Better Next Word Prediction

Shuai Wang

August 23rd, 2015

Next Word Prediction Website

- A website with better next word prediction than SwiftKey https://shuai114.shinyapps.io/JHUDSCapstone
- Accurate: 16% accuracy for the first predicted word, 10% for the second, 4% for the third, and 30% in total
- Super fast: 0.05 seconds in average for prediction on one phrase, at most 0.12 seconds in my test
- Light: Only about 14 MB
- Simple to use: Just input your phrase in the text box in the left panel and click OK. The result will be shown in the right main panel.
- Geared towards prediction for twitter and news, but also OK for other kinds

Prediction Algorithm (Basic)

- Basic n-gram model is used for predicting the next word based on the previous 1, 2, and 3 words.
- Simple backoff model is used to handle unseen n-grams, i.e., first do prediction based on the last 3 words, then on the last 2 words, then on the last one word, until three predicted words are found (Katz's back-off model is not used, because it is more time consuming and does not improve the prediction accuracy).
- Some characters are removed before training and prediction, including unrecognizable UTF-8 codes, special characters, punctuations, numbers, and some profanity words, because I don't want to predict them right now, but hyphens within words are retained.
- Contracted forms are expanded to unify the expression.

Prediction Algorithm (Innovation)

- The main problem in next word prediction is the sparsity of the n-gram patterns.
- One way to solve it is to increase the training data, but computation power is limited, and the size of the application should not be large, as it needs to be installed in mobile devices.
- Stemming is another way to deal with it, but I don't want to predict stemmed words. So stemming is performed on all the training n-grams except their last words.
- Stemming does not unify variants of irregular verbs, or the first / second / third person possessive forms and reflexive pronouns. So they are unified manually to make the frequent n-gram pattern more evident. Again, it is done on all the training n-grams except their last words.

Future Plans And Needs

- There is far more to be done to further improve the prediction in accuracy, speed and size at the same time.
- However, the current version still outperforms SwiftKey. See the website for more details.
- One of the future plans is to develop an app based on this algorithm to be used in mobile devices.
- Another future plan is to further improve the algorithm. Currently the problem is only approached primitively from a pattern recognition perspective. Other techniques, such as document classification, semantic analysis, may be incorporated to make it more accurate.
- Thus your funding is needed for the future. Imagine what it is like to be a sponsor of some company like SwiftKey!