# BiLSTM-CRF & BiGRU-CRF for Thai Segmentation

Armando Banuelos and Nantanick Tantivasadakarn
CS224N Natural Language Processing with Deep Learning, Department of Computer Science

## 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.

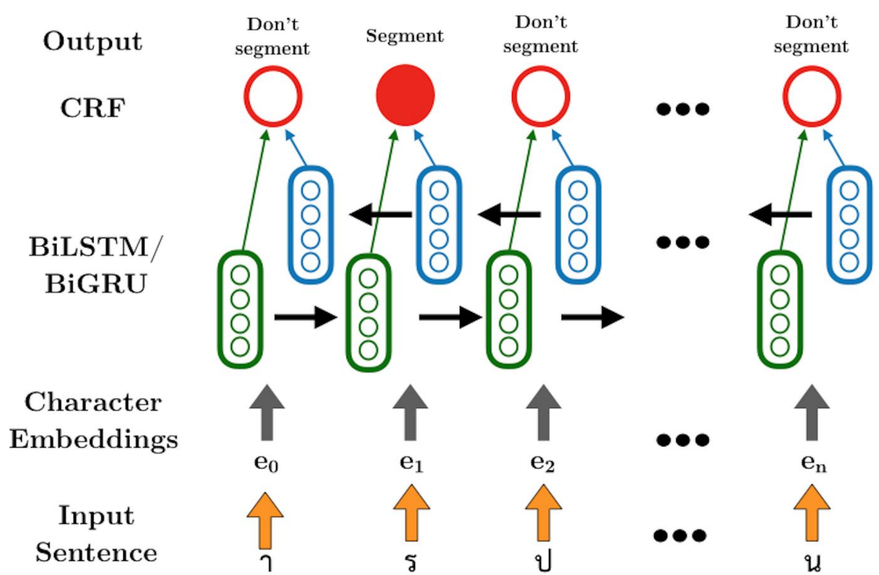
## Data

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.

## BiLSTM-CRF & BiGRU-CRF Model




## Results

| Model | Without named entities | | With named entities | |
|---|---|---|---|---|
| | F1-Micro | F1-Macro | F1-Micro | F1-Macro |
| CutKum | 88.96 | 96.95 | 88.18 | 96.27 |
| DeepCut | 89.95 | **97.99** | 89.61 | 97.80 |
| BiLSTM-CRF | 94.39 | 95.72 | 92.54 | 93.74 |
| BiLSTM-CRF (10%) | 93.47 | 94.40 | 91.11 | 92.24 |
| BiGRU-CRF | **94.78** | **96.26** | - | - |
| BiGRU-CRF (10%) | 93.88 | 95.01 | 91.36 | 92.70 |

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:**
...เท้าให้คล่องแคล่วและว่องไว...
**Correct segmentation:**
...| เท้า | ให้ | คล่องแคล่ว | และ | ว่องไว | ...
**Model segmentation:**
...| เท้า | ให้ | คล่อง | แคล่ว | และ | ว่อง | 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

1. Yao, Yushi & Huang, Zheng (2016). Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation. A. Hirose et al. (eds.): ICONIP 2016, Part IV, LNCS 9950. pp.345–353, 2016.
2. Samih, Younes & Attia, Mohammed (2017). A Neural Architecture for Dialectal Arabic Segmentation Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP). Valencia, Spain, pp.46-54, 2017.