The baseline was computed by guessing the most likeliy result – which is having no accent at all (y=0). And if I geussed it all times the baseline is around 0.699. Any feature that contributes to increasing this accuracy (score) will be taken to the final train.

The following table is the test score ( avaraged of 2 loops of 5-fold of a random subset of 10k)

|  |  |
| --- | --- |
| Without the Feature | Score |
| no\_phrases | 0.739 |
| no word | 0.772 |
| no\_func | 0.773 |
| no\_dial | 0.773 |
| no\_nuc | 0.775 |
| no\_sylls | 0.776 |
| no\_neg | 0.78 |
| no\_kontrast | 0.783 |
| no\_c\_tag | 0.784 |
| no\_tag | 0.785 |

Most of the trials here succesfully overcome the missing feature however, phrases is suspected to crucial due to the relatively high decreased score. The second most contributing but with not a small gap was word.

The following table is the isolated feature test score in a decending order.

|  |  |
| --- | --- |
| With the Feature Only | Score |
| word | 0.738 |
| c\_tag | 0.727 |
| kontrast | 0.723 |
| tag | 0.720 |
| all\_sylls | 0.719 |
| function | 0.703 |
| phrases | 0.703 |
| negation | 0.702 |
| nuc\_info | 0.702 |
| dialect | 0.696 |

The top 5 (without tag) will be used to the bottom up feature selection.

|  |  |
| --- | --- |
| With 2 Feature Only | Score |
| phrases+words | 0.775 |
| Phrases+c\_tags | 0.766 |
| phrases+sylls | 0.743 |
| c\_tags+words | 0.736 |
| kontrast+sylls | 0.735 |
| kontrast+words | 0.731 |
| c\_tags+sylls | 0.729 |
| sylls+words | 0.728 |
| kontrast+phrases | 0.723 |
|  |

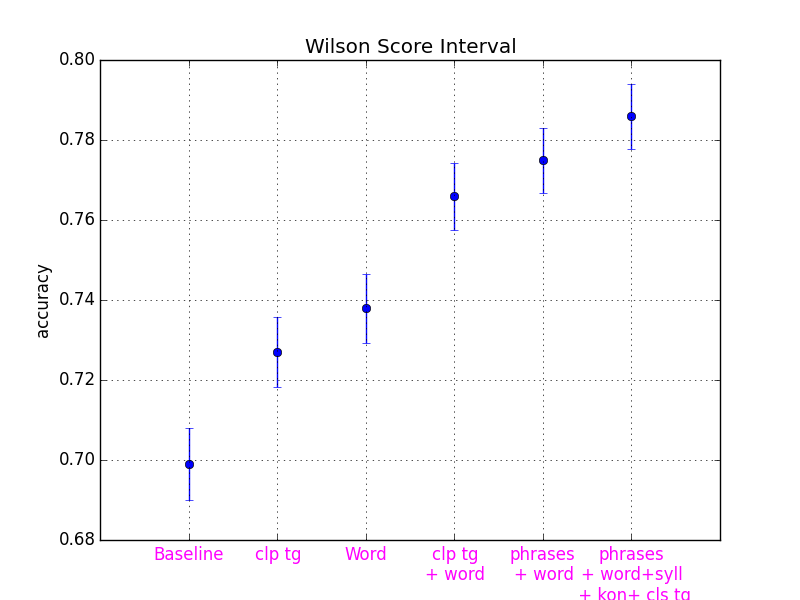
Though phrases isolated are meaningless (table 2) when the were with word or tags they boosted the scores and also improved sylls.

Kontrast improves sylls.

|  |  |
| --- | --- |
| With 3 Feature and more | Score |
| phrases+sylls+words | 0.783 |
|  |  |
| phrases+sylls+words+kontrast | 0.772 |
|  |  |
| phrases+sylls+words+c\_tags | 0.78 |
|  |  |
| phrases+sylls+words+c\_tags+kontrast | 0.786 |
|  |

The final features that were found to be the most contributing are: words, syll, phrases, c\_tags and kontrast

The Wilson confidence interval score plot with selected feature combinations is:



\*The code is in the repository. It seems that using these features demostrated that we were able to improve our model in a way that is higher than just the baseline.