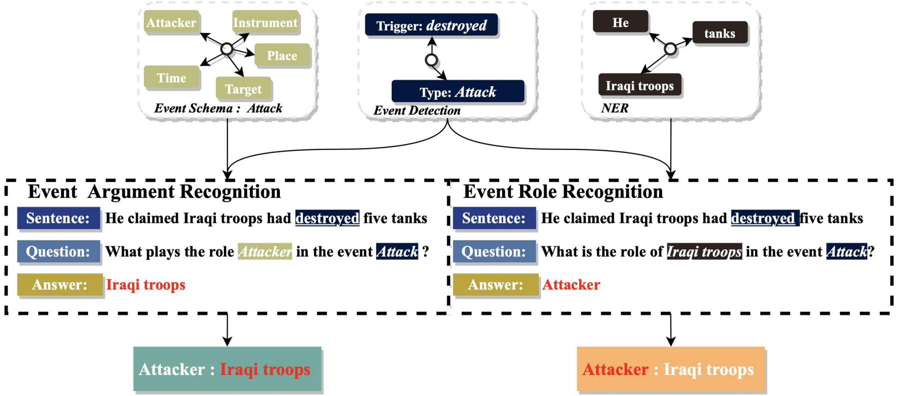
- **Model**
  - Insufficient parameter sharing
    - Previous studies always model different roles separately
  - Insufficient utilizing semantics of the roles
    - Previous studies treat the roles as labels, without allowing the model to understand the meaning of labels.

- **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



中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

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*.

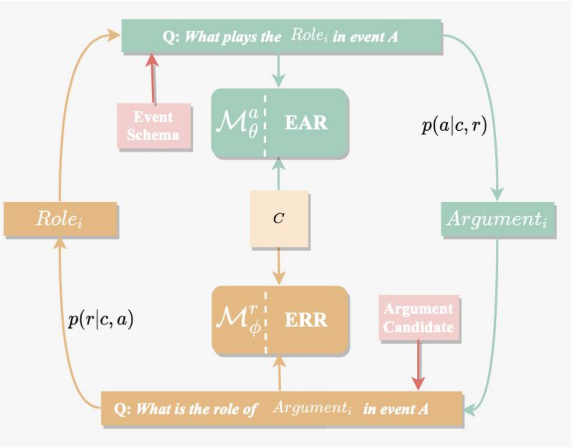
ConceptNet

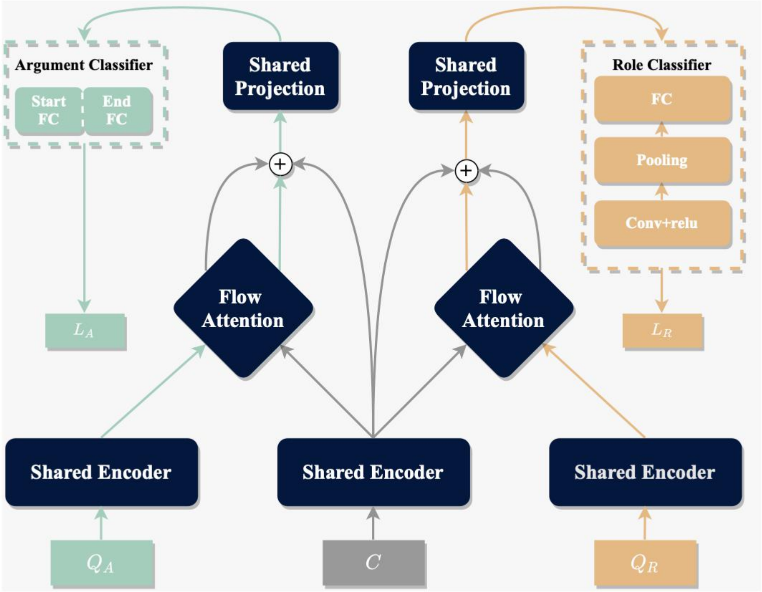
An open, multilingual knowledge graph

Speer, R.; and Havasi, C. 2013. ConceptNet 5: A Large Se- mantic Network for Relational Knowledge. In The People's Web Meets NLP, Collaboratively Constructed Language Re- sources, Theory and Applications of Natural Language Pro- cessing, 161–176. Springer.

Method

- **Instance Encode**
  - BERT
  - Parameter-sharing
- **Flow Attention**
  - Parameter-sharing
- **Classifier**
  - MLP
  - CNN





Method

- **Semi-supervised Dual Training Strategy**
  - Joint Train
    - Optimize alternative
    - Mutual

$$
\begin{aligned}
\mathbf{O}(\theta) &= max(\mathbb{E}_{(c,r,a)\in S_A}[log(p(a|c,r,\theta))]) \\
&= min(-\mathcal{L}_A(S_A,\theta)) \\
&= -min\sum_{k=1}^{|S_A|}(log(p(a_s^k|c^k,r^k,\theta)) \\
&+ log(p(a_e^k|c^k,r^k,\theta))),
\end{aligned}
$$

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

  - 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 $S_U = \{(c_j)\}_{j=1}^{|S_U|}$

while $S_U \neq \emptyset$ **and** *not converge* **do**
  $\mathcal{M}_\theta^a, \mathcal{M}_\phi^r \leftarrow$ Initialize
  $\mathcal{M}_\theta^a, \mathcal{M}_\phi^r \leftarrow$ Joint train using S_A and S_R (Eq. 10)
  for all c_j in S_U **do**
    for all r in event schema of c_j **do**
      $\hat{a} \leftarrow \mathcal{M}_\theta^a(c_j, r)$
      $\hat{r} \leftarrow \mathcal{M}_\phi^r(c_j, \hat{a})$
      if $\hat{a}$ not *neg* and $\hat{r}$ not *neg* and $\hat{r} = r$ **then**
        Append $(c_j, \hat{a}, r)$ to S_A and S_R
      end if
    end for
    for all a in argument candidate of c_j **do**
      $\hat{r} \leftarrow \mathcal{M}_\phi^r(c_j, a)$
      $\hat{a} \leftarrow \mathcal{M}_\theta^a(c_j, \hat{r})$
      if $\hat{a}$ not *neg* and $\hat{r}$ not *neg* and $\hat{a} = a$ **then**
        Append $(c_j, a, \hat{r})$ to S_A and S_R
      end if
    end for
    if all role of c_j and all argument related to c_j has credible answer **then**
      Remove (c_j) from S_U
    end if
  end for
end while

Output: Enhanced $\mathcal{M}_\theta^a$

Experiments

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

Experiments

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

- **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.

Ablation Study

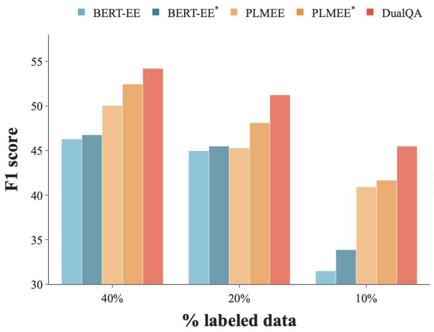
Method/Dataset	ACE			FewFC		
	P	R	F1	P	R	F1
BERT-EE(Devlin et al.)	26.7	38.2	31.4	18.9	35.9	24.8
BERT-EE*	28.3	41.9	33.8	19.4	37.6	25.6
EAR	33.6	42.6	37.5	34.8	28.4	32.0
EAR*	44.2	35.4	39.3	40.0	30.2	34.4
DualQA	49.1	42.3	**45.4**	57.4	34.4	**43.1**

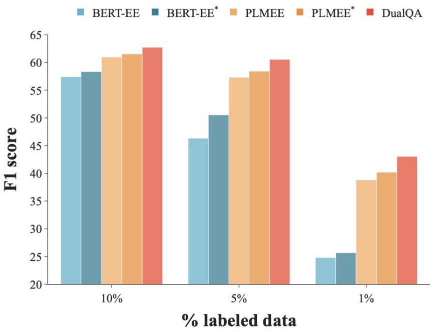
Method/Dataset	ACE			FewFC		
	P	R	F1	P	R	F1
PLMEE(Yang et al.)	36.3	46.8	40.9	52.0	30.9	38.8
PLMEE*	37.6	46.6	41.6	54.1	31.9	40.2
EAR	33.6	42.6	37.5	34.8	28.4	32.0
Joint-EAR-ERR	40.5	42.2	41.4	40.0	43.0	41.5
DualQA	49.1	42.3	**45.4**	57.4	34.4	**43.1**

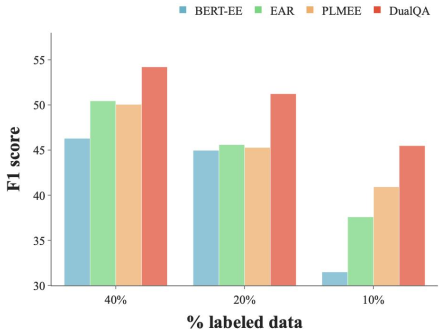
- **The effectiveness of MRC framework**
  - MRC-based methods make significant improvements compare with the sequence labeling model.
- **The effectiveness of dual learning**
  - Our approach is more efficient in benefiting from unlabeled data.
  - Dual learning leads to high precision.
  - Best model will get in the middle of training epoch

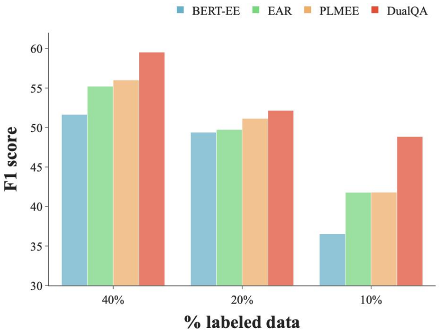
中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

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

中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19