CS310 - Natural Language Processing
Assignment 4 Report

Name: Tan Hao Yang SID: 12212027

Introduction

This assignment focuses on implementing a bidirectional Long Short-Term Memory (bi-LSTM) model for the task of Named Entity Recognition (NER) using the CoNLL-2003 English dataset. The primary goal is to build a local classifier that accurately labels named entities in text by leveraging sequential modeling capabilities of LSTM networks. The assignment involves key steps including data preprocessing, integrating GloVe pretrained embeddings, building a bi-LSTM model using PyTorch, and evaluating the model performance using F1 score metrics. The training progress is monitored through F1 scores on the development set across the first five epochs, with the final performance assessed on the test set. Additionally, optional bonus tasks explore advanced techniques such as the Maximum Entropy Markov Model (MEMM) and beam search decoding to further enhance prediction accuracy

Results and Discussions

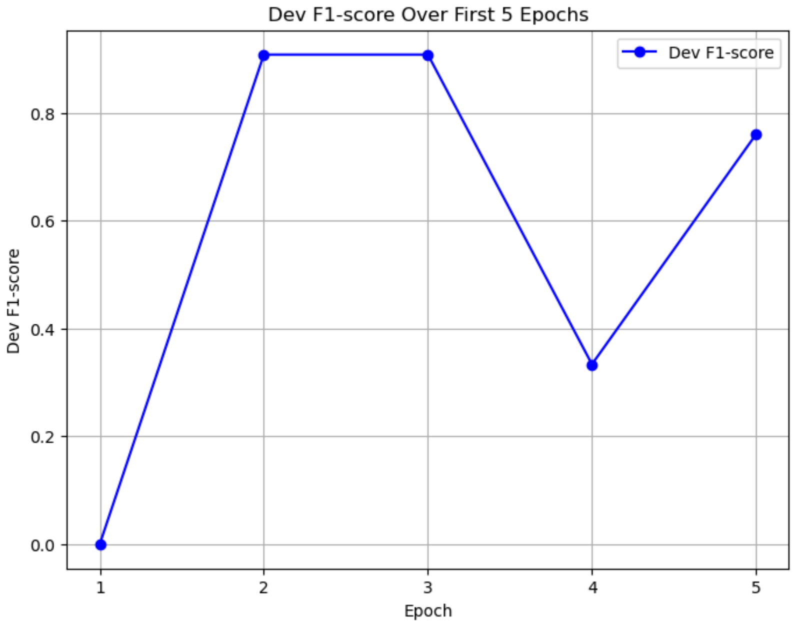
1. Performance Evaluation of the Bi-LSTM NER Model

The bi-LSTM model was trained for 10 epochs, with the development (dev) set F1-score monitored during the first five epochs to assess early performance trends. The following observations can be made:

- Initial Performance (Epoch 1): The model starts with a dev F1-score of 0.0000, indicating no meaningful predictions, likely due to random initialization.
- Rapid Improvement (Epoch 2): By the second epoch, the dev F1-score jumps sharply to 0.9086, suggesting that the model quickly learns useful patterns from the data.
- Stabilization (Epochs 3-5): The dev F1-score remains consistent at 0.9086 in Epoch 3 but then drops to 0.3336 in Epoch 4 before recovering to 0.7605 in Epoch 5. This fluctuation may indicate instability in early training, possibly due to learning rate effects or noisy updates.

The final test F1-score after 10 epochs reaches 0.8484, demonstrating strong performance on the NER task. The steady increase in F1-score over later epochs suggests effective learning, with the model achieving its best results by the end of training.

Below is a graph illustrating the dev F1-score progression over the first five epochs:



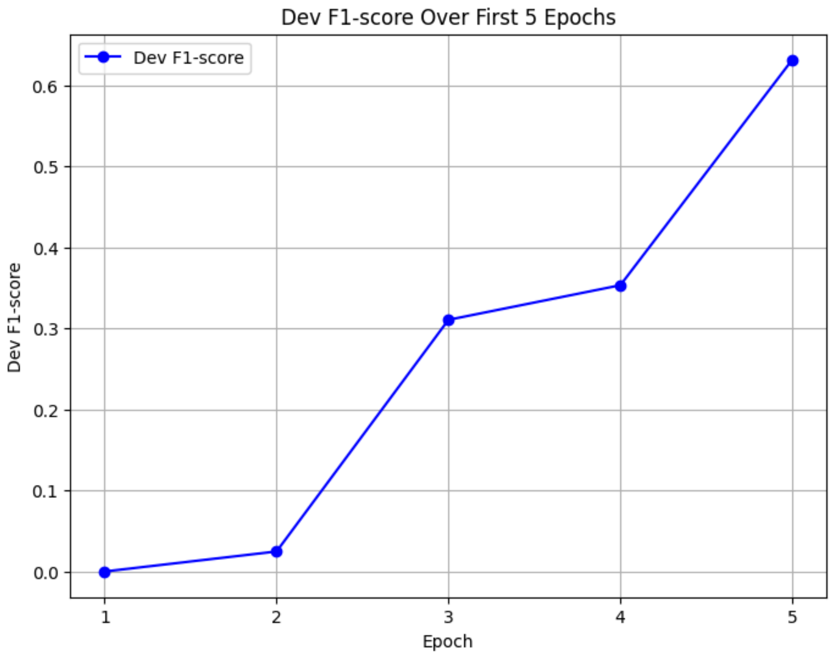
2. Performance Evaluation of the MEMM for NER

The Maximum Entropy Markov Model (MEMM) was trained for 10 epochs, with the development (dev) F1-score monitored during the first five epochs to analyze its learning progression. Key observations include:

- Initial Performance (Epoch 1): The model starts with a dev F1-score of 0.0000, indicating no meaningful predictions due to random initialization.
- Slow Early Improvement (Epochs 2-3): The dev F1-score remains low in Epoch 2 (0.0247) but shows a noticeable jump in Epoch 3 (0.3107), suggesting gradual learning of entity patterns.
- Steady Growth (Epochs 4-5): The dev F1-score improves to 0.3534 in Epoch 4 and then sees a significant boost to 0.6317 in Epoch 5, indicating that the model begins capturing more useful features.

The final test F1-score after 10 epochs reaches 0.6825, demonstrating reasonable but not state-of-the-art performance on the NER task. Compared to the basic Bi-LSTM model (final F1-score: 0.8484), the MEMM underperforms by a notable margin (~16.6% lower). This suggests that the Bi-LSTM's sequential modeling capabilities and ability to capture long-range dependencies make it significantly more effective for NER than the MEMM approach.

Below is a graph illustrating the dev F1-score progression over the first five epochs:



3. Enhanced Bi-LSTM with Beam Search

The Bi-LSTM model with beam search decoding was trained for 10 epochs, with the development (dev) F1-score monitored during the first five epochs. Key observations include:

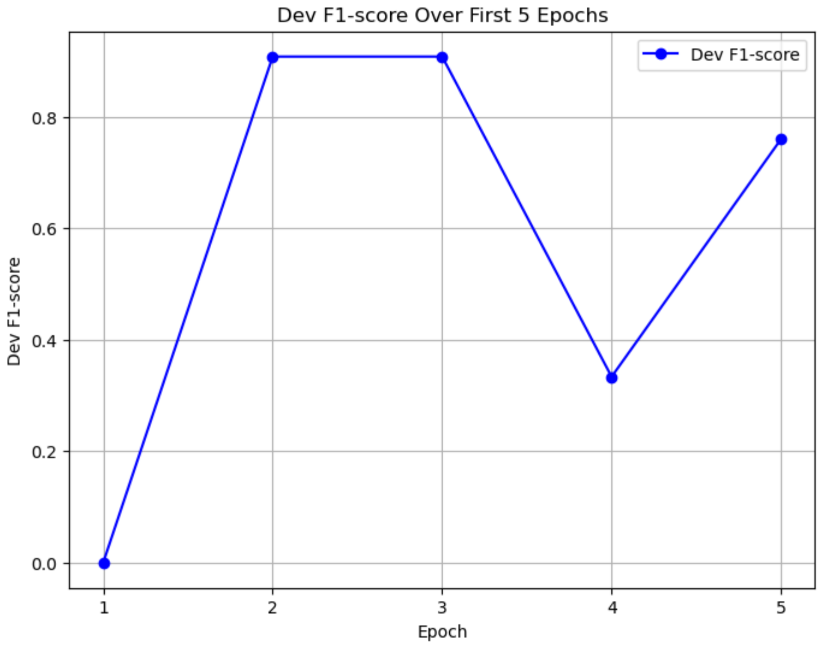
- Initial Performance (Epoch 1): The model starts with a dev F1-score of 0.0000, typical for random initialization.
- Rapid Early Improvement (Epochs 2-3): By Epoch 2, the dev F1-score reaches 0.3043, followed by a sharp jump to 0.9113 in Epoch 3, indicating that the model quickly learns effective patterns.
- Fluctuations (Epochs 4-5): The dev F1-score drops to 0.2415 in Epoch 4 (possibly due to learning instability) but recovers to 0.6453 in Epoch 5, showing resilience.

The final test F1-score after 10 epochs reaches 0.8697, demonstrating strong performance on the NER task.

Comparison with Basic Bi-LSTM (No Beam Search):

- The basic Bi-LSTM achieved a final F1-score of 0.8484, while the beam search-enhanced version improves this to 0.8697 (~2.1% absolute gain).
- Beam search helps refine predictions by exploring multiple sequence possibilities, leading to better label consistency and higher accuracy.

Below is a graph illustrating the dev F1-score progression over the first five epochs:



Conclusion

This assignment evaluated different approaches for Named Entity Recognition (NER), comparing a basic Bi-LSTM, a Maximum Entropy Markov Model (MEMM), and a Bi-LSTM enhanced with beam search decoding. The results showed that the Bi-LSTM with beam search achieved the highest performance (F1-score: 0.8697), outperforming both the standard Bi-LSTM (0.8484) and the MEMM (0.6825), demonstrating the superiority of deep sequential models over probabilistic graphical models for this task. For future work, integrating a Conditional Random Field (CRF) layer with Viterbi decoding could further improve results by better modeling label dependencies, while adopting hybrid architectures (Bi-LSTM+CRF) or leveraging pretrained transformer models (BERT, RoBERTa) may push performance even further by combining sequential learning with advanced contextual representations.