A Joint Neural Model for Information Extraction with Global Features

Ying Lin, Heng Ji, Fei Huang, Lingfei Wu

ACL2020
Tasks

• This work jointly performs three tasks of Information Extraction at sentence-level:
  + Entity Extraction.
  + Relation Extraction.
  + Event Extraction.
Model

- Their model performs all the tasks in four stages:
  + Encoding tokens.
  + Identifying nodes (i.e., triggers or entity mentions).
  + Scoring nodes and edges (i.e., relations or argument roles).
  + Searching for best graph.
Model

Decoding
- Beam search

Classification
- Score vectors

Identification

Encoding
- The earthquake killed 19 people and injured 300 in Kashmir region, India
Model: Encoding Tokens

- Input: a sentence of L words.
- Uses BERT as the encoder.
- Word representations are the averaged vector of their wordpiece representations.
Model: Identifying Nodes

• BERT outputs are fed into a Feed-Forward Net to obtain score vectors.
  \[ \hat{y}_i = \text{FFN}(x_i) \]

• Node identification is formalized as a sequence labeling task (e.g., B-Life:Marry, B-GPE) with a CRF layer.
  \[ s(X, \hat{z}) = \sum_{i=1}^{L} \hat{y}_{i, \hat{z}_i} + \sum_{i=1}^{L+1} A_{\hat{z}_{i-1}, \hat{z}_i}, \quad \hat{z} = \{\hat{z}_1, \ldots, \hat{z}_L\} \]
Model: Scoring Nodes and edges

• At this layer, the model computes node and edge representations.
  + Node: average sum over its component words.
  + Edge: concatenation of node representations.

• Then, score vectors for nodes \( \hat{y}_i^t = \text{FFN}^t(v_i) \) & edges \( \hat{y}_k^t = \text{FFN}^t(v_i, v_j) \) are computed via softmax layers.

• Node that, the model does not make predictions here.
Model: Searching for Best Graph

- With the score vectors obtained from the previous step. They use Beam search to efficiently find the configuration with the highest score.

- Score of a graph is computed by:

\[
\begin{align*}
    s(G) &= s'(G) + uf_G \\
    s'(G) &= \sum_{t \in T} \sum_{i=1}^{N_t} \max \hat{y}_i
\end{align*}
\]

Where:

- \( s'(G) \) is local score: \( s'(G) = \sum_{t \in T} \sum_{i=1}^{N_t} \max \hat{y}_i \)

- \( uf_G \) is global score where \( f_G = \{f_1(G), ..., f_M(G)\} \) global feature vector
Model: Global Features

- Global features are introduced to capture cross-subtask and cross-instance dependencies.

| Category | Description |
|----------|-------------|
| Role     | 1. The number of entities that act as \(<role_i>\) and \(<role_j>\) arguments at the same time.  
          2. The number of \(<event\_type_i>\) events with \(<number>\) \(<role_j>\) arguments.  
          3. The number of occurrences of \(<event\_type_i>, <role_j>, \) and \(<entity\_type_k>\) combination.  
          4. The number of events that have multiple \(<role_i>\) arguments.  
          5. The number of entities that act as a \(<role_i>\) argument of an \(<event\_type_j>\) event and a \(<role_k>\) argument of an \(<event\_type_1>\) event at the same time.  |
| Relation | 6. The number of occurrences of \(<entity\_type_i>, <entity\_type_j>, \) and \(<relation\_type_k>\) combination.  
          7. The number of occurrences of \(<entity\_type_i>\) and \(<relation\_type_j>\) combination.  
          8. The number of occurrences of a \(<relation\_type_i>\) relation between a \(<role_j>\) argument and a \(<role_k>\) argument of the same event.  
          9. The number of entities that have a \(<relation\_type_i>\) relation with multiple entities.  
          10. The number of entities involving in \(<relation\_type_i>\) and \(<relation\_type_j>\) relations simultaneously.  |
| Trigger  | 11. Whether a graph contains more than one \(<event\_type_i>\) event.  |
Training

- Identification loss: negative log-likelihood
  \[ L^I = - \log p(z | X) = -s(X, z) + \log \sum_{\hat{z} \in Z} e^{s(X, \hat{z})} \]

- Classification loss: cross-entropy
  \[ L^t = - \frac{1}{N^t} \sum_{i=1}^{N^t} y_i^t \log \hat{y}_i^t \]

- Global feature constraint: the ground-truth graph \( G \) should be the one with the highest score. Minimize this:
  \[ L^G = s(\hat{G}) - s(G) \]

- Overall loss: \( L = L^I + \sum_{t \in T} L^t + L^G \)
Experiments

• Datasets: ACE, ERE

| Dataset     | Split | #Sents | #Entities | #Rel | #Events |
|-------------|-------|--------|-----------|------|---------|
| ACE05-R     | Train | 10,051 | 26,473    | 4,788| -       |
|             | Dev   | 2,424  | 6,362     | 1,131| -       |
|             | Test  | 2,050  | 5,476     | 1,151| -       |
| ACE05-E     | Train | 17,172 | 29,006    | 4,664| 4,202   |
|             | Dev   | 923    | 2,451     | 560  | 450     |
|             | Test  | 832    | 3,017     | 636  | 403     |
| ACE05-CN    | Train | 6,841  | 29,657    | 7,934| 2,926   |
|             | Dev   | 526    | 2,250     | 596  | 217     |
|             | Test  | 547    | 2,388     | 672  | 190     |
| ACE05-E+    | Train | 19,240 | 47,525    | 7,152| 4,419   |
|             | Dev   | 902    | 3,422     | 728  | 468     |
|             | Test  | 676    | 3,673     | 802  | 424     |
| ERE-EN      | Train | 14,219 | 38,864    | 5,045| 6,419   |
|             | Dev   | 1,162  | 3,320     | 424  | 552     |
|             | Test  | 1,129  | 3,291     | 477  | 559     |
| ERE-ES      | Train | 7,067  | 11,839    | 1,698| 3,272   |
|             | Dev   | 556    | 886       | 120  | 210     |
|             | Test  | 546    | 811       | 108  | 269     |
Experiments

- Monolingual performance on English language:

| Dataset  | Task   | DYGIE++ | Baseline | OneIE |
|----------|--------|---------|----------|-------|
| ACE05-R  | Entity | 88.6    | -        | 88.8  |
|          | Relation | 63.4    | -        | 67.5  |
| ACE05-E  | Entity | 89.7    | 90.2     | 90.2  |
|          | Trig-I | -       | 76.6     | 78.2  |
|          | Trig-C | 69.7    | 73.5     | 74.7  |
|          | Arg-I  | 53.0    | 56.4     | 59.2  |
|          | Arg-C  | 48.8    | 53.9     | 56.8  |
Experiments

- Multilingual performance (with additional English data) on Chinese and Spanish.

| Dataset   | Training | Entity | Relation | Trig-C | Arg-C |
|-----------|----------|--------|----------|--------|-------|
| ACE05-CN  | CN       | 88.5   | 62.4     | 65.6   | 52.0  |
|           | CN+EN    | 89.8   | 62.9     | 67.7   | 53.2  |
| ERE-ES    | ES       | 81.3   | 48.1     | 56.8   | 40.3  |
|           | ES+EN    | 81.8   | 52.9     | 59.1   | 42.3  |