SRCB at TAC KBP 2017 Event Nugget Track

Data Mining Lab
Ricoh Software Research Center (Beijing) Co., Ltd.

(TeamID: srcb)
Preprocessing includes:
- Sentence boundary
- Tokenization
- Vocabulary setup
- Word embedding
EN Detection and Classification

• To detect explicit mentions of relevant events and identify event types and subtypes.
• A sequence labelling problem.

• An ensemble model which combines a neural network model and a Conditional Random Fields (CRFs) model.

He said the killings were in self-defense and he fled the state because he could not get a fair trial.

Leak information about US war crimes, and spend the rest of your life in a military prison.
Neural Network Model

• **Bidirectional LSTM** is adopted to capture both past and future contexts for a given word.

• 18 event labels are defined according to 8 event types and 18 subtypes.

• Continuous words with the same type label are regarded as the same event mention with the specified type value.
CRFs Model

Features:
• Token
• Lemma
• Stemming
• POS tag
• Dependency type
• NER nearby
• Position in sentence
• Sentence position in document
• Trigger word dict
• WordNet

33 labels:
• 1 subtype
• 2 subtypes
The whole training dataset is denoted as D.

**En_BiLSTM**
- The dataset D is split into 10 parts. 10 BiLSTM models are trained separately using corresponding 9 of 10 as training data and the remaining as validation data. The training dataset used for model i is denoted as Di.
- For each Di, *over-sampling* technique is adopted to increase the number of event labels with fewer instances.
- **Voting strategy** is adopted to combine outputs of the 10 BiLSTM models.

**Ensemble Model**
- Combine: En_BiLSTM and CRFs outputs.
- Strategy: If conflict happens between two models, the results of CRFs are kept.
To identify three REALIS values for event mentions (Actual, Generic, Other)

He said the killings were in self-defense and he fled the state because he could not get a fair trial.

Leak information about US war crimes, and spend the rest of your life in a military prison.
A SVM model is used to classify REALIS value for event mentions.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Word itself</td>
</tr>
<tr>
<td>POS tag</td>
<td>POS tag of current word</td>
</tr>
<tr>
<td>NER nearby</td>
<td>NER tags for nearby words</td>
</tr>
<tr>
<td>Tense</td>
<td>Whether the word is ended with “ed” or not</td>
</tr>
<tr>
<td>WordNet</td>
<td>WordNet index</td>
</tr>
</tbody>
</table>
EN Coreference

To Identify the coreference links between event mention instances within a document.

While U.S. space program is reduced to begging the Russians for rides Guest post by Investor's Business Daily Unexceptional …

Meanwhile, U.S. astronauts have to ride Russian spacecraft to fix toilets on the International Space Station.
Classification problem: 2-pass sieve method. The method was proved significantly efficient by UTD team @TAC2016.

1) training pair’s subtype is the same as test pair’s;
2) training pair’s lemmatized nuggets are the same as test pair’s;
3) the sentence distance between two nuggets of training pair is in the range that of test pair plus or minus $m$. 

1) nuggets of test pair have the same lemma;
2) nuggets of neighbor training pair have the same lemma, but not necessarily identical with test pair’s;
3) training pair’s subtype is the same as test pair’s.
Experiments

Datasets

- For development, LDC2017E02 (2014 and 2015), LDC2015E29 and LDC2015E68 are used as training data, and LDC2017E02 (2016) as testing data.
- For evaluation on 2017 datasets, LDC2016E31 and LDC2017E02 (2016) are further included as training data.

Neural network training

- Construct one vocabulary including most frequent words in documents. Words that are not in vocabulary are labeled by a special token “UNK”.
- Word embedding are pre-trained using Wikipedia English corpus.
- The training stage of each model took about 1.5 hours.
## Performance on development data

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Micro Average</th>
<th>Macro Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td><strong>Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain</td>
<td>61.81</td>
<td>59.88</td>
</tr>
<tr>
<td>mention_type</td>
<td>51.92</td>
<td>51.02</td>
</tr>
<tr>
<td>realis_status</td>
<td>43.14</td>
<td>44.15</td>
</tr>
<tr>
<td>mention_type+realis_status</td>
<td>37.20</td>
<td>35.90</td>
</tr>
<tr>
<td><strong>Overall Average CoNLL score</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compared with other systems developed in 2016 TAC KBP, our model got better scores on plain and mention_type.
Experiments

On development data

- BiLSTM model outperforms CRFs in F1 measure.
- However, CRFs can identify events in higher precision.
Experiments

On development data

• The ensemble of 10 BiLSTM models (En_BiLSTM) outperform one single BiLSTM model.

• Then ensemble of En_BiLSTM and CRF models outperform En_BiLSTM model.
Results

We submitted 3 runs to En detection and coreference. The best performance of the 3 runs on 2017 official evaluation data are listed below.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Micro Average</th>
<th>Macro Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>plain</td>
<td>68.04</td>
<td>66.53</td>
</tr>
<tr>
<td>mention_type</td>
<td>56.83</td>
<td>55.57</td>
</tr>
<tr>
<td>realis_status</td>
<td>47.95</td>
<td>46.89</td>
</tr>
<tr>
<td>mention_type+realis_status</td>
<td>39.69</td>
<td>38.81</td>
</tr>
<tr>
<td>Overall Average CoNLL score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Our ensemble model significantly outperforms other systems on EN plain and EN mention_type.
- For EN coreference, sieve-based method (srcb1) performs better than ME-based method (srcb2).
THANK YOU!