Final Project report

50.040 Natural Language Processing

Lee Jet Xuen 1004365

Tay Sze Chang 1004301

Brandon Chong Wah Jin 1004104

Folder Structure

– Final Project Submission
    – final_project_part1-4.ipynb
    – final_project_part5.ipynb
    – final_project_part6i.ipynb
    – final_project_part6ii.ipynb
    – README.md
    – requirements.txt
    – conlleval.py
    – eval.py
    – dataset
        – dev.p2.out
        – dev.p4.out
        – dev.p5.out
        – test.p6.model.out
        – test.p6.CRF.out

Instructions to run the code

Refer to the readme.md

Evaluation Results

The conlleval.py is taken from the source on github[1].

Part 2

processed 3809 tokens with 154 phrases; found: 210 phrases; correct: 71.
accuracy: 43.06%; (non-O)
accuracy: 93.25%; precision: 33.81%; recall: 46.10%; FB1: 39.01
        negative: precision: 13.85%; recall: 36.00%; FB1: 20.00 65
        neutral: precision: 12.50%; recall: 33.33%; FB1: 18.18 8
        positive: precision: 44.53%; recall: 48.41%; FB1: 46.39 137
((33.80952380952381, 46.103896103896105, 39.010989010989015), 0)

Part 4

processed 3809 tokens with 149 phrases; found: 210 phrases; correct: 68.

accuracy: 50.66%; (non-O)
accuracy: 93.49%; precision: 32.38%; recall: 45.64%; FB1: 37.88
negative: precision: 15.38%; recall: 52.63%; FB1: 23.81 65
neutral: precision: 0.00%; recall: 0.00%; FB1: 0.00 8
positive: precision: 42.34%; recall: 44.62%; FB1: 43.45 137
((32.38095238095238, 45.63758389261745, 37.883008356545965), 0)

Part 5

processed 3809 tokens with 44 phrases; found: 210 phrases; correct: 14.
accuracy: 50.00%; (non-O)
accuracy: 92.41%; precision: 6.67%; recall: 31.82%; FB1: 11.02
negative: precision: 0.00%; recall: 0.00%; FB1: 0.00 65
neutral: precision: 0.00%; recall: 0.00%; FB1: 0.00 8
positive: precision: 10.22%; recall: 32.56%; FB1: 15.56 137
((6.666666666666667, 31.818181818181817, 11.023622047244094), 0)

Part 6i with dev.in

processed 3809 tokens with 56 phrases; found: 210 phrases; correct: 20.
accuracy: 31.82%; (non-O)
accuracy: 92.26%; precision: 9.52%; recall: 35.71%; FB1: 15.04
negative: precision: 4.62%; recall: 25.00%; FB1: 7.79 65
neutral: precision: 12.50%; recall: 16.67%; FB1: 14.29 8
positive: precision: 11.68%; recall: 42.11%; FB1: 18.29 137
((9.523809523809524, 35.714285714285715, 15.037593984962406), 0)

Part 6ii with dev.in

processed 3809 tokens with 177 phrases; found: 210 phrases; correct: 81.
accuracy: 44.35%; (non-O)
accuracy: 93.12%; precision: 38.57%; recall: 45.76%; FB1: 41.86
negative: precision: 16.92%; recall: 36.67%; FB1: 23.16 65
neutral: precision: 0.00%; recall: 0.00%; FB1: 0.00 8
positive: precision: 51.09%; recall: 47.95%; FB1: 49.47 137
((38.57142857142858, 45.76271186440678, 41.860465116279066), 0)

Approaches

Part 6i

We created a Bigram and Trigram model to calculate their respective transition values. We first get the states (labels) that are tied to the input file. Afterwards, we calculate the number of bigrams and trigrams in the dataset, followed by their transition value respectively.

Dictionaries created based on the bigrams' transitions values are created, which are further parsed into the trigrams' transition dictionary calculations. With the final dictionary created, it is parsed into the Viterbi Algorithm, chopping off the "START" and "STOP" sign in the sequences we parse.

Part 6ii

In the last part of this project, we implemented a fairly simple architecture, namely BiLSTM-CRF. While Long Short Term Memory(LSTM) is sufficient for the task such as named entity recognition, adding a layer of CRF will certainly improve the result compared to pure Bi-LSTM. It has also proved that CRF improved the result of such tasks significantly[2].

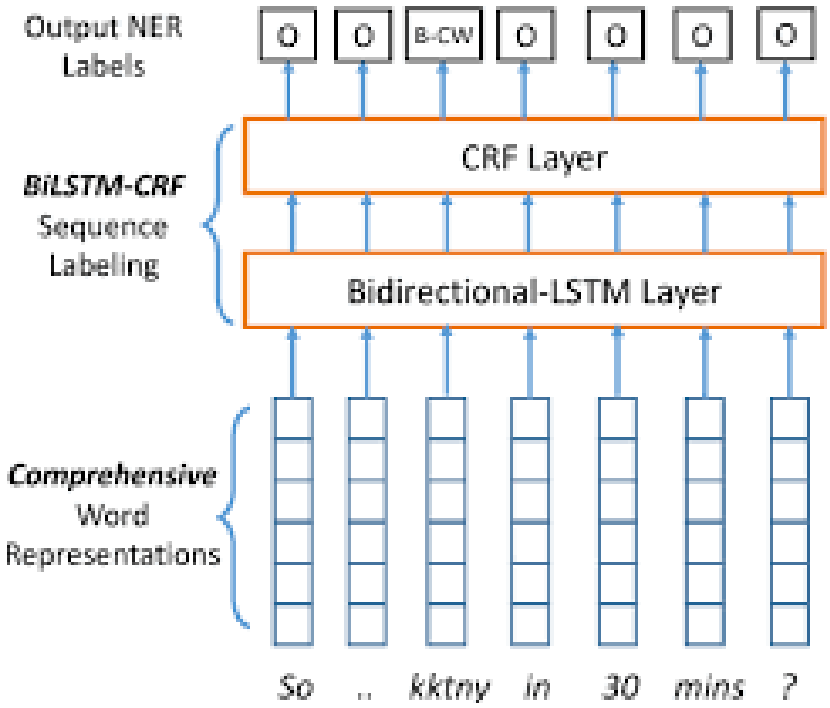

Figure 1. Architecture of a BiLSTM-CRF model

The overall score is as followed:

$$P(x|y) = \frac{exp(Score(x, y))}{\sum (Score(x, y'))}$$

In each of the Bi-LSTM cells, we have both the transition and emission scores. Hence, the Score(x,y) is dependent on both scores:

$$Score(x, y) = \sum_i \left(\log\left(P_{emission}(y_i \rightarrow x_i) \right) + \log\left(P_{transition}(y_{i-1} \rightarrow y_i) \right) \right)$$

References:

1. Colleval Package: https://github.com/sighsmile/conlleval
2. Panchendrarajan, Rrubaa, and Aravindh Amaresan. 1AD. Review of *Bidirectional LSTM-CRF for Named Entity Recognition*. In . Hong Kong.