Assignment-4
Submitted by – Harshita Jhavar
2566267, s8hajhav@stud.uni-saarland.de

In the submission folder, one can find the output folder for each of the tests sets. Also, there are separate files of 'Labelled' and 'Unlabelled' performance measures as recorded from EVALB.

Berkley Parser -1.7

Test Set	Precision		Recall		F1-Score	
	Unlabelled	Labelled	Unlabelled	Labelled	Unlabelled	Labelled
1.plain (1346 sentences)	91.81%	90.63%	90.74%	89.58%	91.27%	90.10%
2.plain (1300 sentences)	87.63%	85.89%	87.53%	85.78%	87.58%	85.83%
3.plain (869 sentences)	89.53%	87.13%	89.20%	86.80%	89.37%	86.97%
Overall test set	89.67%	88.02%	89.14%	87.51%	89.40%	87.76%

Stanford Parser – 3.7.0

Test Set	Precision		Recall		F1-Score	
	Unlabelled	Labelled	Unlabelled	Labelled	Unlabelled	Labelled
1.plain (1346 sentences)	88.11%	81.40%	86.58%	79.99%	87.34%	80.69%
2.plain (1300 sentences)	84.28%	77.44%	84.14%	77.30%	84.21%	77.37%
3.plain (869 sentences)	86.62%	77.09%	86.82%	77.28%	86.72%	77.19%
Overall test set (3515 sentences)	86.25%	78.95%	85.62%	78.38%	85.93%	78.66%

Significance Test – Paired t-test

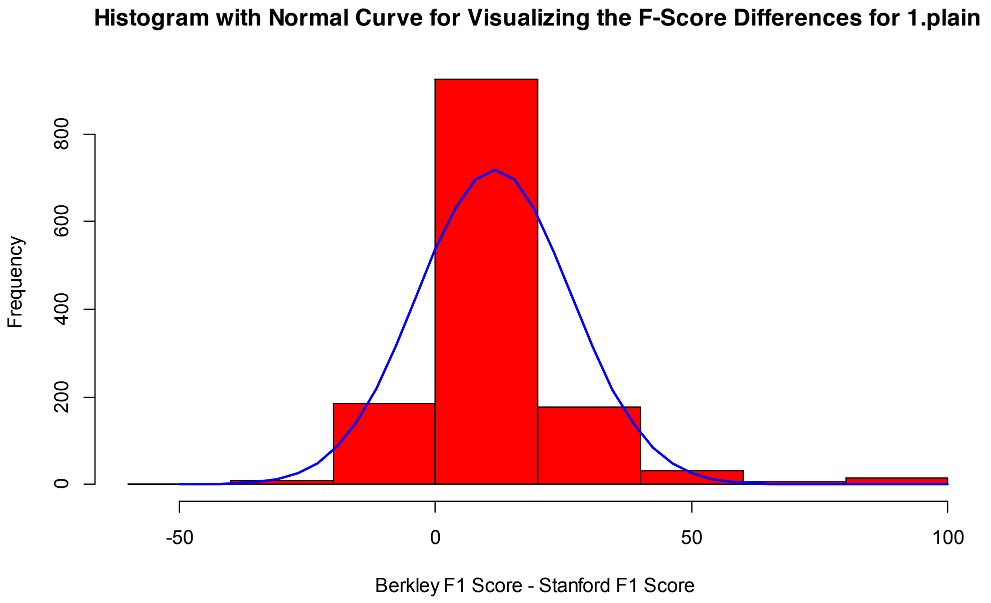
I used Paired t-test as the F-Values were obtained after parsing on the same data. These are two different observations which are subjected to the same conditions. Here, we are trying to confirm if the difference between the F-Scores is significant or not. I have used R for this purpose whose script and the values recorded can be found in the 'SignificanceTestFileInR' folder.

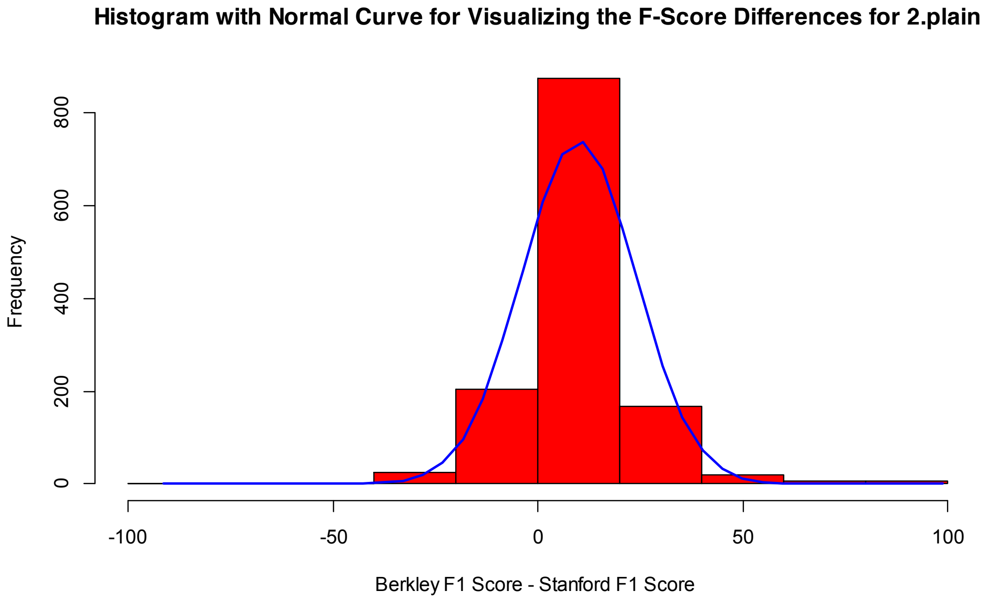
The observations recorded are:

Paired t-test Results:

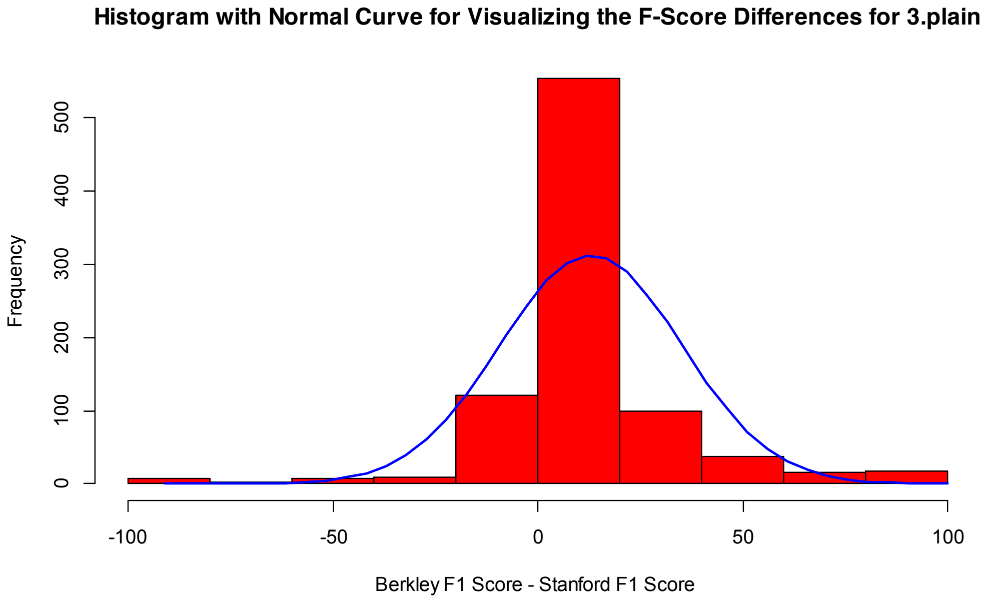
Test Set	Observed p-value	Mean of the Difference of their F-Scores	95 percent confidence interval	Result if p < 0.05 is significant	Result if p < 0.01 is significant
1.plain (1346 sentences)	p-value < 2.2e-16	11.46777	10.66842 – 12.26712	Significant	Significant
2.plain (1300 sentences)	p-value < 2.2e-16	9.903206	9.140602 - 10.665811	Significant	Significant
3.plain (869 sentences)	p-value < 2.2e-16	12.9344	11.45688 - 14.41193	Significant	Significant
Overall test set (3515 sentences)	p-value < 2.2e-16	11.25172	10.69717 - 11.80626	Significant	Significant

Below are the histogram with the normal distribution for the difference between the F scores recorded for the different test data. The plot to the left of 0 tells the value with which the Stanford Parser F Score was more. The plot to the right of 0 tells the value with which the Berkley Parser F Score was more.

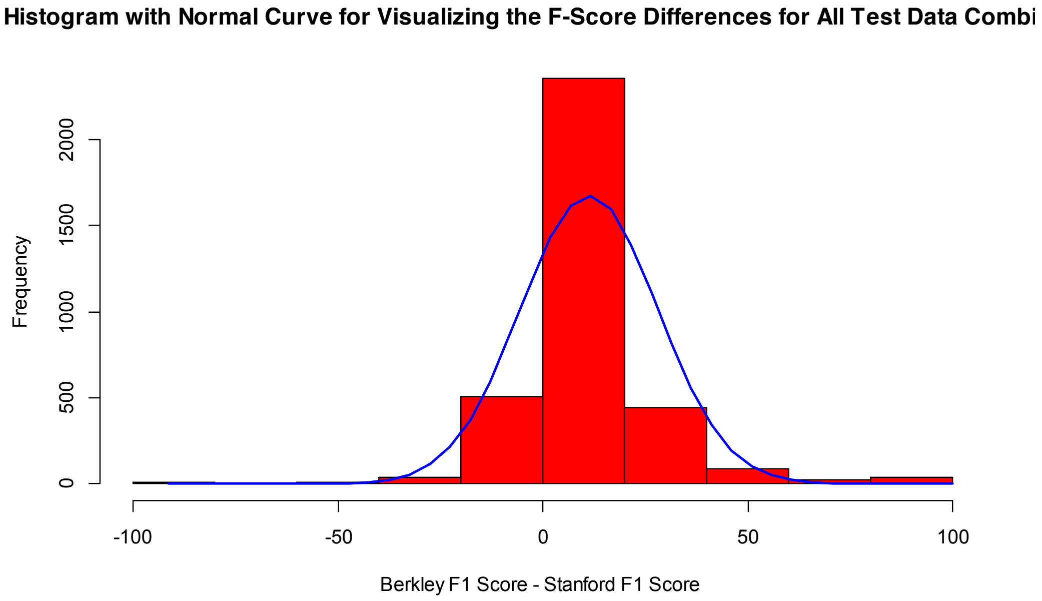




Histogram with Normal Curve for Visualizing the F-Score Differences for 2.plain



Histogram with Normal Curve for Visualizing the F-Score Differences for 3.plain



Histogram with Normal Curve for Visualizing the F-Score Differences for All Test Data Combi

Data Sources Speculation

From google, I found:

1.plain – Source is Wall Street journal 10/16/89

2.plain – Brown corpus http://www.csi.uottawa.ca/tanka/Brown/original/br-g01

3.plain – Novel - Stranger in a Strange Land