2020Fall COMP4901K Project Report

Weiqi Wang, Yiduo Yu, Kaixing Wu

{wwangbw}@connect.ust.hk

1. Introduction

Named entity recognition is a common task in the field of natural language processing, where each word in the training/validation set is labelled by a tag and the task is to predict the label of words in the testing set that either appeared or never appeared in the training set. Current approaches mainly include Ontology-based NER and deep learning NER. While ontology-based NER requires a significant level of details of the ontology, the result of NER can be very comprehensive or specific to a particular topic. Deep learning approach, however, is much more accurate than ontology as it is capable to assemble words by word embedding. It is capable of understanding the semantic and syntactic relationship between various words. In this project, we'll implement two deep learning approaches to solve the NER in COVID-19 and report the results correspondingly.

2. Methodology

Our initial approach is **BERT classification**. BERT is the state of the art language model for NLP, by applying the bidirectional training of transformers, which is a popular attention mode, to language modelling, BERT demonstrated a deeper sense of language context and flow than single-direction language models. Using the pre-trained BERT model released by Hugging Face, we can encode and decode each word with the BERT Tokenizer. By adding a token classification layer on top of the traditional BERT model, we can achieve the goal of solving NER tasks with BERT.

However, since BERT is a pre-trained model based on a certain range of corpus, and the task we are trying to solve in the project is mainly related to COVID-19, BERT is not an ultimate choice as few, or even none of the pretraining corpus is related to COVID19. This led to our second approach, where we used pre-trained **Glove** as the word embedding and a **Bidirectional LSTM** model plus a conditional random field layer as the best methodology. In the next section, we'll introduce more detail in both approaches.

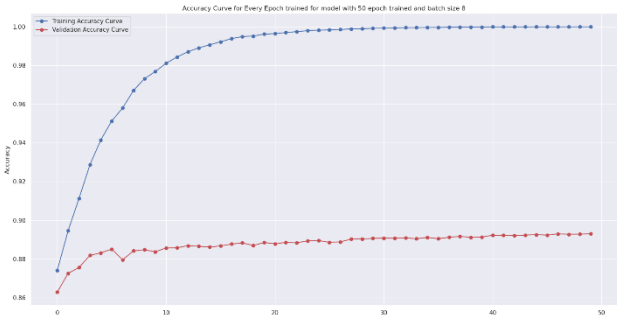
3. Experiment setup

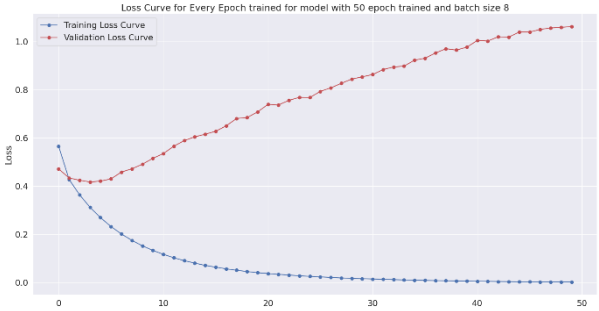
In our initial approach, we used transformer 3.4.0 and PyTorch 1.7.0 as our development framework. The transformer package released by Hugging Face contains the BERT tokenizer which enables us to encode all the words into unique integers, with these integers as inputs and tags as outputs, we can fine-tune the pre-trained BERT model, after which we are capable of generating the prediction for the encoded-words in the test dataset. As we introduced in the previous section, the BERT is pre-trained on a large corpus where COVID-19 is not likely to appear. Considering that the data in this problem is a domain knowledge and BERT may be difficult in dealing with those words, we need our second approach, which is Glove+BiLSTM. We used the 6B_100d Glove released in https://nlp.stanford.edu/projects/glove/ as our glove

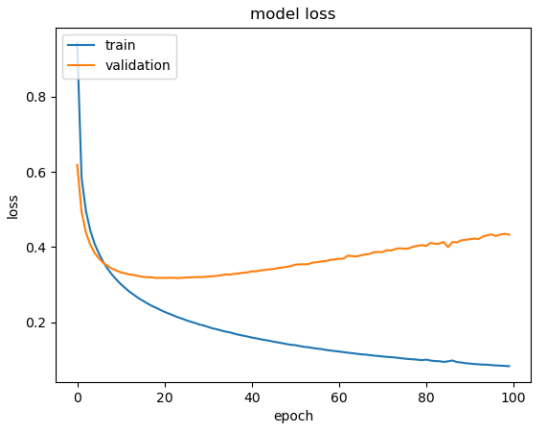
vocabulary, it is reported that 76% words can be found in this embedding dictionary and 100 dimensions are sufficient for us to deal with this task. For words that have not appeared in Glove, such as some COVID-19 proper nouns, we adopt a random initialization method to randomly initialize a one-hundred-dimensional vector for these words. Excessive dimensionality may lead to redundant computing power requirements, and more inaccurate word vector results, so instead of choosing 300 dimensions, we chose 100 dimensions. The deep learning framework is implemented under Tensorflow 2.3.1 and Keras 2.4.3, where embedding layer, Bidirectional layer, LSTM layer are all supported, we only need to do some preprocessing to fit our neural network. In the next section, we'll demonstrate the results and performance of both approaches.

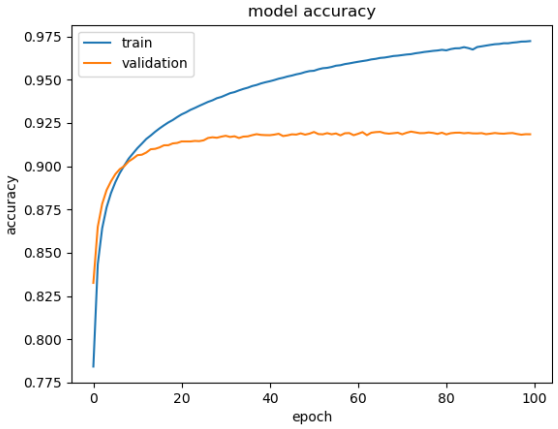
4. Results

To further enhance the model's performance, We tuned the hyperparameters of the models, including learning rate, weight decay, learning rate decay, epochs, batch size, optimizer, etc. It is reported that the BERT achieved the highest accuracy of 89.37% on the validation set while Glove+BiLSTM achieved 92.07%. Some figures for the highest accuracy model during the training process are attached below.









Figures in the first row are the accuracy and loss for BERT classification and the second row is for Glove+BiLSTM respectively. We can see that the model is overfitting and the accuracy on the validation set is fluctuating after a certain number of epochs. However, the result of the validation set is still promising, thus we can choose Glove+BiLSTM as our final result.

All codes and datasets are available at https://github.com/wwangbw/COMP4901K-Project.