Assignment 3 NLP

1 n-gram Language Model

1.2 n-gram Language Modeling

	Corpus Length	Unique n-grams
Unigram Model	1622905	26602
Bigram Model	1684434	510392
Trigram Model	1684433	1116160

Figure 1: Model Information

	Train Perplexity	Validation Perplexity	Test Perplexity
Unigram Model	1105.5403	1009.9143	1015.3158
Bigram Model	77.0735	inf	inf
Trigram Model	7.8279	inf	inf

Figure 2: Model Perplexity Scores (unsmoothed)

Train perplexity decreases as the history increases, thereby suggesting better performance in the model from unigram to trigram. As UNK token was used in the unigram model, probability would be non-zero for the validation and test perplexity. In contrast, since the probability of certain bigram/trigram were found to be 0, the perplexity obtained is infinite for their respective validation and test perplexity

1.3 Smoothing

	Train Perplexity	Validation Perplexity	Test Perplexity
$\lambda_{1,2,3}$ = 0.01, 0.10, 0.99	7.6296	2126.3137	2121.6016
$\lambda_{1,2,3}$ = 0.05, 0.15, 0.95	7.8492	935.2956	934.4221
$\lambda_{1,2,3}$ = 0.10, 0.10, 0.80	9.3821	739.8144	739.4445
$\lambda_{1,2,3}$ = 0.10, 0.30, 0.50	14.4137	793.1056	793.7961
$\lambda_{1,2,3}$ = 0.66, 0.66, 0.66	10.0358	239.5573	240.1129
$\lambda_{1,2,3}$ = 0.10, 0.30, 0.60	**12.0533**	**736.4886**	**736.9305**

Figure 3: Model Perplexity Scores and Hyperparameters

1.3.3 If you use half of the training data, would it increase or decrease the perplexity on previously unseen data? Why? Provide empirical experimental evidence if necessary.
The perplexity would increase as the training data decreases since there is less information and tokens to be utilized by the model. Below are the perplexity scores trained on half the data set using interpolation smoothing

	Train Perplexity	Validation Perplexity	Test Perplexity
$\lambda_{1,2,3}$ = 0.10, 0.30, 0.60	NA	468.3740	471.4108

Figure 4: Model Perplexity Scores Based on ½ Training Set

1.3.4 If you convert all tokens that appeared less than 5 times to (a special symbol for out-of-vocabulary tokens), would it increase or decrease the perplexity on the previously unseen data compared to an approach that converts only a fraction of words that appeared just once to ? Why? Provide empirical experimental evidence if necessary.
More UNK tokens would lead to a model that has less perplexity and is more generalizable. Below are the perplexity scores after changing the UNK Threshold from 3 to 5 on a unigram model.

	Train Perplexity	Validation Perplexity	Test Perplexity
Unigram Model	909.6217	853.7755	856.9755

Figure 5: Unigram Model Perplexity Score with UNK Threshold of 5

The Unigram model with the higher UNK threshold showed a decrease in perplexity compared to Figure 2.

2 Neural Language Modeling, based on Eisenstein 6.10 (p. 136)

2.1.1 Fully describe your model architecture, hyperparameters, and experimental procedure.

The neural language model is trained using vanilla RNN with the epoch=6, learning rate=0.001, and batch size=128.

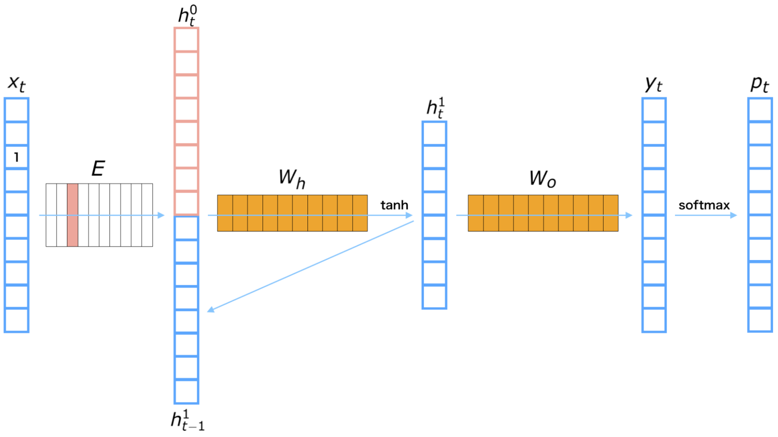

Figure 6: One layer of the RNN Language Model[1]

The process used to get a next word prediction from the i-th input word x_t is as follows:

1) Get the embedding vector: $h_t^{(0)} = Ex_t$
2) Calculate the hidden layer: $h_t^{(1)} = \tanh\left(\mathbf{W}_h \begin{bmatrix} \mathbf{h}_t^{(0)} \\ \mathbf{h}_{t-1}^{(1)} \end{bmatrix}\right)$
3) Calculate the output layer: $y_t = W_0 h_t^{(1)}$
4) Transform to probability: $p_t = \text{softmax}(y_t)$

2.1.2 After each epoch of training, compute your LM's perplexity on the development data. Plot the development perplexity against # of epochs. Additionally, compute and report the perplexity on test data.

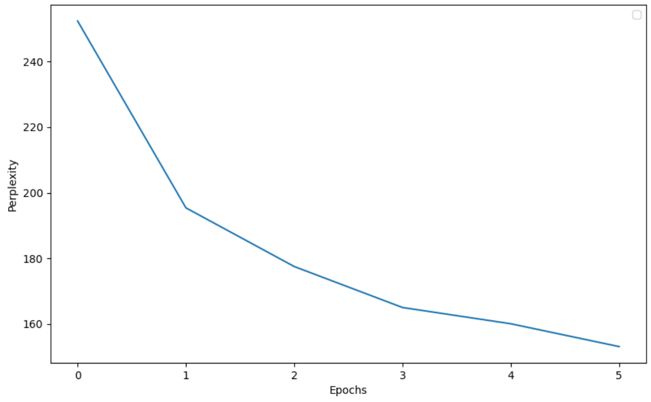

Figure 7: Vanilla RNN Language Model Training Perplexity

The perplexity on the test dataset was computed to be 142.65

[1] https://docs.chainer.org/en/stable/examples/ptb.html

3 What Can Language Models Do?

I am more or less convinced by the paper’s argument. Based on the paper’s definition of linguistic meaning to be “the relation between a linguistic form and communicative intent”, it would mean there requires some level of sentience in order to have an intent in trying to communicate something before just probabilities that a machine has learnt. In our current state of what machines can do, we are far from achieving that level of autonomy and self-awareness that is required to attain intent in speech. Beyond the technical study of NLP, I believe there needs to be a broader study on the sense of what it means for a machine to be able to communicate and what that means if a machine can do so in relation to humans.