Abstract

Thai is one of the languages that does not have explicit segmentation similar to languages such as Chinese, Japanese, and Arabic. In this project, we aim to create a BiLSTM-CRF and BiGRU-CRF model to learn Thai segmentation.

Data

Our data set is the BEST2010 corpus. This corpus contains 5 million segmented Thai words and includes named entity and abbreviation tagging.

Methods

Our models are tested on both data sets with named entities and sets with the entities collapsed into a new token. We also vary the size of the training sets (full data set or 10% dataset). Results are evaluated using character level F1.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Without named entities</th>
<th>With named entities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1-Micro</td>
<td>F1-Macro</td>
</tr>
<tr>
<td>CutKum</td>
<td>88.96</td>
<td>96.95</td>
</tr>
<tr>
<td>DeepCut</td>
<td>89.95</td>
<td>97.99</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>94.39</td>
<td>95.72</td>
</tr>
<tr>
<td>BiLSTM-CRF (10%)</td>
<td>93.47</td>
<td>94.40</td>
</tr>
<tr>
<td>BiGRU-CRF</td>
<td><strong>94.78</strong></td>
<td><strong>96.26</strong></td>
</tr>
<tr>
<td>BiGRU-CRF (10%)</td>
<td>93.88</td>
<td>95.01</td>
</tr>
</tbody>
</table>

Existing non-published models: CutKum (RNN), DeepCut (CNN). Models trained on 10% of the training data are denoted by (10%).

Analysis

Given the high micro F1 scores, we reason that our model does considerably well on smaller training sets.

Our model has difficulty with named entities and compound words.

Input sentence:

...เท้าให้คล่องแคลวและวูงไว...

Correct segmentation:

...| เทา | ให | คล่องแคลว | และ | วูงไว | ...

Model segmentation:

...| เทาน | ให | คลอง | แคลว | และ | วูง | ไว |...

Conclusion

Using the BiLSTM-CRF and BiGRU-CRF shows that an effective Thai segmentation model can be constructed despite the size of segmented Thai data. Hyperparameter tuning and dropout can be applied to future work.

References