Multilingual Part-Of-Speech Tagging

Charity King
Carnegie Mellon University
csking@andrew.cmu.edu

Abstract

This paper evaluates a baseline Bidirectional Long Short-Term Memory (BiLSTM) model for Multilingual Part-Of-Speech Tagging and provides a comparison on accuracy/loss rates for both high and low-resource languages.

1 Introduction

Languages categorized as low-resource are a major focus and challenge within the natural language processing (NLP) community, which enables human-to-human communication, human-machine communication, and language analysis [3]. The majority of the world's approximately 7,000 languages are low-resource languages, typified by lacking a large monolingual or parallel corpora for machine translation and largely lacking other linguistic resources required for NLP development [4]. Sequence labeling is a critical NLP task, which seeks to categorize an observed text sequence through pattern recognition. Part-of-Speech tagging is a popular form of sequence labeling, involving the prediction of a corresponding part-of-speech tag such as 'NOUN' or 'VERB' for any word within a sentence. This project will evaluate a benchmark Bidirectional Long Short-Term Memory (BiLSTM) model on Universal Dependency (UD) treebank data (comprised of parts-of-speech, morphological features and syntactic dependencies) for eight different languages across the resource spectrum and analyze the model's strengths and weaknesses.

2 Methods

Recurrent Neural Networks such as a BiLSTM model are typically used for sequential learning problems where context is critical for a learning model. Additionally, the BiLSTM has been shown to be an effective learner for NLP tasks such as part-of-speech tagging [1]. For this project, we used eight different UD language data-sets for the baseline BiLSTM model; English (en), Afrikaans (af), Czech (cs), Spanish (es), Arabic (ar), Lithuanian (lt), Armenian (hy), and Tamil (ta). We define a high-resource data-set as containing at least 60,000 non-unique tokens. Of the eight languages, only English, Czech, Spanish, and Arabic are considered high-resource languages (*see Table 1*).

Lang	Training	Total Tokens
English (en)	12,543	204,586
Afrikaans (af)	1,315	33,894
Czech (cs)	41,559	719,317
Spanish (es)	14,187	382,436
Arabic (ar)	6,174	225,853
Lithuanian (lt)	2,341	47,605
Armenian (hy)	1,975	42,105
Tamil (ta)	400	6,329

Table 1: Universal Dependency Training Data-set Description

Each UD language data-set was split into a training, testing, and validation set with varying splits per data set (a potential design flaw for this benchmark assessment), but all languages having at least approximately 70 percent of the total data used for training purposes. Our UD data-sets contain sequences of sentences of a target language, along with a label sequence of the same length indicating the Part-of-Speech tag associated to each word or token within a sequence. The BiLSTM model's input is a sequence of word tokens, where the model predicts a sequence of Part-of-Speech labels of the same length. Due to this neural network providing contextual access during both the forward and backward propagation iterations, our model has increased access to prior and after data for any token within a sequence.

We trained our baseline model against each lan-

Language	Model 1	Model 2
English (en)	91.58%	91.51%
Afrikaans (af)	88.53%	91.23%
Czech (cs)	94.05%	94.20%
Spanish (es)	93.30%	93.30%
Arabic (ar)	94.24%	94.61%
Lithuanian (lt)	75.65%	77.04%
Armenian (hy)	80.02%	81.01%
Tamil (ta)	39.84%	47.38%

Table 2: Model 1 benchmark and comparison Model 2 Test Accuracy performance

guage training data-set to determine loss rate and accuracy scores. This model was configured with a 2-layer neural network, dropout rate of .25, and learning over 10 epochs. Due to the scarcity of training data for the resource poor languages, we ran the baseline model with a few parameter adjustments to determine if this would increase performance. We trained a second iteration of the baseline model (Model 2) with a 3-layer neural network, dropout rate of .50, and learning over 20 epochs to address our languages with fewer labeled training data. Additionally by increasing our dropout rate to .5, we are regularizing our network in order to avoid over-fitting during our training cycle by increasing the probability that a random neuron will be deactivated. Our goal for Model 2 is to increase accuracy for our data-scarce languages while maintaining accuracy for our other language models.

3 Evaluation

Our benchmark model (Model 1) had varying accuracy across the languages with our best results for the Arabic and Czech languages around approximately 94% accuracy, followed by Spanish with 93.3% accuracy. With the addition of English, all of the high-resource languages scored at least 91% accuracy (*see Table 2*). In contrast, our worst performing model was for the Tamil language, with an accuracy rate of 39.84%. This poor performance is not surprising since our Tamil training data-set consisted of only 400 labeled training instances with 6,329 tokens. The next smallest data-set Afrikaans has 1,315 labeled instances and 33,894 tokens. In comparison, our largest data-set Czech had 41,559 training labels and 719, 317 total tokens.

We ran another iteration of our baseline model with parameter changes (Model 2) to include in-

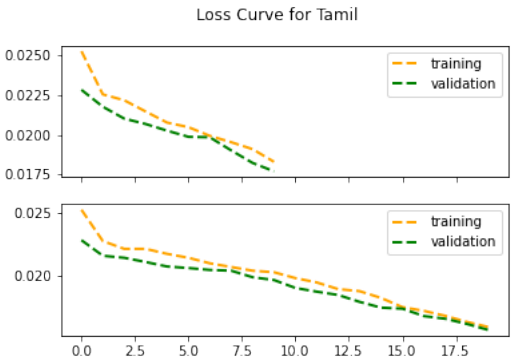

Figure 1: Loss Curve for Model 1 across 10 epochs (top) and Model 2 across 20 epochs (bottom)

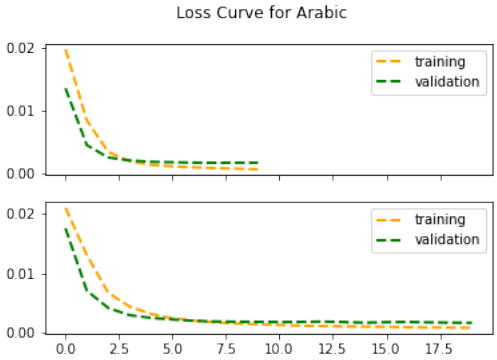

Figure 2: Loss Curve for Model 1 across 10 epochs (top) and Model 2 across 20 epochs (bottom)

creasing our random node dropout-rate, adding another internal neural network layer, and increasing our training epochs from 10 to 20. These configuration changes were made to determine if accuracy can be improved with more epoch training and regularization of the network in order to avoid an over-fitting problem for our models with more scare training data-sets. Across all languages, the mean accuracy increase was 1.63%. If we analyze the mean average for high-resource vs low-resource languages, we obtain a .11% versus a 3.16% increase respectively. However, our model increased in accuracy by 7.54% for Tamil, our data-set with the smallest training samples and tokens available.

Over-fitting was a large concern for model 2. For our primary target language Tamil, our loss curve between both models indicate that our validation testing continues to decrease with our training loss rate, a good indication that our model is fitted in a balanced way (*see Figure 1*). Additionally, both loss curves maintain stability across epochs in a downward slope, indicating there is still more learning the model can accomplish. Our best per-

forming language across both models demonstrates potentially slight over-fitting for both models with a validation curve that slightly increases with more experience. This paired with minimal model accuracy increases indicates that for learning a model with large data-sets, longer training iterations result in negligible performance increases and potential over-fitting.

4 Discussion

Overall, the method of simply tweaking learning parameters and extending the learning time per model had modest results for most of the languages, but more analysis is required on the actual error rates per model to determine which part-of-speech tags are responsible for the majority of mislabeled tags and better understand these distributions per language. A longer training period aided with an increased drop-out rate to control regularization resulted in a substantial increase in accuracy for our most sparse data-set and should be incorporated in more development when annotated data is scarce. This comparison high-lighted the importance of training data for bringing up baseline accuracy levels.

Further development on BiLSTM models that incorporate linear conditional random fields (CRFs) [2] have shown great promise in sequential text data, enabling a model to account for conditional probability distributions for a set of random observations.

References

[1] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks.

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc.

[3] Graham Neubig. Low-resource NLP tasks.

[4] Yulia Tsvetkov. Opportunities and challenges in working with low-resource languages.