Multilingual HW1 Report

Phoebe Li

Language	Test Accuracy	Test Loss
en	91.85 %	0.266
cs	94.35 %	0.186
es	93.30 %	0.278
ar	94.24 %	0.168
af	88.53 %	0.388
lt	75.65 %	0.721
hy	80.02 %	0.620
ta	39.84 %	1.862

Table 1: Model Using optim.Adam

Abstract

This report analyzes how BiLSTM model's accuracy in predicting different language tags is affected by the language family, typology and data set size. It also exploits potentials in using optimizer to improve the model accuracy.

1 Introduction

The BiLSTM (Bi-directional Long Short Term Memory) model combines forward LSTM and backward LSTM. Its functionality of encoding the information from back to front enables it to better capture bidirectional semantics.

The major task for this assignment involves using the BiLSTM to predict the sequence of labels for each token in the sentence. The model tested eight different languages and returned different results.

2 Baseline Model Result

The Baseline model contains an embedding layer, a Bi-directional LSTM layer, followed by a liner layer with dropout. After testing the model in different languages, I discovered significant variations in model performance. (Table 1)

Based on the observations from Table 1, I divided the results into three groups: languages with an accuracy rate over 90%, languages with an accuracy rate between 60% to 89%, and languages with an accuracy rate below 60%. The results indicated that Tamil, a language spoken in North India, has a significantly low accuracy rate at 39.84 percent. As the rest of the languages are majorly spoken in Europe, I suspected that geographic location could have an impact on Tamil's low accuracy.

3 Baseline Model Result Analysis

3.1 Performance changes across three languages group

The result indicates that Tamil prediction accuracy is significantly lower than other languages. As we only have 400 training samples for Tamil, the small data set size could have caused a negative effect on the prediction. Through examining other data sets, I found that the size of the training set has a positive correlation with the test accuracy. As English and Spanish sets have over 10,000 training sets, their test accuracy was significantly higher than others.

From a linguistic perspective, Tamil belongs to the Dravidian languages family, which is spoken by 220 million people living mainly in southern India, northeast Sri Lanka, and South Asia. Dravidian languages have many branches, and Tamil has the largest number of speakers. Unlike other languages, Tamil does not have an equivalent for the existential verb "to be"; it is included in the translations only to convey the meaning. The negative existential verb, "to be not", however, does exist in the form of illai and goes at the end of the sentence (does not change with the number, gender, or tense). The verb "to have" in the meaning "to possess" is not translated directly as well. For instance, to say "I have a horse" in Tamil, a construction equivalent is "There is a horse to me" or "There exists a horse to me ."This special structure could be the reason that leads to the low accuracy.

3.2 Performance changes across different tags

By comparing the prediction and test data set, we can see that the model also performs differently

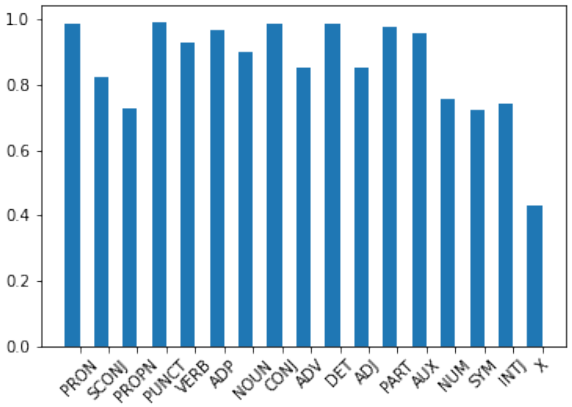

Figure 1: English Tags Accuracy Difference

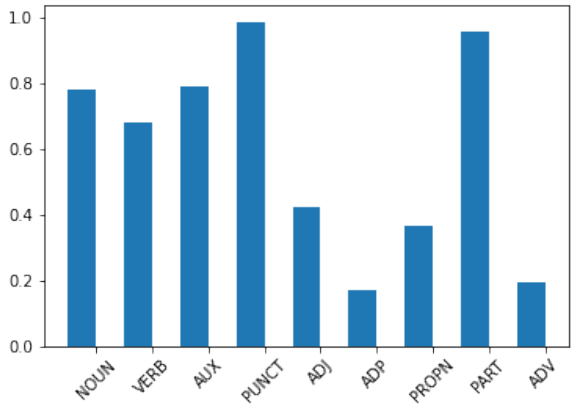

Figure 2: Tamil Tags Accuracy Difference

on different language tags. For example, English (with over 90 percent accuracy) has higher accuracy on PUNCT (99.32%) and lower accuracy on NUM (76.59%). There are 14 punctuation marks that are commonly used in English grammar, including period, question mark, exclamation point, comma, semicolon, colon, dash, hyphen, parentheses, brackets, braces, apostrophe, quotation marks, and ellipsis. English's punctuation is relatively simple compared to other languages. On the other hand, Spanish has paired punctuation ¿? and exclamation mark ¡!, and French has different quotation marks: « ». German also has double quotation marks:" ". As for the numerical system, English number words include numerals and have various words derived from them, as well as many words that were borrowed from other languages.

Tamil has fewer languages tags than other languages. It also shows that the Punctuation prediction performs well on the model. However, for Tamil's ADP, the accuracy rate is low as 16.92%.

Tamil uses postpositional elements for ADP. Postpositional elements are frequently found in head-final languages such as Basque, Estonian, Finnish, Georgian, Korean, Japanese, Hindi, Urdu, Bengali, and Tamil. The word or other morpheme that corresponds to an English preposition occurs after its complement, hence the name postposition.

According to Schiffman's Tamil Case System analysis, Tamil is using missionary grammars (NMG) as consisting of a finite number of cases.), to some of which postpositional suffixes may be added. The syntax of the Tamil is usually dealt with in approximately the same manner as the inventory of morphemes: we are told that the case in question is "governed" by various constraints, such as that the accusative is "governed" by the presence of certain transitive verbs, and that nouns that are the objects of verbs are marked accusative. (Schiffman, Schiffman,) This unique case and syntax system may cause the poor performance on the Tamil language tag predictions.

In addition Tamil also has identical person pronouns system. The first person plural pronouns in Tamil distinguish between inclusive and exclusive "we". In Tamil, plural terminators are used for honorific addressing. It could be noted in both 2nd and 3rd persons. The pronouns "we" could represents the one person with honor, rather than a group of people. That could also leads the lower accuracy on PROPN than other languages. (Schiffman, 1999)

4 Improvement

Instead of using Adam optimizer, I choose to implement a new model that uses Adagrad as the optimizer. The result indicates that the new model performance has significantly increased on ta, hy, af. However, other languages did not show a significant influence. The new model impacted ta the most, with a significant improvement from 39.84% to 61.82%, which ultimately lowered the test loss for all languages. (Table 2)

On the other hand, Adagrad also eliminates the need to manually tune the learning rate.s, which leads to faster convergence and less sensitivity to the size of the master step. With the optimizer, I am able to significantly improve the result despite the fact that Tamil has limited number of training data sets.

5 Summary

Although BiLSTM model does an excellent job in predicting language tags for certain languages, the

Language	Test Accuracy	Test Loss
en	91.29 %	0.274
cs	93.96 %	0.186
es	93.27 %	0.273
ar	94.23 %	0.167
af	90.91 %	0.309
lt	76.91 %	0.689
hy	81.34 %	0.524
ta	61.82 %	1.218

Table 2: Model Using optim.AdaGrad

model accuracy could be affected by many factors, including language family, typology and data set size. The accuracy can be significantly improved by using optimizers such as adagrad.

References

Gunner, J. (n.d.). What are the 14 punctuation marks in English grammar? Grammar. Retrieved February 9, 2022, from https://grammar.yourdictionary.com/punctuation/what/fourteen-punctuation-marks.html

Schiffman , H. F. (n.d.). The Tamil Case System - University of Pennsylvania. Retrieved February 9, 2022, from https://ccat.sas.upenn.edu/ haroldfs/public/h$_s ch_9 a.pdf$