NLP Ass4

Name: Ziheng Wang

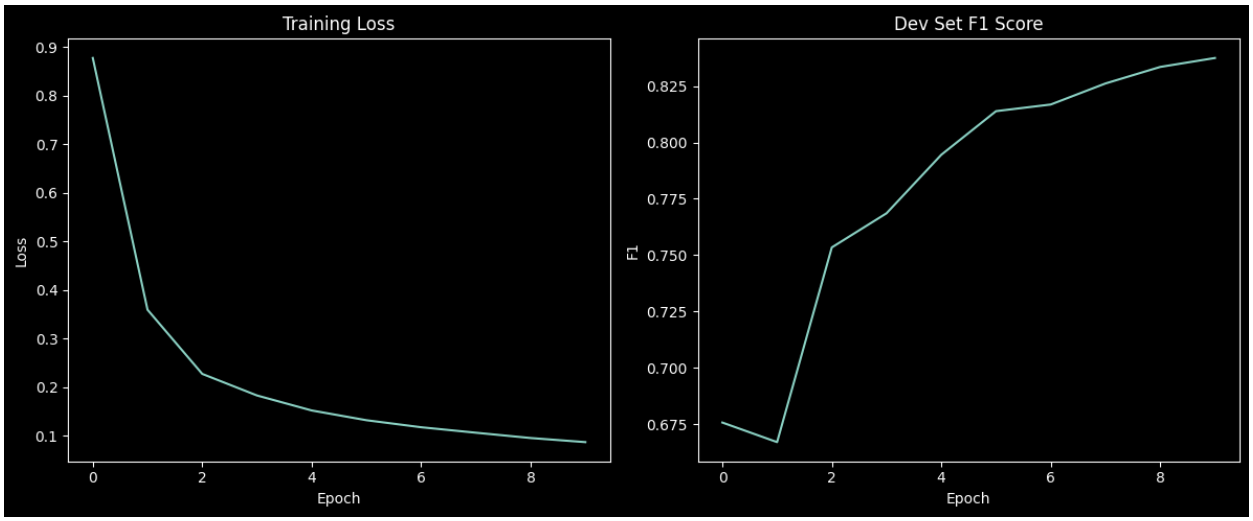
SID: 12310401

All the result can be found in the notebook, Make Sure you have download the glove datasets and unzip them.

Step 123: Implication and F1 score of the BILSTM model

Dataset/Epoch	F1 score
Dev/1	0.6758
Dev/2	0.6671
Dev/3	0.7535
Dev/4	0.7686
Dev/5	0.7945
Test/Final	0.7860

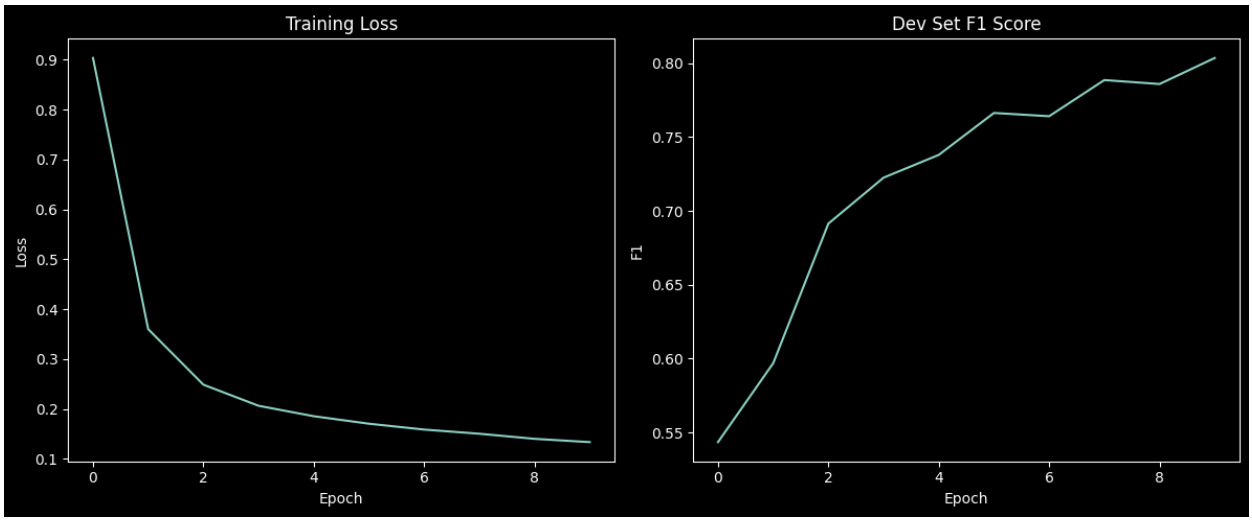
```
开始训练...
训练: 100%|██████████| 110/110 [00:07<00:00, 13.89it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.15it/s]
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:5: RuntimeWarnin
  return x[1, 1]/(x[1, 1] + x[0, 1])
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:9: RuntimeWarnin
  return x[1, 1] / (x[1, 0] + x[1, 1])
Epoch 1/10, Loss: 0.8772, Dev set F1: 0.6758
Epoch 1 Dev set F1 score: 0.6758
训练: 100%|██████████| 110/110 [00:07<00:00, 14.19it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.15it/s]
Epoch 2/10, Loss: 0.3592, Dev set F1: 0.6671
Epoch 2 Dev set F1 score: 0.6671
训练: 100%|██████████| 110/110 [00:07<00:00, 13.94it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.08it/s]
Epoch 3/10, Loss: 0.2272, Dev set F1: 0.7535
Epoch 3 Dev set F1 score: 0.7535
训练: 100%|██████████| 110/110 [00:07<00:00, 14.20it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.25it/s]
Epoch 4/10, Loss: 0.1828, Dev set F1: 0.7686
Epoch 4 Dev set F1 score: 0.7686
训练: 100%|██████████| 110/110 [00:07<00:00, 14.72it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.13it/s]
Epoch 5/10, Loss: 0.1519, Dev set F1: 0.7945
Epoch 5 Dev set F1 score: 0.7945
训练: 100%|██████████| 110/110 [00:07<00:00, 14.68it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 11.76it/s]
Epoch 6/10, Loss: 0.1318, Dev set F1: 0.8139
训练: 100%|██████████| 110/110 [00:07<00:00, 14.74it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.52it/s]
Epoch 7/10, Loss: 0.1174, Dev set F1: 0.8168
训练: 100%|██████████| 110/110 [00:07<00:00, 14.15it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.22it/s]
Epoch 8/10, Loss: 0.1061, Dev set F1: 0.8262
训练: 100%|██████████| 110/110 [00:07<00:00, 14.12it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.08it/s]
Epoch 9/10, Loss: 0.0952, Dev set F1: 0.8335
训练: 100%|██████████| 110/110 [00:07<00:00, 14.09it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 12.27it/s]
Epoch 10/10, Loss: 0.0868, Dev set F1: 0.8374
Evaluating on the test set...
评估: 100%|██████████| 27/27 [00:01<00:00, 14.02it/s]
Test set F1 score: 0.7860
```



Bonus1 Implication of MEMM

You can find my code at A4_memm.ipynb.

```
开始MEMM模型训练...
训练: 100%|██████████| 110/110 [00:10<00:00, 10.19it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 10.18it/s]
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:5: RuntimeW
  return x[1, 1]/(x[1, 1] + x[0, 1])
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:9: RuntimeW
  return x[1, 1] / (x[1, 0] + x[1, 1])
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:15: Runtime
  return ((1 + beta**2)*precision*recall)/(beta**2 * precision +
Epoch 1/10, Loss: 0.9035, Dev set F1: 0.5435, TF ratio: 1.00
训练: 100%|██████████| 110/110 [00:12<00:00,  9.12it/s]
评估: 100%|██████████| 26/26 [00:03<00:00,  8.39it/s]
Epoch 2/10, Loss: 0.3602, Dev set F1: 0.5969, TF ratio: 0.90
训练: 100%|██████████| 110/110 [00:13<00:00,  8.16it/s]
评估: 100%|██████████| 26/26 [00:03<00:00,  8.49it/s]
Epoch 3/10, Loss: 0.2489, Dev set F1: 0.6914, TF ratio: 0.80
训练: 100%|██████████| 110/110 [00:13<00:00,  8.09it/s]
评估: 100%|██████████| 26/26 [00:03<00:00,  8.51it/s]
Epoch 4/10, Loss: 0.2066, Dev set F1: 0.7225, TF ratio: 0.70
训练: 100%|██████████| 110/110 [00:14<00:00,  7.84it/s]
评估: 100%|██████████| 26/26 [00:03<00:00,  8.41it/s]
Epoch 5/10, Loss: 0.1857, Dev set F1: 0.7379, TF ratio: 0.60
训练: 100%|██████████| 110/110 [00:13<00:00,  8.08it/s]
评估: 100%|██████████| 26/26 [00:03<00:00,  8.45it/s]
Epoch 6/10, Loss: 0.1707, Dev set F1: 0.7663, TF ratio: 0.50
训练: 100%|██████████| 110/110 [00:13<00:00,  8.24it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 10.30it/s]
Epoch 7/10, Loss: 0.1590, Dev set F1: 0.7640, TF ratio: 0.40
训练: 100%|██████████| 110/110 [00:12<00:00,  9.07it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 10.80it/s]
Epoch 8/10, Loss: 0.1506, Dev set F1: 0.7885, TF ratio: 0.30
训练: 100%|██████████| 110/110 [00:12<00:00,  9.09it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 10.74it/s]
Epoch 9/10, Loss: 0.1404, Dev set F1: 0.7858, TF ratio: 0.20
训练: 100%|██████████| 110/110 [00:12<00:00,  9.07it/s]
评估: 100%|██████████| 26/26 [00:02<00:00, 10.49it/s]
Epoch 10/10, Loss: 0.1338, Dev set F1: 0.8035, TF ratio: 0.10
```



Bonus2 Implication of Beam Search

You can find my code at A4_ner_beam.ipynb.

```
Compare Decode method:
-----------------------------------------------------------------------
Model + Decode method          Dev set F1      Test set F1
-----------------------------------------------------------------------
Evaluate with Greedy: 100%|██████████| 26/26 [00:02<00:00, 12.61it/s]
Evaluate with Greedy: 100%|██████████| 27/27 [00:01<00:00, 14.29it/s]
BiLSTM + Greedy search         0.8374          0.7860
Evaluate with Beam: 100%|██████████| 26/26 [00:49<00:00,  1.91s/it]
Evaluate with Beam: 100%|██████████| 27/27 [00:44<00:00,  1.65s/it]
BiLSTM-CRF + Beam search (width=3) 0.8437         0.7952
Evaluate with Beam: 100%|██████████| 26/26 [01:22<00:00,  3.16s/it]
Evaluate with Beam: 100%|██████████| 27/27 [01:13<00:00,  2.73s/it]
BiLSTM-CRF + Beam search (width=5) 0.8443         0.7960
Evaluate with Beam: 100%|██████████| 26/26 [02:39<00:00,  6.15s/it]
Evaluate with Beam: 100%|██████████| 27/27 [02:22<00:00,  5.30s/it]
BiLSTM-CRF + Beam search (width=10) 0.8445         0.7952
```

It is a little bit better than the greedy search. And the width of the beam search 3,5,10 have no significant difference on f1 score but the time cost is much longer while the width is larger.

Bonus3 Implication of CRF

You can find my code at A4_ner_CRF_Viterbi.ipynb.

The CRF model's F1 score is lower than the previous BILSTM model at the first epoch, but it quickly surpasses the BILSTM model and achieves a higher F1 score at the end of training.

```
Train CRF: 100%|██████████| 110/110 [00:13<00:00,  8.42it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:03<00:00,  8.40it/s]
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:5: RuntimeWa
  return x[1, 1]/(x[1, 1] + x[0, 1])
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:9: RuntimeWa
  return x[1, 1] / (x[1, 0] + x[1, 1])
/home/stu_12310401/nlp/SUSTech-NLP25/Ass4/metrics.py:15: RuntimeW
  return ((1 + beta**2)*precision*recall)/(beta**2 * precision +
Epoch 1/10, Loss: 11.5426, Dev set F1: 0.5564
Train CRF: 100%|██████████| 110/110 [00:12<00:00,  8.61it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:03<00:00,  8.33it/s]
Epoch 2/10, Loss: 5.2235, Dev set F1: 0.6394
Train CRF: 100%|██████████| 110/110 [00:12<00:00,  8.62it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 10.67it/s]
Epoch 3/10, Loss: 3.4265, Dev set F1: 0.7305
Train CRF: 100%|██████████| 110/110 [00:11<00:00,  9.96it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 10.98it/s]
Epoch 4/10, Loss: 2.6634, Dev set F1: 0.7713
Train CRF: 100%|██████████| 110/110 [00:10<00:00, 10.01it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.03it/s]
Epoch 5/10, Loss: 2.2818, Dev set F1: 0.8028
Train CRF: 100%|██████████| 110/110 [00:11<00:00,  9.98it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.06it/s]
Epoch 6/10, Loss: 1.9735, Dev set F1: 0.8025
Train CRF: 100%|██████████| 110/110 [00:10<00:00, 10.07it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.04it/s]
Epoch 7/10, Loss: 1.7476, Dev set F1: 0.8202
Train CRF: 100%|██████████| 110/110 [00:11<00:00,  9.95it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.03it/s]
Epoch 8/10, Loss: 1.5591, Dev set F1: 0.8295
Train CRF: 100%|██████████| 110/110 [00:10<00:00, 10.05it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.05it/s]
Epoch 9/10, Loss: 1.4060, Dev set F1: 0.8397
Train CRF: 100%|██████████| 110/110 [00:11<00:00, 10.00it/s]
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.09it/s]
Epoch 10/10, Loss: 1.2792, Dev set F1: 0.8453
```

```
Compare BiLSTM with BILSTM-CRF:
-------------------------------------------------------------------
Model + Decode method          Dev set F1      Test set F1
-------------------------------------------------------------------
Evaluate : 100%|██████████| 26/26 [00:02<00:00, 12.69it/s]
Evaluate : 100%|██████████| 27/27 [00:01<00:00, 14.04it/s]
BiLSTM + Greedy search          0.8374          0.7860
Evaluate CRF: 100%|██████████| 26/26 [00:02<00:00, 11.11it/s]
Evaluate CRF: 100%|██████████| 27/27 [00:02<00:00, 12.04it/s]
BiLSTM-CRF + Viterbi            0.8453          0.7927
```