A Joint Neural Model for Information Extraction with Global Features

Ying Lin, Heng Ji , Fei Huang, Lingfei Wu

ACL2020

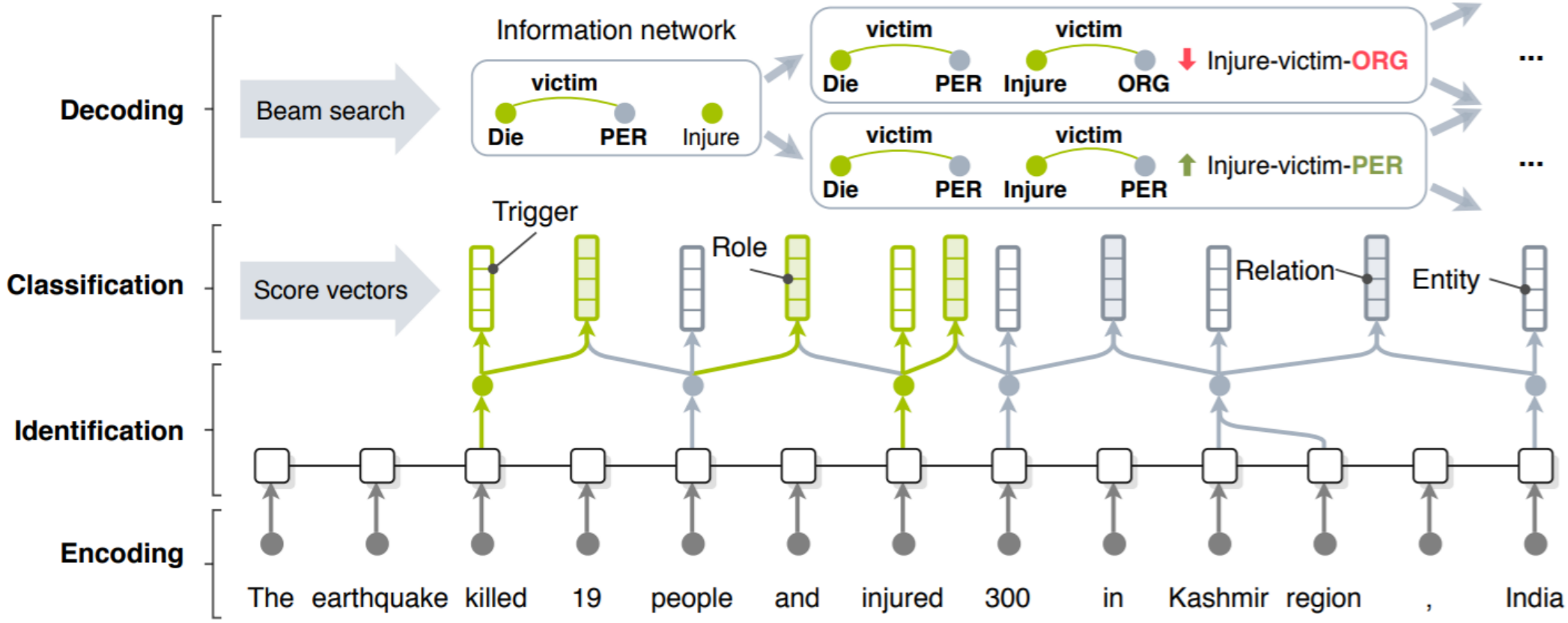
Tasks

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

Model

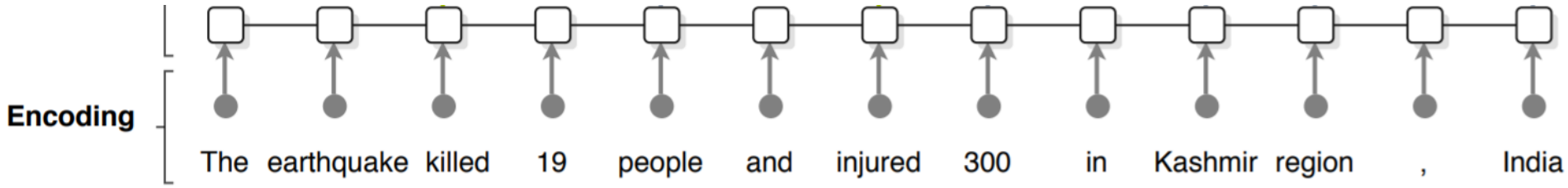
- Their model performs all the tasks in four stages:
  + Encoding tokens.
  + Identifying nodes (i.e., triggers or entity mentions).
  + Scoring nodes and edges (i.e., relations or argument roles).
  + Searching for best graph.

Model


Decoding
Classification
Identification
Encoding
Beam search
Information network
victim
Die
PER
Injure
ORG
Injure-victim-ORG
Injure-victim-PER
Score vectors
Trigger
Role
Relation
Entity
The
earthquake
killed
19
people
and
injured
300
in
Kashmir
region
,
India

Model: Encoding Tokens

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



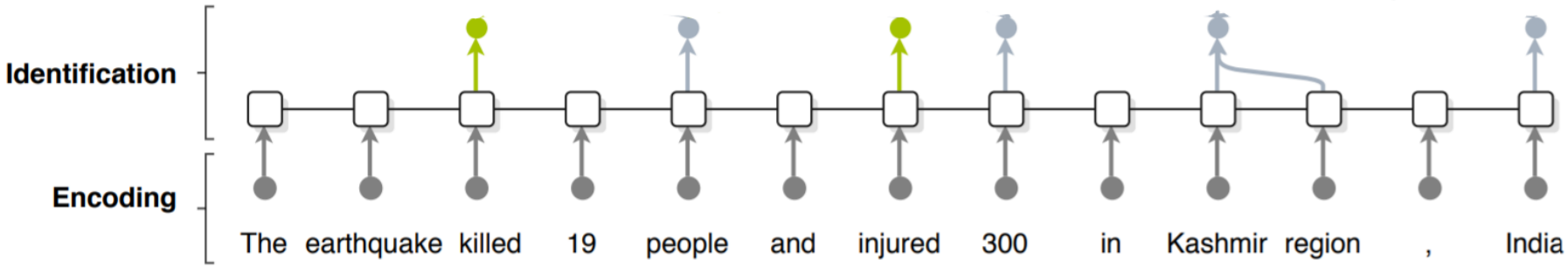
Model: Identifying Nodes

- BERT outputs are fed into a Feed-Forward Net to obtain score vectors.

$$\hat{\boldsymbol{y}}_i = \mathrm{FFN}(\boldsymbol{x}_i)$$

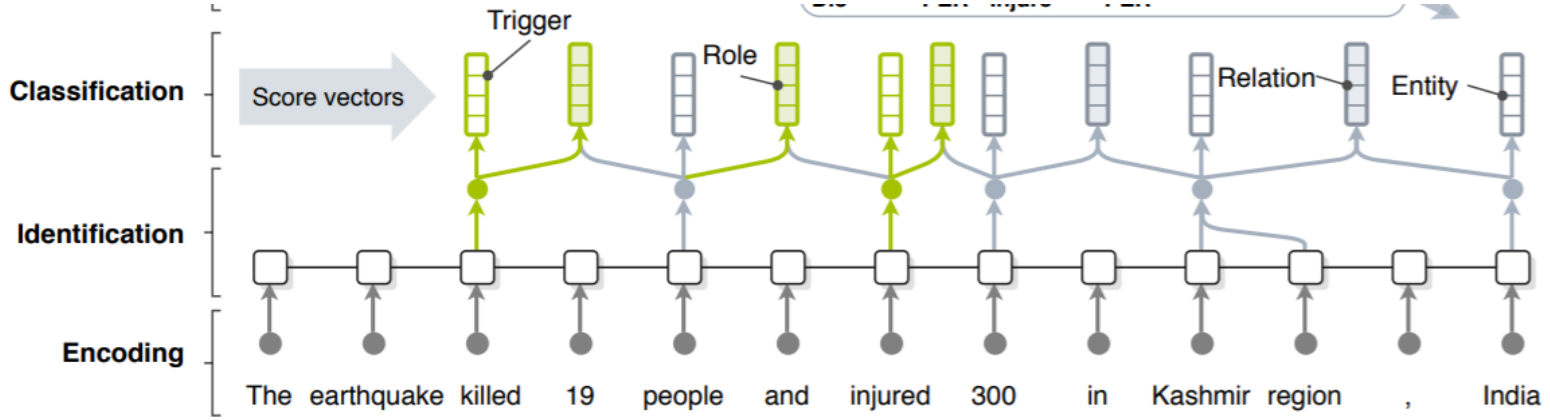
- Node identification is formalized as a sequence labeling task (e.g., B-Life:Marry, B-GPE) with a CRF layer.

$$s(\boldsymbol{X}, \hat{\boldsymbol{z}}) = \sum_{i=1}^{L} \hat{y}_{i,\hat{z}_i} + \sum_{i=1}^{L+1} A_{\hat{z}_{i-1},\hat{z}_i}, \quad \hat{\boldsymbol{z}} = \{\hat{z}_1, \ldots, \hat{z}_L\}$$



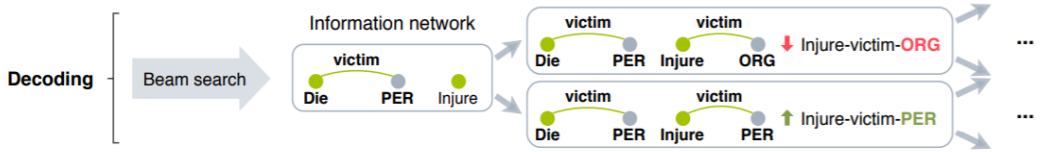
Model: Scoring Nodes and edges

- 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 ($\hat{\boldsymbol{y}}_i^t = \mathrm{FFN}^t(\boldsymbol{v}_i)$) & edges ($\hat{\boldsymbol{y}}_k^t = \mathrm{FFN}^t(\boldsymbol{v}_i, \boldsymbol{v}_j)$) 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:

$$s(G) = s'(G) + \boldsymbol{u}\boldsymbol{f}_G$$

Where:

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

+ $\boldsymbol{u}\boldsymbol{f}_G$ is global score where $\boldsymbol{f}_G = \{f_1(G), ..., f_M(G)\}$ global feature vector

Model: Global Features

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

Categary	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$_l$> 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

$$\mathcal{L}^{\mathrm{I}} = -\log p(\boldsymbol{z}|\boldsymbol{X}) = -s(\boldsymbol{X}, \boldsymbol{z}) + \log \sum_{\hat{\boldsymbol{z}} \in Z} e^{s(\boldsymbol{X}, \hat{\boldsymbol{z}})}$$

- Classification loss: cross-entropy

$$\mathcal{L}^{\mathrm{t}} = -\frac{1}{N^t} \sum_{i=1}^{N^t} \boldsymbol{y}_i^t \log \hat{\boldsymbol{y}}_i^t$$

- Global feature constraint: the ground-truth graph G should be the one with the highest score. Mimize this:

$$\mathcal{L}^{\mathrm{G}} = s(\hat{G}) - s(G)$$

- Overall loss: $\mathcal{L} = \mathcal{L}^I + \sum_{t \in T} \mathcal{L}^t + \mathcal{L}^G$

Experiments

- Datasets: ACE, ERE

Dataset	Split	#Sents	#Entities	#Rels	#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