Part 3:

Considerations:

For the embedding, we initialized full random weight matrix like before, then for each word in our vocabulary from train, that has a corresponding pre-trained vector, we put the pre-trained vector instead the random one.
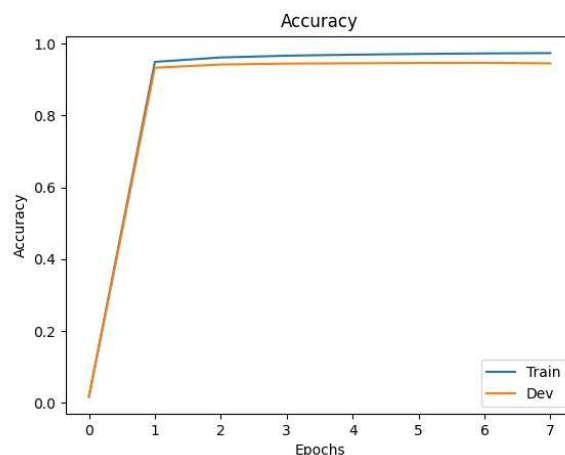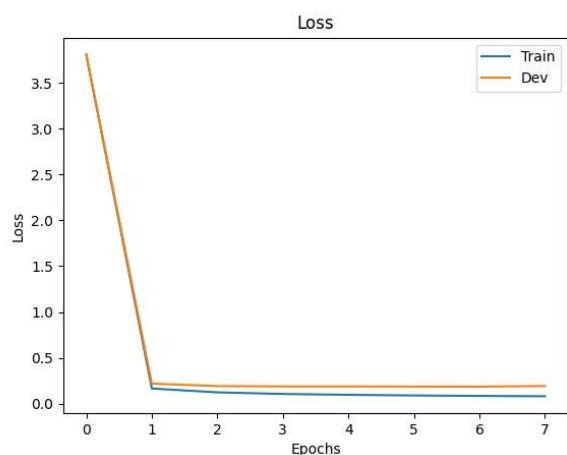
We handled the lower-casing issue by lower-casing all words the moment we read them. In test, we also saved original words for writing in the output file.

All the rest remained as part1.

The results did not get better than part one, but we reached the peak faster, because part of the training was already done for us.

Result graphs:

Pos: 95% on dev

Ner: 96% on dev