Named Entity Recognition with PyTorch

Harrison Jansma
harrison.jansma@utdallas.edu

Abstract

In this project I was tasked with implementing a machine learning application to perform the classic NLP task of Named Entity Recognition on the CoNLL 2003 dataset. A logistic regression and LSTM models were trained on this data, achieving F1 scores of 0.926 and 0.899 respectively.

1 Introduction

Named Entity Recognition is a classic Natural Language Processing task wherein a model is created to associate the words in a document as references to a specific person, place, or thing. In this project we identified five different tags for a word.

- “O” – No named entity recognized
- “ORG” – Organization
- “PER” – Person
- “LOC” - Location
- “MISC” – Miscellaneous

Each word in the training set is labelled as one of these five categories. Given this prelabelled data, we trained logistic regression and LSTM models to detect the presence of a named entity within a document. Each model reached high performance on the data, with the logistic regression being the top performer.

2 Data Extraction

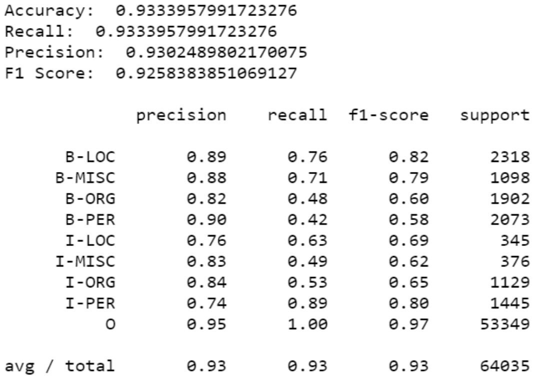
From the prelabelled data, we were able to extract various features that were useful in training machine learning classifiers. Of note, we gathered data on the Parts of Speech (POS) tag for the word in question, as well as the POS tag for the prior word. We also gathered words that had some syntactic relationship to the word contained in the document.

Unfortunately, POS tags and syntactic relationship features were only included in the training of the logistic regression model, The LSTM model was limited to training on the Word2Vec embedding of the word contained within the document. This may account for the higher performance of the logistic regression model. In future implementations of Named Entity Recognition, I would like to attempt concatenation of derived features with embeddings obtained from the Word2Vec model.

Feature	Logistic Regression	LSTM
Word	One Hot Encoded word representation	Word2Vec representation
POS Tag	OHE Encoded	Not Included
Prior POS Tag	OHE Encoded	Not Included
Syntactically Related Words	OHE word representation	Not Included

3 Logistic Regression Results

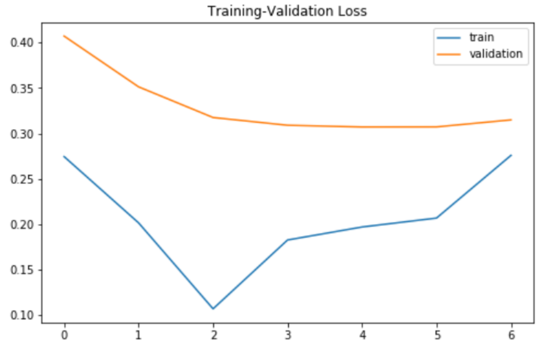
The logistic regression model outperformed the LSTM. This may be due to the logistic regression model having a richer feature representation of the document. (POS tags, syntactically related words) Results of the logistic regression are as follows.

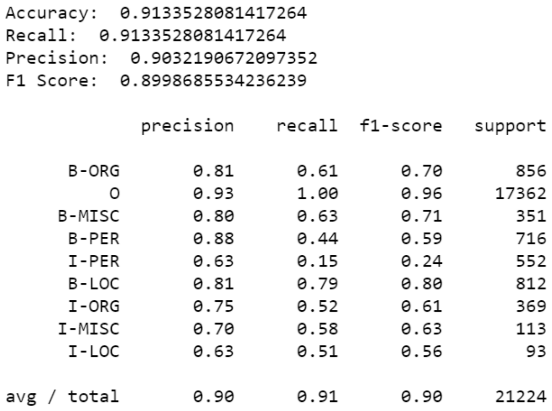


```
Accuracy:  0.9333957991723276
Recall:  0.9333957991723276
Precision:  0.9302489802170075
F1 Score:  0.9258383851069127

             precision    recall  f1-score   support

      B-LOC       0.89      0.76      0.82      2318
     B-MISC       0.88      0.71      0.79      1098
      B-ORG       0.82      0.48      0.60      1902
      B-PER       0.90      0.42      0.58      2073
      I-LOC       0.76      0.63      0.69       345
     I-MISC       0.83      0.49      0.62       376
      I-ORG       0.84      0.53      0.65      1129
      I-PER       0.74      0.89      0.80      1445
          O       0.95      1.00      0.97     53349

avg / total       0.93      0.93      0.93     64035
```

4 LSTM Results

The LSTM model was implemented with PyTorch. The model made use of Word2Vec embeddings to represent the textual data within the dataset. Results are as follows.





```
Accuracy:  0.9133528081417264
Recall:  0.9133528081417264
Precision:  0.9032190672097352
F1 Score:  0.8998685534236239

             precision    recall  f1-score   support

      B-ORG       0.81      0.61      0.70       856
          O       0.93      1.00      0.96     17362
     B-MISC       0.80      0.63      0.71       351
      B-PER       0.88      0.44      0.59       716
      I-PER       0.63      0.15      0.24       552
      B-LOC       0.81      0.79      0.80       812
      I-ORG       0.75      0.52      0.61       369
     I-MISC       0.70      0.58      0.63       113
      I-LOC       0.63      0.51      0.56        93

avg / total       0.90      0.91      0.90     21224
```

Conclusion

Named Entity Recognition is a challenging task studied extensively within the literature. Our implementation utilized Scikit-Learn and PyTorch to train machine learning models to classify the presence of Organizations, Persons, Locations, or Miscellaneous named entities within the CoNLL 2003 dataset.

Specifically, logistic regression and LSTM models were trained to predict one of five categories for each word in the training dataset. Results showed that the logistic regression model outperformed the LSTM, achieving an F1-score of 0.926 while the LSTM attained a F1 score of 0.899. This difference in performance was most likely due to derived features that were included in the training set of the logistic regression and not the LSTM.

References

Tomas Mikolov. 2013. Distributed Representations of Words and Phrases and their Compositionality American Psychological Association. 1983.

https://pytorch.org/

https://scikit-learn.org/stable/