STATISTICAL NATURAL LANGUAGE PROCESSING
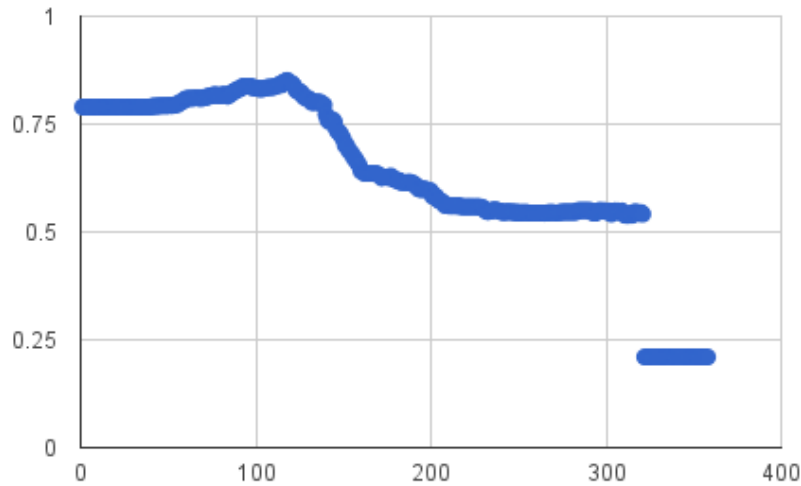
BRANDEIS UNIVERSITY

FALL 2015

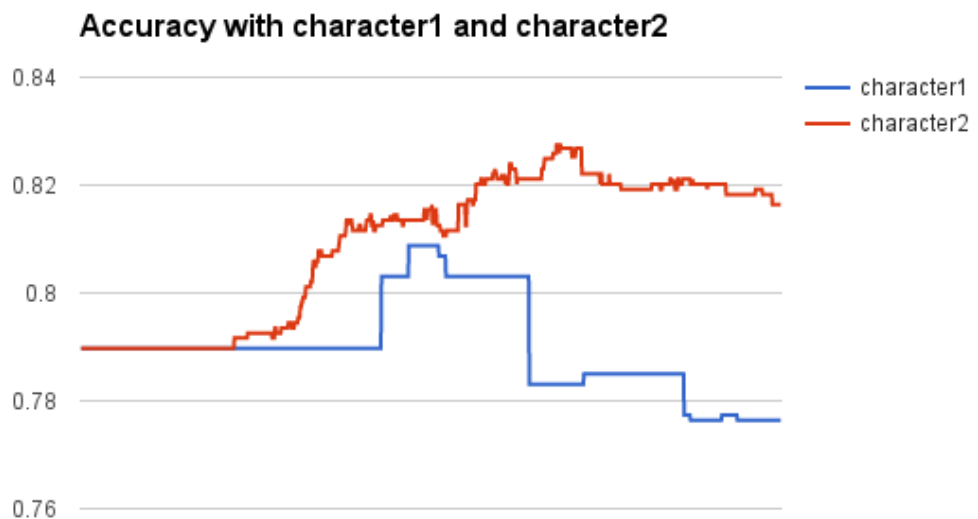# PA3: Thai Word Segmenter - Linear-Chain Conditional Random Fields

Huilin Gang

November 17, 2015

1 Conditional Random Fields

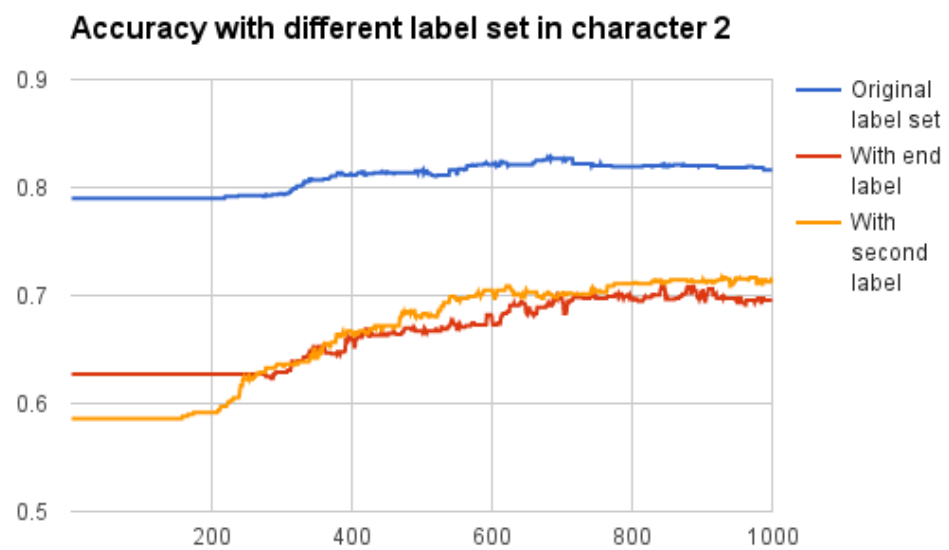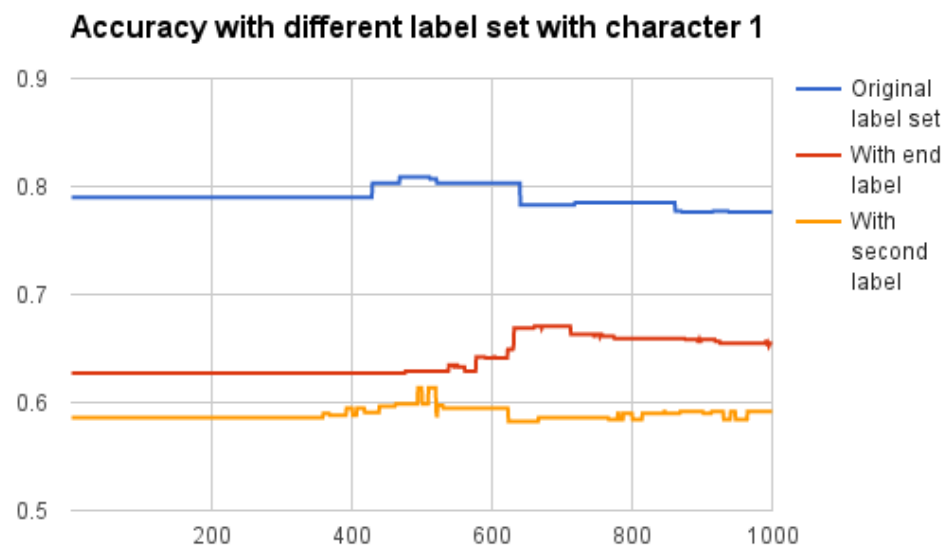The unregularised standard CRF performs overfitting and overflow during the experiment.

Therefore, I implemented base log to avoid overflow in CRFs. After implemented base logarithm, the accuracy of CRFs looks like below:

2 Experiment

Label Sets The current label set only distinguishes the beginning of the word and the non-beginning of the word. For the experiment, I tag the end of the word and also the

second place of the word. I tested both of the label set and the accuracy looks like below:

Accuracy with different label set with character 1

Accuracy with different label set in character 2

From what I have experimented, the best result came from the original label set.