9319 proj report

Yu Feng z5094935

May 2017

1 feature selection

According to this paper [1], the most useful feature is part of speech, type of the syllable, number of syllables in the word, stress marks of three preceding syllables, the first grapheme of the current syllable, the last grapheme of the current syllable and position of the current syllable in the word.

For testing, we can not get stress marks of three preceding syllables. I found most of words(more than 99%) in training set is NPP, so I did not select part of speech. I use remain entries as features. It's hard to get which is first and last grapheme in the syllable, so I select the consonant before and after vowel as first and last grapheme.

In short, I select type of the syllable(vowel), number of vowels in the word, the consonant before and after each vowel. I use trigram.

I have tried use suffix and prefix of the word, but the result is not ideal.

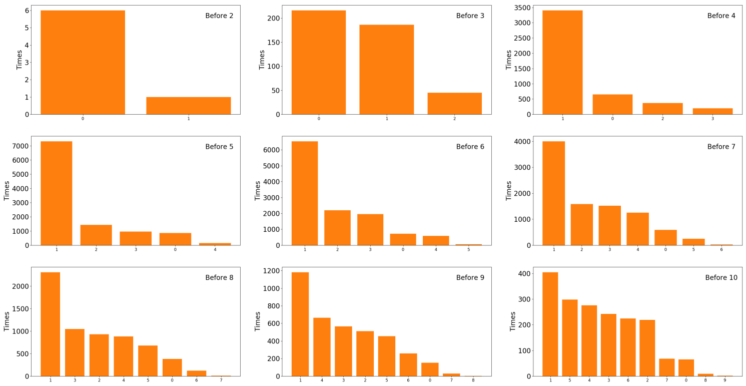

Figure 1: the number of stress vowel in different position for different total length

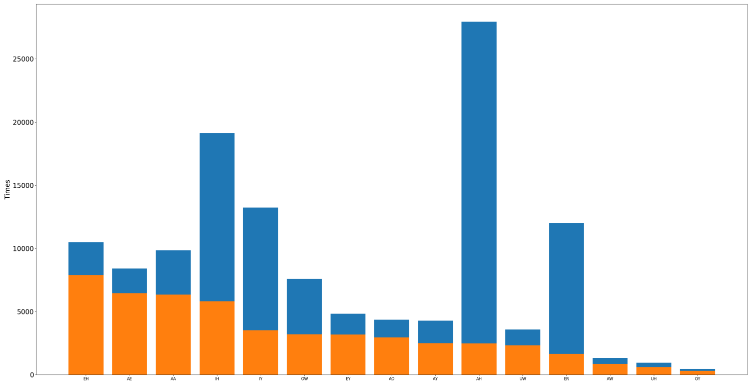

Figure 2: number of stress vowel and total number of vowel

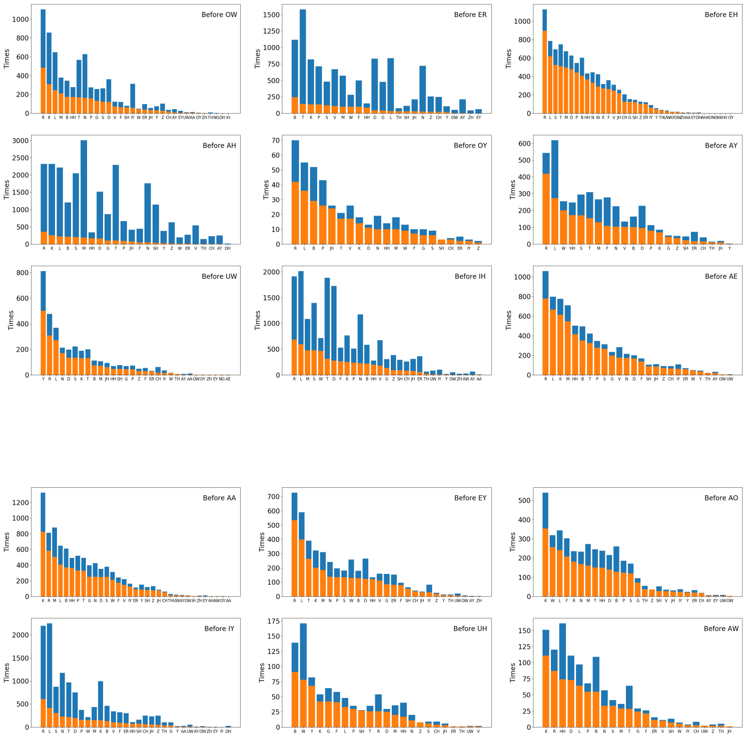

Figure 3: number of consonant before stress vowel

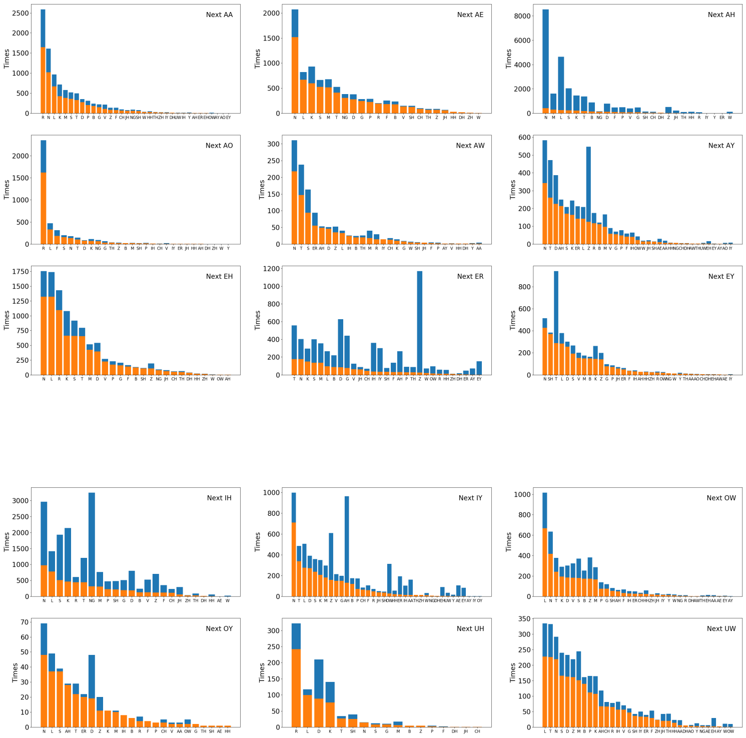

Figure 4: number of consonant after stress vowel

2 experiment and improvement

At beginning, I only use position and type of vowel as features. I tried mutil-bayesian and decision tree. Decision tress performed better which accuracy is 0.68. so I decide use decision tree classifier.

After visualized data and read papers, I found the consonant before and after vowel is also important. So I add these two feature, the accuracy is 0.86.

But f1 score is only 0.68. After checking out the prediction and real value, I found that the classifier performed bad in 4 vowel word. Because the data is unbalanced – the number of words which stressed in the fourth vowel is low. So

I separate words according to the length(number of vowels). Each length has a decision tree classifier. The f1 score become 0.70 by doing this.

then I remap the phonetic to integer ordered by the frequency, oversampling the unbalanced data, adjust the decision tree depth to avoid overfitting, set minority data larger weight. After doing these thing, I got 0.75 f1 score.

TODO: I should add more features to the data.

References

[1] Stephen James. Learning english stress rules-using a machine learning approach, 1999.