HW1_Named Entity Recognition (NER)

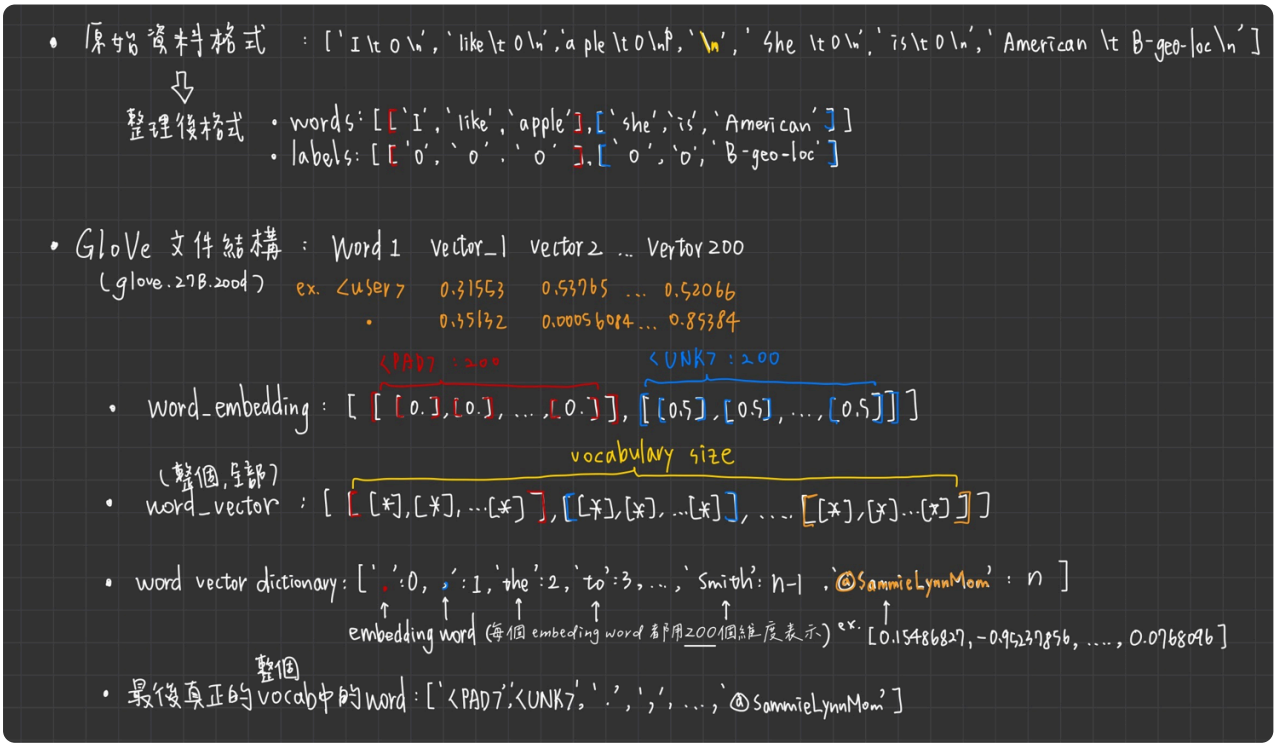
NER tasks using GloVe, BiLSTM, CRF [GloVe] (https://nlp.stanford.edu/projects/glove/), [CRF] (http://www.jionlp.com/lecture/logistics_crf)

Data Loading

- **Dataset:** WNUT-2016
  targets: person, geo-location, company, facility, product, music artist, movie, sports team, TV show, and other
- **Evaluation tools:** conlleval
  https://raw.githubusercontent.com/sighsmile/conlleval/master/conlleval.py (https://raw.githubusercontent.com/sighsmile/conlleval/master/conlleval.py)
- **Pre-trained word vectors:** glove.twitter.27B.200d
  https://nlp.stanford.edu/projects/glove/ (https://nlp.stanford.edu/projects/glove/)

Pre-Processing

1. **Process raw data into a specific format**
   - **word:** [["word"], ["word1", "word2", "word3"]]
   - **target:** [["target"], ["target1", "target2", "target3"]]



2. **GloVe**

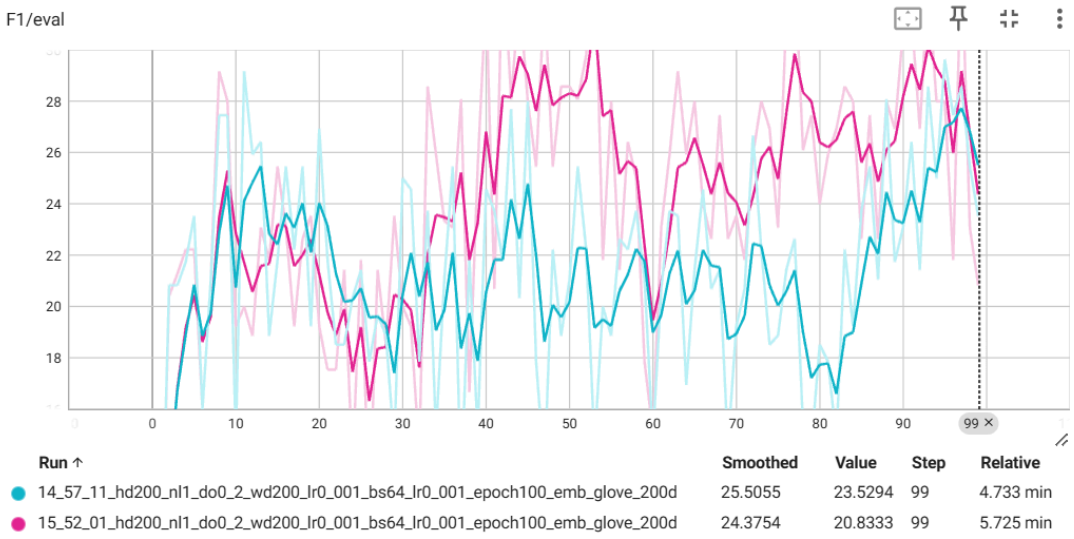
   使用先前做好的 glove_embeddings 初始化 nn.Embedding 層的權重矩陣，使得在自然語言處理任務中，每個 word vector 能反映出事先訓練的 GloVe 所能獲得的資訊

   使用 `glove.twitter.27B.200d`

   - 載入並建立 `glove_embedding` 字典方便後續使用
   - 挑出維度不為 200 的向量

3. **Data Cleaning**

   - 將資料集中的使用者特定文字以 `glove.twitter.27B.200d` 原先擁有的特殊 Token 替代
     - "RT"–>"<retweet>"
     - "http\S+|www\S+"–>"<url>"
     - "@\S+"–>"<user>"
     - "#\S+"–>"<hashtag>"
   - 若該 word 之小寫形式在 `glove.twitter.27B.200d` 中，則替換為小寫
   - 是否替換數字為標籤 <number> ?
     - 粉色: 替換 (後選用)
     - 藍色: 無



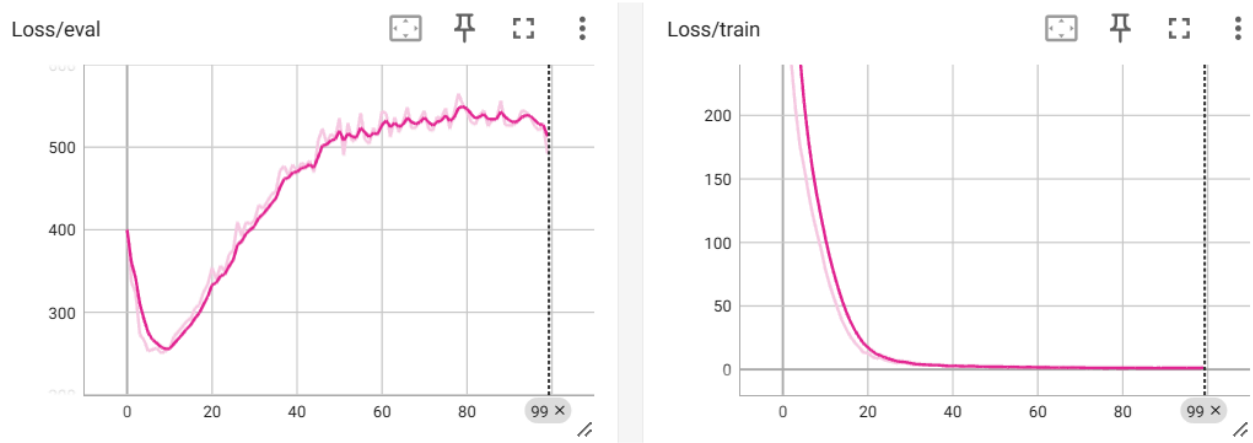
6. **word, tag <–> index**

    - word 的 index 0 位置加入<PAD> (向量全0)、index 1加入<UNK> (向量全0.5)
    - tag 的 index 0 加入 <PAD>
7. **Padding the sentence**
   方便 batch 運算，填充使所有句子長度一致

Model Architecture

- Parameter:
    - embedding_dim = 200
    - hidden_dim = 200
    - batch_size = 64
    - epochs = 100
    - nn.Dropout(p=0.2)
- BiLSTM
    - embedding
    - nn.LSTM(embedding_dim=200, hidden_dim=200, num_layers=1, dropout=0.2, bidirectional = True)
    - nn.Linear(hidden_dim * 2, target_size=22)
- CRF
- Optimizer: AdamW, Learning rate = 0.001

Training Process



Evaluation

- 評估工具: `conlleval.py`

```
processed 17261 tokens with 661 phrases; found: 528 phrases; correct: 189.
accuracy:  28.01%; (non-O)
accuracy:  94.04%; precision:  35.80%; recall:  28.59%; FB1:  31.79
          company: precision:  26.83%; recall:  28.21%; FB1:  27.50  41
         facility: precision:  35.00%; recall:  18.42%; FB1:  24.14  20
          geo-loc: precision:  40.98%; recall:  43.10%; FB1:  42.02  122
            movie: precision:   0.00%; recall:   0.00%; FB1:   0.00  9
      musicartist: precision:   0.00%; recall:   0.00%; FB1:   0.00  4
            other: precision:   5.05%; recall:   3.79%; FB1:   4.33  99
           person: precision:  54.10%; recall:  57.89%; FB1:  55.93  183
          product: precision:   8.33%; recall:   5.41%; FB1:   6.56  24
       sportsteam: precision:  75.00%; recall:  21.43%; FB1:  33.33  20
           tvshow: precision:   0.00%; recall:   0.00%; FB1:   0.00  6
```

Predict Result

在檔案 `result.txt` 中