COSI 114 Final Project Report

I have done two experiments about the superchunk task. In both experiments, I use the same feature set to train the model, which is:
For each word w_i:

- The word w_{i-2} ,w_{i-1},w_i , w_{i+1}, w_{i+2}
- The POS-tag of w_{i-2} ,w_{i-1},w_i , w_{i+1}, w_{i+2}
- The Chunk tag of w_{i-2} ,w_{i-1},w_i , w_{i+1}, w_{i+2}
- Whether the word is First word of the sentence.

And I test the performance of the feature set on two classifiers Naive Bayes and CRF++. The results is:

The performance of Naive Bayes:

	precision	recall	F-measure
B-SNP	0.3547	0.8054	0.4925
I-SNP	0.6715	0.6795	0.6755
O	0.8787	0.7698	0.8206

The performance of CRF:

	precision	recall	F-measure
B-SNP	0.8134	0.5891	0.6833
I-SNP	0.7895	0.6261	0.6984
O	0.8436	0.9319	0.8855

Since what we concern is only the B-SNP and I-SNP tag, I only compare the performance on these two labels. We can see that by using the sequential model, which is more reasonable for this task, the F-measure of B and I tag has been improved, especially for the tag B-SNP. It proves that the simple assumption used by Naive Bayes is not so reasonable for the super chunk identification task.