Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning

Nanyun Peng, Mark Dredze

Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, Maryland USA
Bloomberg LP, New York, NY

Named Entity Recognition
- Detecting boundaries and classifying types of text chunks that correspond to entities:
  ○ persons, organizations, locations

Chinese Word Segmentation
- Detecting word boundaries. A pre-processing step for Chinese language processing tasks.
- Helps NER: NE boundaries are consistent with word boundaries.
- Domain mismatch: No annotations for Chinese word segmentation in social media, but annotations in news domain are abundant. ⇒ Instead of using pre-trained CWS models, we do multi-task joint training.

Datasets
- Segmentation: SIGHAN 2005 shared task, PKU portion, 43,963 sentences for training and 4,278 sentences as dev.
- Weibo NER: 1890 Weibo messages annotated with named entities and nominal mentions.

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Mentions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geo-political</td>
<td>243</td>
<td>243</td>
</tr>
<tr>
<td>Location</td>
<td>88</td>
<td>126</td>
</tr>
<tr>
<td>Organization</td>
<td>224</td>
<td>255</td>
</tr>
<tr>
<td>Person</td>
<td>721</td>
<td>1,357</td>
</tr>
</tbody>
</table>

Multi-task Joint Learning
- Segmentation Module:
  \[ \mathcal{L}_s(y_s; x_s, \Theta) = \frac{1}{K} \sum_k \left[ \log \frac{1}{Z(x_s)^k} \right] + \sum_i \left( T_s(y_{i-1}^k, y_i^k) + s(y_i^k; x_s^k, \Lambda_s) \right) \]
- NER Module:
  \[ \mathcal{L}_n(y_n; x_n, \Theta) = \frac{1}{K} \sum_k \left[ \log \frac{1}{Z(x_n)^k} \right] + \sum_j \Lambda_j F_j(y_n^k, x_n^k, e_w, h_w) \]

Code and data available at: https://github.com/hltcoe/goldenhorse

Segmentation results for named mentions (left) and name + nominal mentions (right) on weibo data.