COMP6490 Document Analysis IE Assignment

Q2 Effect of the training size on the test set

1/4 Training sentences classifier on Test A

```
processed 46167 tokens with 3076 phrases; found: 2929 phrases; correct: 1917.
accuracy:  95.52%; precision:  65.45%; recall:  62.32%; FB1:  63.85
              LOC: precision:  71.11%; recall:  65.62%; FB1:  68.26  886
             MISC: precision:  43.52%; recall:  18.01%; FB1:  25.47  108
              ORG: precision:  60.42%; recall:  69.12%; FB1:  64.48  1397
              PER: precision:  73.61%; recall:  62.46%; FB1:  67.58  538
```

2/4 Training sentences classifier on Test A

```
processed 46167 tokens with 3076 phrases; found: 2915 phrases; correct: 2033.
accuracy:  96.05%; precision:  69.74%; recall:  66.09%; FB1:  67.87
              LOC: precision:  72.51%; recall:  71.15%; FB1:  71.82  942
             MISC: precision:  51.54%; recall:  25.67%; FB1:  34.27  130
              ORG: precision:  67.28%; recall:  69.21%; FB1:  68.23  1256
              PER: precision:  74.62%; recall:  69.09%; FB1:  71.74  587
```

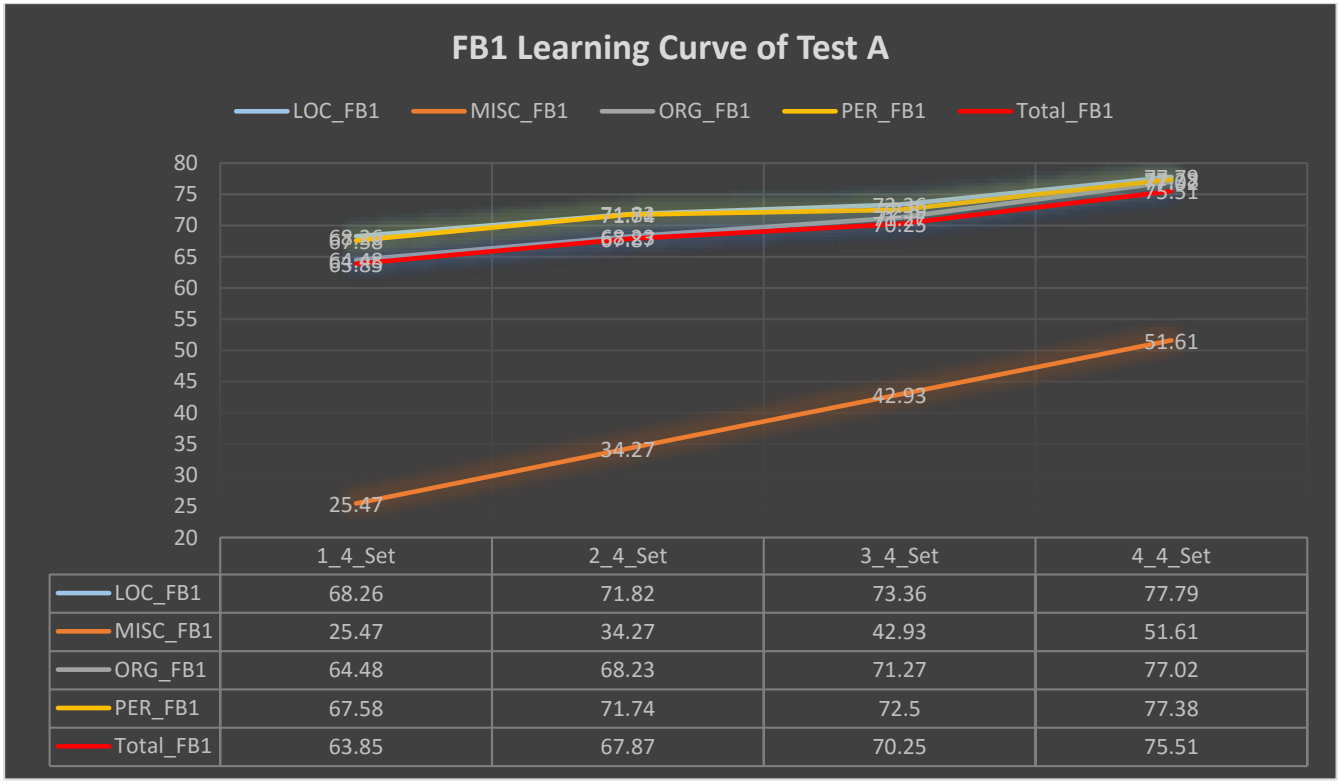
3/4 Training sentences classifier on Test A

```
processed 46167 tokens with 3076 phrases; found: 2923 phrases; correct: 2107.
accuracy:  96.34%; precision:  72.08%; recall:  68.50%; FB1:  70.25
              LOC: precision:  74.02%; recall:  72.71%; FB1:  73.36  943
             MISC: precision:  59.06%; recall:  33.72%; FB1:  42.93  149
              ORG: precision:  70.64%; recall:  71.91%; FB1:  71.27  1243
              PER: precision:  75.34%; recall:  69.87%; FB1:  72.50  588
```

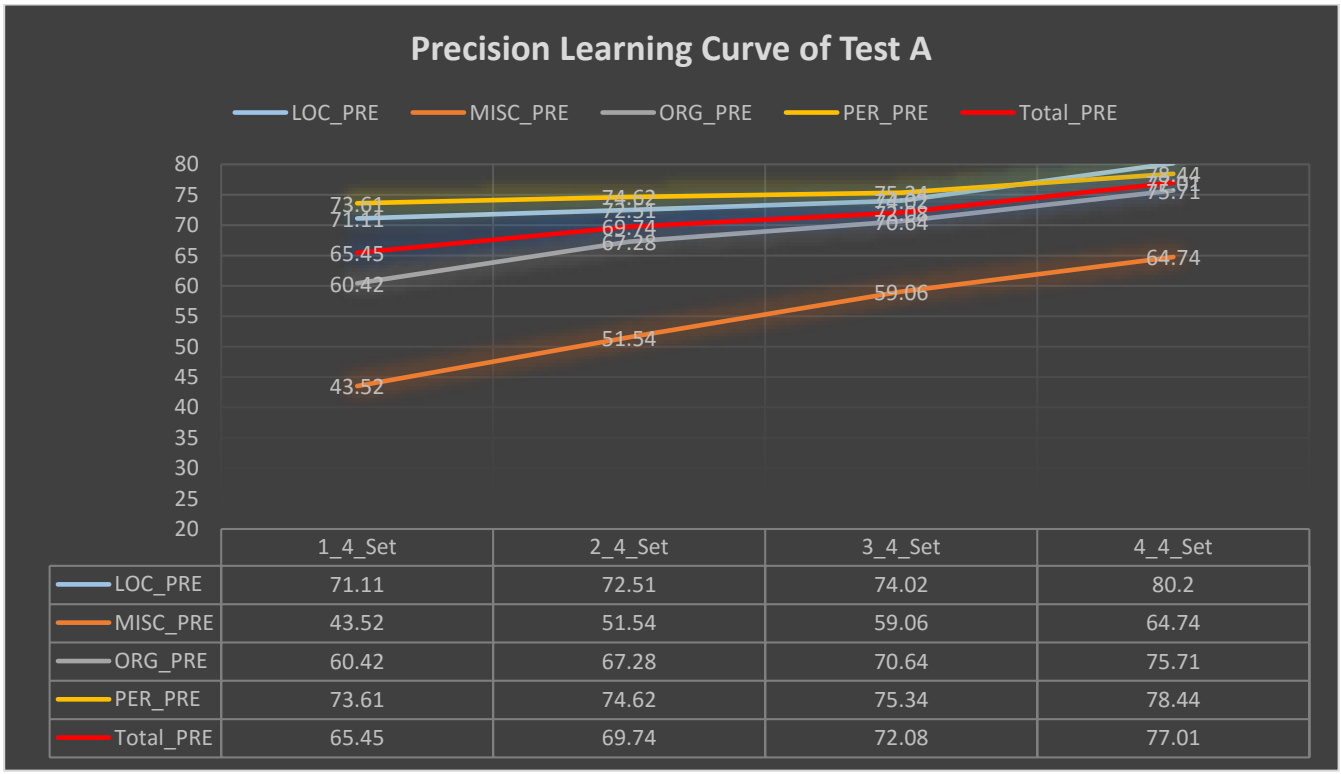
4/4 Training sentences classifier on Test A

```
processed 46167 tokens with 3076 phrases; found: 2958 phrases; correct: 2278.
accuracy:  97.06%; precision:  77.01%; recall:  74.06%; FB1:  75.51
              LOC: precision:  80.20%; recall:  75.52%; FB1:  77.79  904
             MISC: precision:  64.74%; recall:  42.91%; FB1:  51.61  173
              ORG: precision:  75.71%; recall:  78.38%; FB1:  77.02  1264
              PER: precision:  78.44%; recall:  76.34%; FB1:  77.38  617
```

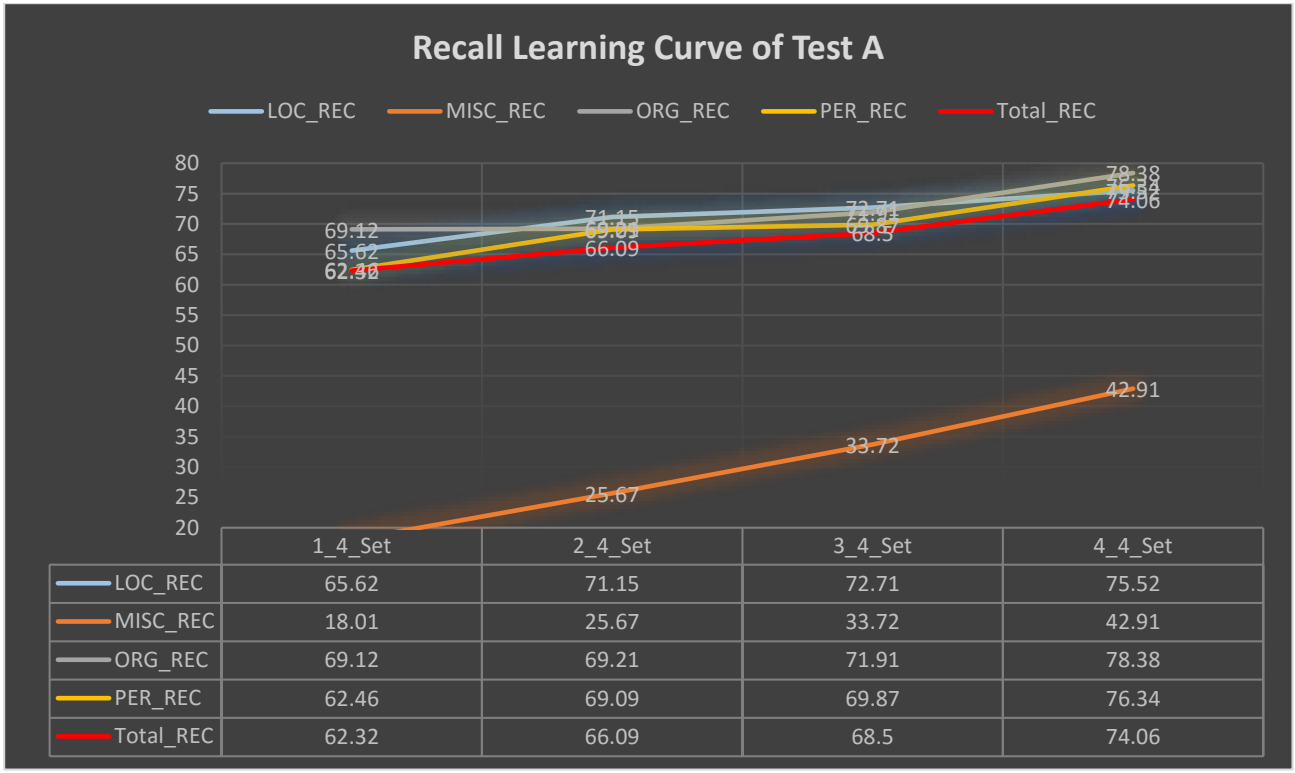
Learning Curve of Test A:



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_FB1	68.26	71.82	73.36	77.79
MISC_FB1	25.47	34.27	42.93	51.61
ORG_FB1	64.48	68.23	71.27	77.02
PER_FB1	67.58	71.74	72.5	77.38
Total_FB1	63.85	67.87	70.25	75.51



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_PRE	71.11	72.51	74.02	80.2
MISC_PRE	43.52	51.54	59.06	64.74
ORG_PRE	60.42	67.28	70.64	75.71
PER_PRE	73.61	74.62	75.34	78.44
Total_PRE	65.45	69.74	72.08	77.01



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_REC	65.62	71.15	72.71	75.52
MISC_REC	18.01	25.67	33.72	42.91
ORG_REC	69.12	69.21	71.91	78.38
PER_REC	62.46	69.09	69.87	76.34
Total_REC	62.32	66.09	68.5	74.06

1/4 Training sentences classifier on Test B

```
processed 47696 tokens with 3877 phrases; found: 3618 phrases; correct: 2170.
accuracy:  93.44%; precision:  59.98%; recall:  55.97%; FB1:  57.91
              LOC: precision:  53.02%; recall:  65.98%; FB1:  58.79  1094
             MISC: precision:  37.14%; recall:  13.68%; FB1:  20.00  140
              ORG: precision:  60.19%; recall:  62.82%; FB1:  61.48  1575
              PER: precision:  72.93%; recall:  53.20%; FB1:  61.52  809
```

2/4 Training sentences classifier on Test B

```
processed 47696 tokens with 3877 phrases; found: 3605 phrases; correct: 2288.
accuracy:  94.12%; precision:  63.47%; recall:  59.01%; FB1:  61.16
              LOC: precision:  53.31%; recall:  71.44%; FB1:  61.06  1178
             MISC: precision:  44.56%; recall:  22.63%; FB1:  30.02  193
              ORG: precision:  67.95%; recall:  60.97%; FB1:  64.27  1354
              PER: precision:  74.32%; recall:  58.97%; FB1:  65.76  880
```

3/4 Training sentences classifier on Test B

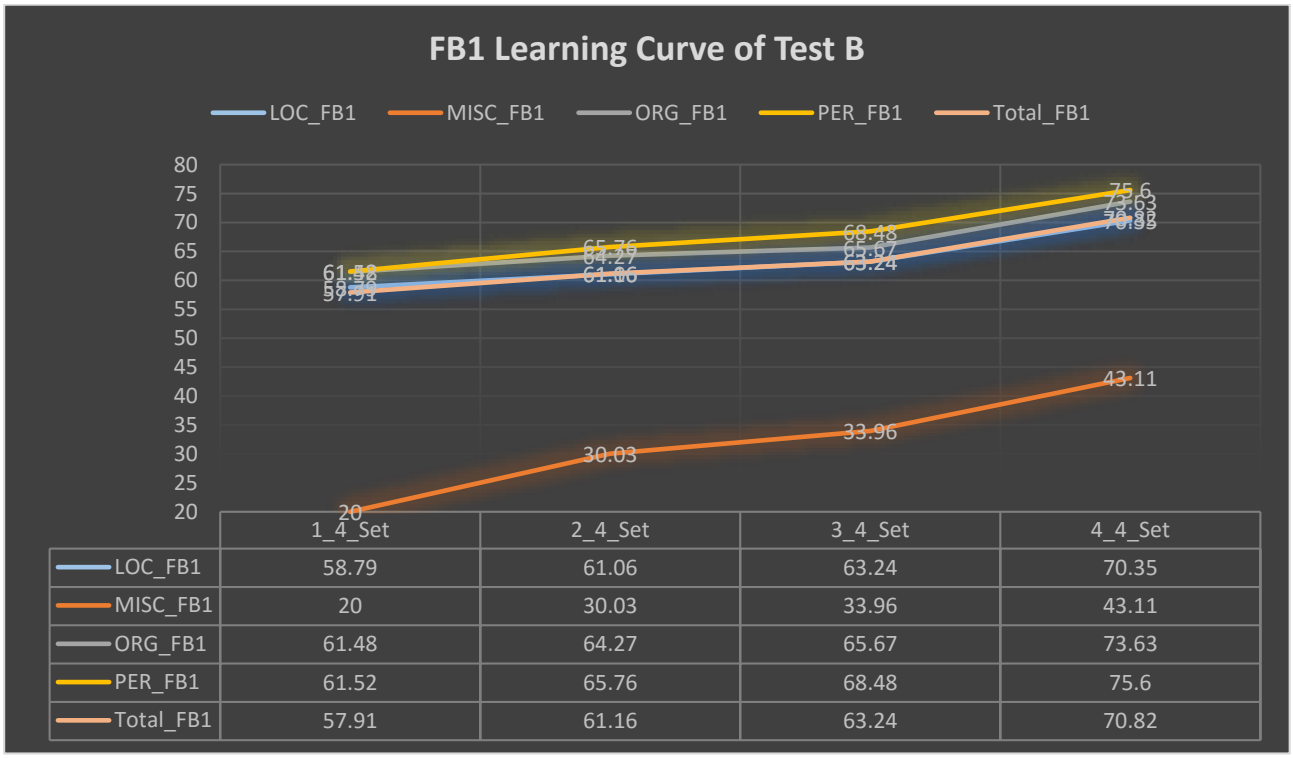
```
processed 47696 tokens with 3877 phrases; found: 3587 phrases; correct: 2360.
accuracy:  94.57%; precision:  65.79%; recall:  60.87%; FB1:  63.24
              LOC: precision:  54.74%; recall:  74.86%; FB1:  63.24  1202
             MISC: precision:  47.85%; recall:  26.32%; FB1:  33.96  209
              ORG: precision:  70.89%; recall:  61.17%; FB1:  65.67  1302
              PER: precision:  77.69%; recall:  61.23%; FB1:  68.48  874
```

4/4 Training sentences classifier on Test B

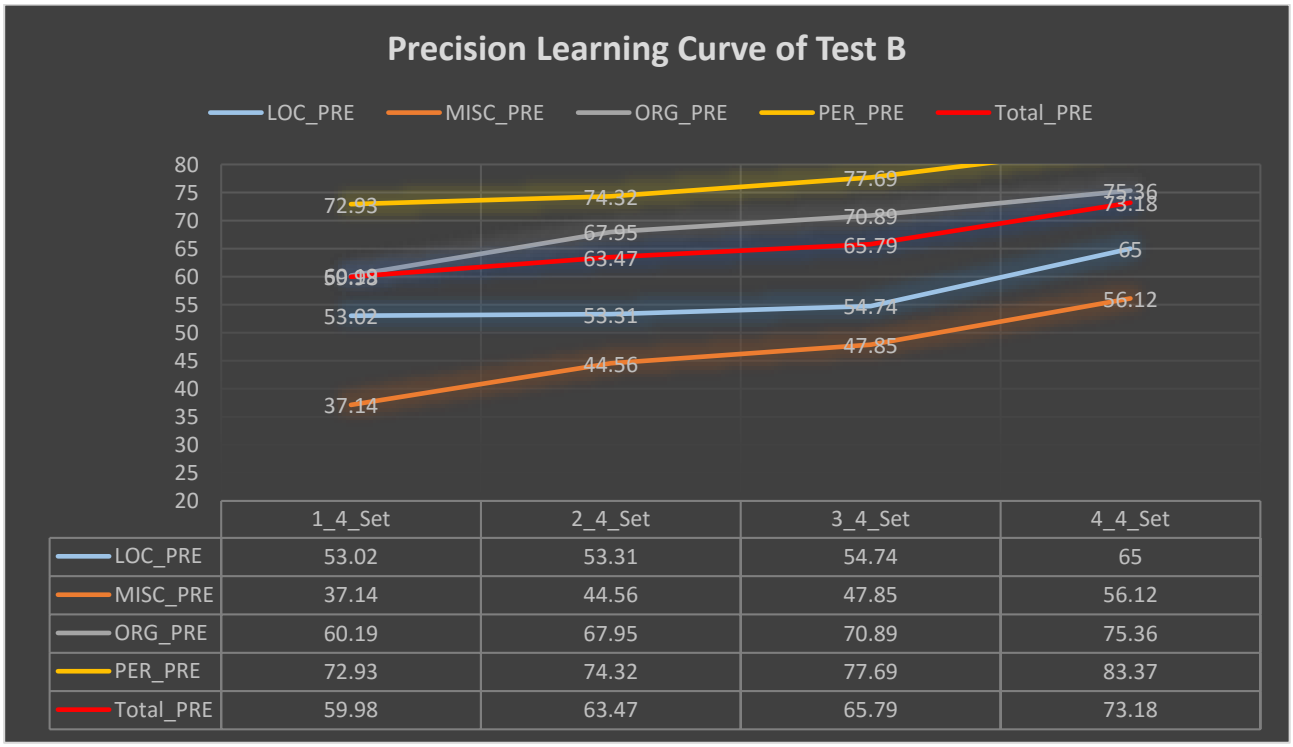
```
processed 47696 tokens with 3877 phrases; found: 3635 phrases; correct: 2660.
accuracy:  95.51%; precision:  73.18%; recall:  68.61%; FB1:  70.82
              LOC: precision:  65.00%; recall:  76.68%; FB1:  70.35  1037
             MISC: precision:  56.12%; recall:  35.00%; FB1:  43.11  237
              ORG: precision:  75.36%; recall:  71.97%; FB1:  73.63  1441
              PER: precision:  83.37%; recall:  69.16%; FB1:  75.60  920

```

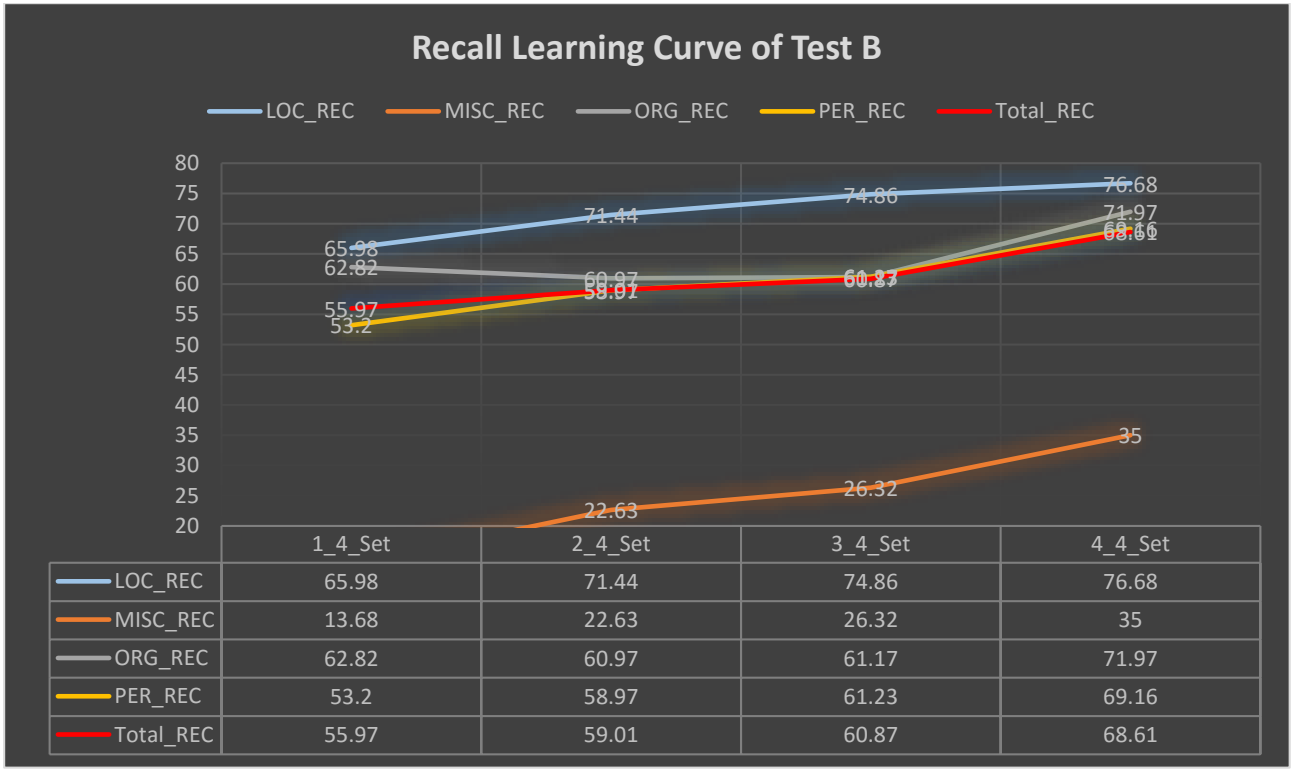
Learning Curve of Test B:



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_FB1	58.79	61.06	63.24	70.35
MISC_FB1	20	30.03	33.96	43.11
ORG_FB1	61.48	64.27	65.67	73.63
PER_FB1	61.52	65.76	68.48	75.6
Total_FB1	57.91	61.16	63.24	70.82



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_PRE	53.02	53.31	54.74	65
MISC_PRE	37.14	44.56	47.85	56.12
ORG_PRE	60.19	67.95	70.89	75.36
PER_PRE	72.93	74.32	77.69	83.37
Total_PRE	59.98	63.47	65.79	73.18



	1_4_Set	2_4_Set	3_4_Set	4_4_Set
LOC_REC	65.98	71.44	74.86	76.68
MISC_REC	13.68	22.63	26.32	35
ORG_REC	62.82	60.97	61.17	71.97
PER_REC	53.2	58.97	61.23	69.16
Total_REC	55.97	59.01	60.87	68.61

Conclusion:

According to the learning curves on test set A and test set B, we can draw the conclusion that the NER performance will improve with the increase of training size.

Q3 Re-substitution performance

Re-substitution is the simplest re-sampling technique to implement. Re-substitution involves using the entire dataset and the test set. [1] The performance of train data using the model trained from itself can be called re-substitution performance.
Re-substitution is usually treated as a poor estimate of generalisation performance because all cases used in the testing have contributed to the data-mining. If the data set is quite representative, re-substitution provides good estimates. The re-substitution method is the simplest re-sampling method to implement and is often used to provide early rough estimates of performance. [1]
In this question, I set c value to 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 10.0. And get the following result of the re-substitution performance on train set.
From the result data and performance curve, we can learn that:

1. The re-substitution performance is always pretty high compared with normal performance on other test set.
2. The re-substitution performance will improve with the increase of c value. However, the improvement is much more obvious when c value increase from less 0.1 to 1.0.
   From 1.0, the contribution of increase of c value becomes flat, which because the performance has reached a pretty high state.

Re-substitution Performance on c value 0.1:

```
processed 237201 tokens with 16431 phrases; found: 15785 phrases; correct: 14017.
accuracy:  98.53%; precision:  88.80%; recall:  85.31%; FB1:  87.02
              LOC: precision:  87.63%; recall:  86.74%; FB1:  87.18  4388
             MISC: precision:  87.21%; recall:  60.87%; FB1:  71.70  1204
              ORG: precision:  87.66%; recall:  88.96%; FB1:  88.31  6474
              PER: precision:  92.69%; recall:  88.52%; FB1:  90.56  3719

```

Re-substitution Performance on c value 0.2:

```
processed 237201 tokens with 16431 phrases; found: 16030 phrases; correct: 15151.
accuracy:  99.27%; precision:  94.52%; recall:  92.21%; FB1:  93.35
              LOC: precision:  93.77%; recall:  92.98%; FB1:  93.37  4396
             MISC: precision:  94.68%; recall:  78.43%; FB1:  85.80  1429
              ORG: precision:  93.57%; recall:  94.18%; FB1:  93.88  6421
              PER: precision:  96.93%; recall:  94.20%; FB1:  95.55  3784
```

Re-substitution Performance on c value 0.3:

```
processed 237201 tokens with 16431 phrases; found: 16170 phrases; correct: 15722.
accuracy:  99.62%; precision:  97.23%; recall:  95.68%; FB1:  96.45
              LOC: precision:  96.32%; recall:  96.28%; FB1:  96.30  4431
             MISC: precision:  97.76%; recall:  88.52%; FB1:  92.91  1562
              ORG: precision:  96.75%; recall:  96.47%; FB1:  96.61  6361
              PER: precision:  98.87%; recall:  96.89%; FB1:  97.87  3816
```

Re-substitution Performance on c value 0.5:

```
processed 237201 tokens with 16431 phrases; found: 16295 phrases; correct: 16079.
accuracy:  99.81%; precision:  98.67%; recall:  97.86%; FB1:  98.26
              LOC: precision:  98.43%; recall:  97.81%; FB1:  98.12  4405
             MISC: precision:  99.15%; recall:  94.67%; FB1:  96.86  1647
              ORG: precision:  98.21%; recall:  98.17%; FB1:  98.19  6376
              PER: precision:  99.51%; recall:  98.82%; FB1:  99.16  3867
```

Re-substitution Performance on c value 1.0:

```
processed 237201 tokens with 16431 phrases; found: 16411 phrases; correct: 16292.
accuracy:  99.91%; precision:  99.27%; recall:  99.15%; FB1:  99.21
              LOC: precision:  99.10%; recall:  99.01%; FB1:  99.05  4429
             MISC: precision:  99.65%; recall:  97.74%; FB1:  98.68  1692
              ORG: precision:  98.97%; recall:  99.29%; FB1:  99.13  6400
              PER: precision:  99.82%; recall:  99.72%; FB1:  99.77  3890
```

Re-substitution Performance on c value 1.5:

```
processed 237201 tokens with 16431 phrases; found: 16434 phrases; correct: 16330.
accuracy:  99.93%; precision:  99.37%; recall:  99.39%; FB1:  99.38
              LOC: precision:  99.14%; recall:  99.21%; FB1:  99.18  4436
             MISC: precision:  99.82%; recall:  98.43%; FB1:  99.12  1701
              ORG: precision:  99.11%; recall:  99.47%; FB1:  99.29  6402
              PER: precision:  99.85%; recall:  99.87%; FB1:  99.86  3895
```

Re-substitution Performance on c value 2.0:

```
processed 237201 tokens with 16431 phrases; found: 16437 phrases; correct: 16337.
accuracy:  99.93%; precision:  99.39%; recall:  99.43%; FB1:  99.41
              LOC: precision:  99.10%; recall:  99.32%; FB1:  99.21  4443
             MISC: precision:  99.82%; recall:  98.49%; FB1:  99.15  1702
              ORG: precision:  99.20%; recall:  99.45%; FB1:  99.33  6395
              PER: precision:  99.85%; recall:  99.92%; FB1:  99.88  3897
```

Re-substitution Performance on c value 2.5:

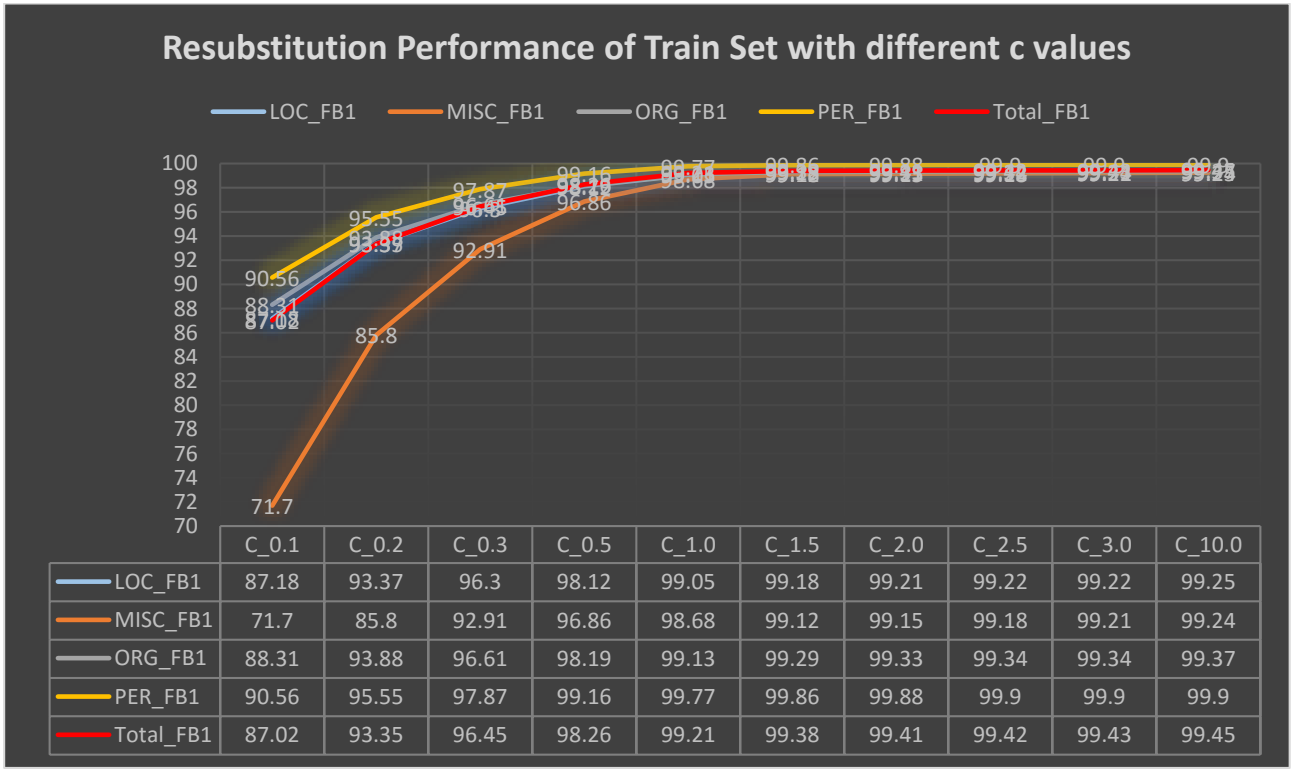
```
processed 237201 tokens with 16431 phrases; found: 16432 phrases; correct: 16337.
accuracy:  99.93%; precision:  99.42%; recall:  99.43%; FB1:  99.42
              LOC: precision:  99.32%; recall:  99.12%; FB1:  99.22  4424
             MISC: precision:  99.65%; recall:  98.72%; FB1:  99.18  1709
              ORG: precision:  99.14%; recall:  99.55%; FB1:  99.34  6405
              PER: precision:  99.90%; recall:  99.90%; FB1:  99.90  3894
```

Re-substitution Performance on c value 3.0:

```
processed 237201 tokens with 16431 phrases; found: 16435 phrases; correct: 16339.
accuracy:  99.93%; precision:  99.42%; recall:  99.44%; FB1:  99.43
              LOC: precision:  99.26%; recall:  99.19%; FB1:  99.22  4430
             MISC: precision:  99.77%; recall:  98.67%; FB1:  99.21  1706
              ORG: precision:  99.17%; recall:  99.51%; FB1:  99.34  6401
              PER: precision:  99.85%; recall:  99.95%; FB1:  99.90  3898
```

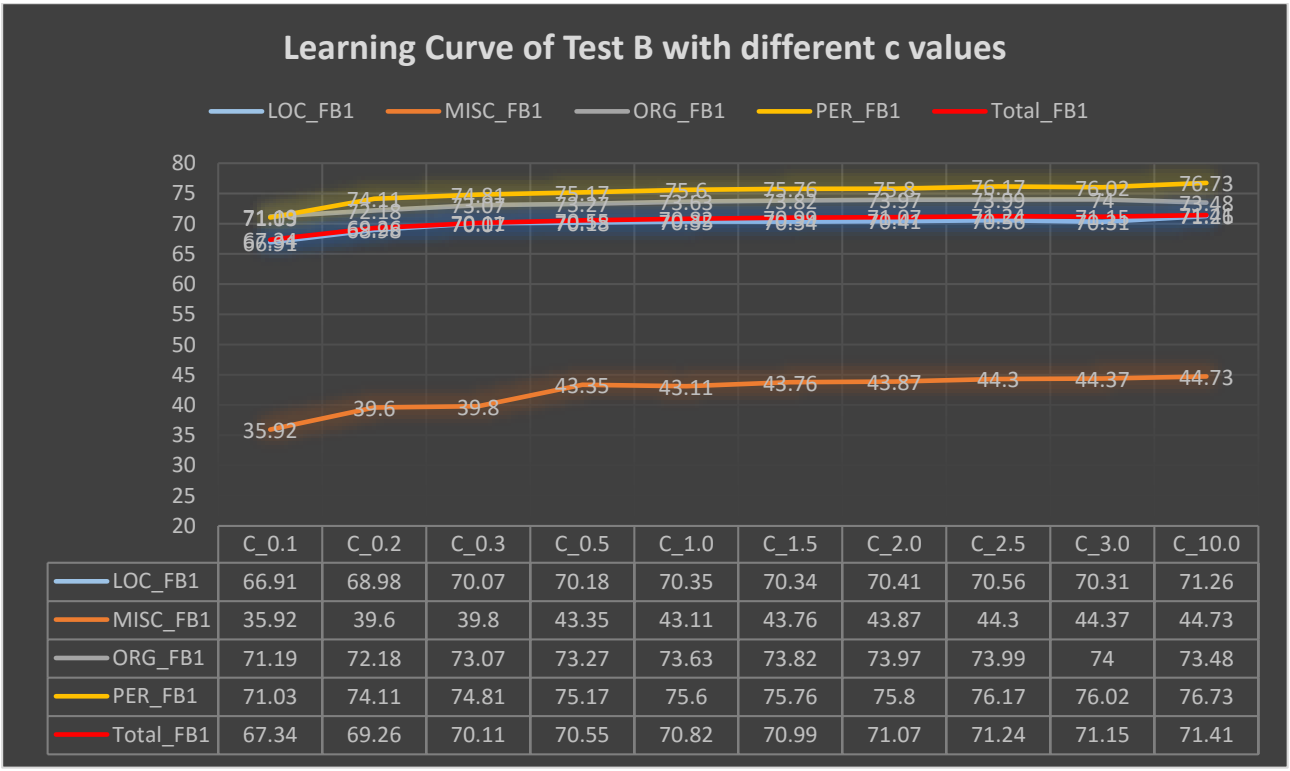
Re-substitution Performance on c value 10.0:

```
processed 237201 tokens with 16431 phrases; found: 16439 phrases; correct: 16344.
accuracy:  99.93%; precision:  99.42%; recall:  99.47%; FB1:  99.45
              LOC: precision:  99.14%; recall:  99.35%; FB1:  99.25  4442
             MISC: precision:  99.48%; recall:  99.01%; FB1:  99.24  1717
              ORG: precision:  99.31%; recall:  99.42%; FB1:  99.37  6386
              PER: precision:  99.90%; recall:  99.90%; FB1:  99.90  3894
```



	C_0.1	C_0.2	C_0.3	C_0.5	C_1.0	C_1.5	C_2.0	C_2.5	C_3.0	C_10.0
LOC_FB1	87.18	93.37	96.3	98.12	99.05	99.18	99.21	99.22	99.22	99.25
MISC_FB1	71.7	85.8	92.91	96.86	98.68	99.12	99.15	99.18	99.21	99.24
ORG_FB1	88.31	93.88	96.61	98.19	99.13	99.29	99.33	99.34	99.34	99.37
PER_FB1	90.56	95.55	97.87	99.16	99.77	99.86	99.88	99.9	99.9	99.9
Total_FB1	87.02	93.35	96.45	98.26	99.21	99.38	99.41	99.42	99.43	99.45

From the bellowing learn curve, we can see that the improvement on re-substitution performance have some positive effect on general performance. But not obvious enough.



	C_0.1	C_0.2	C_0.3	C_0.5	C_1.0	C_1.5	C_2.0	C_2.5	C_3.0	C_10.0
LOC_FB1	66.91	68.98	70.07	70.18	70.35	70.34	70.41	70.56	70.31	71.26
MISC_FB1	35.92	39.6	39.8	43.35	43.11	43.76	43.87	44.3	44.37	44.73
ORG_FB1	71.19	72.18	73.07	73.27	73.63	73.82	73.97	73.99	74	73.48
PER_FB1	71.03	74.11	74.81	75.17	75.6	75.76	75.8	76.17	76.02	76.73
Total_FB1	67.34	69.26	70.11	70.55	70.82	70.99	71.07	71.24	71.15	71.41

Q4 HMM for POS tagging

1. HMM is a generative probabilistic model, in which a sequence of observable X variables is generated by a sequence of internal hidden states Z. [2]

   In this question, the observations can be the single word in the Finnish text, and the hidden states are the six labels: *noun*, *adjective*, *pronoun*, *numeral*, *particle*, *verb*, and *other*.

2. The emission probabilities in HMM model refers to the conditional distribution of the observed variables from the specific state. If the observed values Xn are discrete, the probabilities ∅ is a K*D table of probabilities, of K hidden states and D symbols (words).

   $\mathrm{B} = (b_{ij}) = P(y_i|x_j)$

   Note: x_j is the hidden state, y_i is the observation derived from the hidden state.

   In this case, emission probabilities refer to the probabilities of a certain Finnish Symbol derived from a given label out of the six.

3. Transition probability refers to the probability of observing a particular state given that the hidden model is in a particular hidden state.

   $\mathrm{A} = (a_{ij}) = P(x_{it}|x_{jt-1})$

   Note: x_j is the hidden state, x_{jt-1} is the previous hidden state.

   As described in the question, in Finnish, adjectives that define a noun tend to occur before the noun. Which means, in Finnish, the transition probability

   adj → noun >> noun → adj

$P(noun|adj) = p_1 \gg P(adj|noun) = p_2$.

In Portuguese, on the contrary, the adjective defined a noun tend to occur after the noun.

Then we have:

noun → adj >> adj → noun

$P(adj|noun) = p_3 \gg P(noun|adj) = p_4$

Then we can expect that, the transition probability from an adjective to a noun in Finnish is greater than that in Portuguese, rather than smaller.

References

[1] A. S. Zeleznikow, “Chapter 9 Evaluation, Deployment,” in *Knowledge Discovery from Legal Databases And Related Issues*, Dordrecht, Springer, 2005, p. 174.

[2] “Tutorial,” hmmlearn developers (BSD License), 2016. [Online]. Available: http://hmmlearn.readthedocs.io/en/latest/tutorial.html. [Accessed 13 10 2016].