1 Word-Level Neural Bi-gram Language Model

1.a

First let us denote by t the index in which y has the 1. For simplicity, we assume that the base of the log is e. We get that

$$CE(\hat{y}) = -\sum_i y_i log(\hat{y}_i) = -log(\hat{y}_t) = -log(\frac{exp(\theta_t)}{\sum_j exp(\theta_j)}) = -\theta_t + log(\sum_j exp(\theta_j))$$

Let us look at the right term of the last equation. We get that

$$\frac{\partial log(\sum_j exp(\theta_j))}{\partial \theta_i} = \frac{exp(\theta_i)}{\sum_j exp(\theta_j)} = \hat{y}_i$$

For the term we that $\frac{\partial - \theta_t}{\partial \theta_i} = -1$ if $i = t$ and 0 otherwise. To summarize, we get

$$\frac{\partial CE}{\partial \theta_i} = \begin{cases} \hat{y}_i - 1, \ i = t \\ \hat{y}_i, \ i \neq t \end{cases} = \hat{y} - y$$

Note that the partial derivative for θ_t is negative, which makes sense since when applying SGD we subtract the derivative from the current value of θ_t, hence we would enlarge θ_t and minimize the loss. For all other $i \neq t$ the derivative is positive.

1.b

We will use the chain rule. First, let us denote:

$$f_1 = xW_1 + b_1$$

$$f_2 = hW_2 + b_2$$

By the chain rule we get:

$$\frac{\partial CE}{\partial x} = \frac{\partial CE}{\partial softmax} * \frac{\partial softmax}{\partial f_2} * \frac{\partial f_2}{\partial h} * \frac{\partial h}{\partial \sigma} * \frac{\partial \sigma}{\partial f_1} * \frac{\partial f_1}{\partial x}$$

Now let us calculate each of the partial derivatives & then we can substitue. As we've seen in the previous question:

$$\frac{\partial CE}{\partial softmax} * \frac{\partial softmax}{\partial f_2} = \begin{cases} \hat{y}_i - 1, \ i = t \\ \hat{y}_i, \ i \neq t \end{cases} = \hat{y} - y = softmax(f_2) - y$$

where $\hat{y} = softmax(f_2)$ and t is the index in which y has the 1.

$$\frac{\partial f_2}{\partial h} = W_2^T$$

$$\frac{\partial h}{\partial \sigma} = \sigma(1 - \sigma)$$

$$\frac{\partial f_1}{\partial x} = W_1^T$$

We end up with:

$$(softmax(f_2) - y) * W_2^T * (\sigma(f_1)(1 - \sigma(f_1))) * W_1^T$$

Note that we need to turn $(\sigma(f_1)(1 - \sigma(f_1)))$ into a square and diagonal matrix with size D_h^2.

1.d

We computed a perplexity of: 112.1805 (after training for 40000 epochs)

2 Theoretical Inquiry of a Simple RNN Language Model

2.a

The notation used in the question is incorrect, we assume all vectors are "column" vectors e.g. $h^{(t)} \in \mathbb{R}^{D_h}$, and that the appropriate transposed were omitted in the question.

2.a.i $\frac{\partial J^{(t)}}{\partial U}(\theta)$

First we compute $\frac{\partial J^{(t)}}{\partial U}(\theta)$. Recall that because $y^{(t)}$ is a one-hot vector our loss simplifies to:

$$J^{(t)}(\theta) = -{y_k}^{(t)} \log \hat{y}_k^{(t)} \tag{1}$$

where k is the index of the target word in the vocabulary.

Next, let us denote ${o^{(t)}}^T = {h^{(t)}}^T U + b_2^T$, recall our previous computation of the differential of the cross entropy loss with a one-hot vector:

$$\frac{\partial J^{(t)}}{\partial o^{(t)}}(\theta) = \hat{y}^{(t)} - y^{(t)} \tag{2}$$

Furthermore we have:

$$\frac{\partial o_l^{(t)}}{\partial U_{i,j}}(\theta) = \mathbb{I}_{\{l=j\}} h_i^{(t)} \tag{3}$$

Combining the last results yields:

$$\frac{\partial J^{(t)}}{\partial U_{i,j}}(\theta) = h_i^{(t)} (\hat{y}^{(t)} - y^{(t)})_j \tag{4}$$

which is recognizable as the outer product:

$$\boxed{\frac{\partial J^{(t)}}{\partial U}(\theta) = h^{(t)} \otimes (\hat{y}^{(t)} - y^{(t)})} \tag{5}$$

2.a.ii $\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}}(\theta)$

Next we compute $\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}}(\theta)$, note that this gradient must be back-propagated through time, but we defer it to section b. Again applying the chain rule we have:

$$\frac{\partial J^{(t)}}{\partial L_{i,j}}(\theta) = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial o^{(t)}} \frac{\partial o^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial L_{i,j}} \tag{6}$$

where $\frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial o^{(t)}} = \frac{\partial J^{(t)}}{\partial o^{(t)}} = \hat{y}^{(t)} - y^{(t)}$ was already computed in (2). The third term in (6) is immediate by the definition of the differential of a linear function:

$$\frac{\partial o_l^{(t)}}{\partial h_m^{(t)}} = (U^T)_{l,m} = U_{m,l} \tag{7}$$

Lastly we derive the fourth term in (6) using the chain-rule without unfolding through time. Let us denote $a^{(t)} = H^T h^{(t-1)} + I^T e^{(t)} + b_1^T$, then:

$$\frac{\partial h^{(t)}}{\partial L_{i,j}} = \frac{\partial \sigma}{\partial a^{(t)}} \left(\frac{\partial}{\partial L_{i,j}} \left(L_{x^{(t)}} I \right) + \frac{\partial}{\partial L_{i,j}} \left({h^{(t-1)}}^T H \right) \right) \tag{8}$$

For this section, we make the simplifying assumption that $x^{(t_1)} \neq x^{(t_2)}$ for any $t_1 \neq t_2$, as implied in the question, which implies that $\frac{\partial}{\partial L_{x^{(t)}}} \left({h^{(t-1)}}^T H \right) = 0$.

Recalling the derivative of the sigmoid yields:

$$\frac{\partial \sigma_k(a^{(t)})}{\partial a_l^{(t)}}(\theta) = \sigma(a_k^{(t)}) \left(1 - a_k^{(t)} \right) \frac{\partial a_k^{(t)}}{\partial a_l^{(t)}} \tag{9}$$

Thus we have:

$$\frac{\partial \sigma(a^{(t)})}{\partial a^{(t)}}(\theta) = \text{diag}\left[h^{(t)} \right] \text{diag}\left[\vec{1} - h^{(t)} \right] \tag{10}$$

A simple computation yields:

$$\frac{\partial}{\partial L_{x^{(t)}}} \left(L_{x^{(t)}} I \right) = I^T \tag{11}$$

Plugging the above computations into (6) gives:

$$\boxed{\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}}(\theta) = \left(\hat{y}^{(t)} - y^{(t)} \right)^T U^T \text{diag}\left[h^{(t)} \right] \text{diag}\left[\vec{1} - h^{(t)} \right] I^T} \tag{12}$$

2.a.iii $\frac{\partial J^{(t)}}{\partial I}(\theta)$

In a similar fashion to (6), we apply the chain rule:

$$\frac{\partial J^{(t)}}{\partial I_{i,j}}(\theta) = \frac{\partial J^{(t)}}{\partial \hat{y}_k^{(t)}} \frac{\partial \hat{y}_k^{(t)}}{\partial o_l^{(t)}} \frac{\partial o_l^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial I_{i,j}} \tag{13}$$

Note that only the last term in (17) differs from (6), futhermore in a similar fashion to (8):

$$\frac{\partial h^{(t)}}{\partial I_{i,j}} = \frac{\partial \sigma}{\partial a^{(t)}} \left(\frac{\partial}{\partial I_{i,j}} \left({e^{(t)}}^T I \right) + \frac{\partial}{\partial I_{i,j}} \left({h^{(t-1)}}^T H \right) \right) \tag{14}$$

Again a simple computation yields:

$$\frac{\partial}{\partial I_{i,j}} \left({e^{(t)}}^T I \right) = e_i^{(t)} \cdot \mathbf{e}_j \tag{15}$$

Plugging this into (13), assuming $h^{(t-1)}$ is constant in I:

$$\left. \frac{\partial J^{(t)}}{\partial I_{i,j}} \right|_{(t)} (\theta) = \left(\hat{y}^{(t)} - y^{(t)} \right)^T U^T \text{diag}\left[h^{(t)} \right] \text{diag}\left[\vec{1} - h^{(t)} \right] e_i^{(t)} \cdot \mathbf{e}_j \tag{16}$$

Recognizing the above form as the outer product:

$$\boxed{\left. \frac{\partial J^{(t)}}{\partial I} \right|_{(t)} (\theta) = e^{(t)} \otimes \left(\left(\hat{y}^{(t)} - y^{(t)} \right)^T U^T \text{diag}\left[h^{(t)} \right] \text{diag}\left[\vec{1} - h^{(t)} \right] \right)^T} \tag{17}$$

2.a.iv $\frac{\partial J^{(t)}}{\partial H}(\theta)$

In a similar fashion to (6), we apply the chain rule:

$$\frac{\partial J^{(t)}}{\partial H_{i,j}}(\theta) = \frac{\partial J^{(t)}}{\partial \hat{y}_k^{(t)}} \frac{\partial \hat{y}_k^{(t)}}{\partial o_l^{(t)}} \frac{\partial o_l^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial H_{i,j}} \tag{18}$$

Note that only the last term in (17) differs from (6), futhermore in a similar fashion to (8):

$$\frac{\partial h^{(t)}}{\partial H_{i,j}} = \frac{\partial \sigma}{\partial a^{(t)}} \left(\frac{\partial}{\partial H_{i,j}} \left({h^{(t-1)}}^T H \right) \right) \tag{19}$$

where

$$\frac{\partial}{\partial H_{i,j}} \left({h^{(t-1)}}^T H \right) = H^T \left(\frac{\partial h^{(t-1)}}{\partial H_{i,j}} \right) \; + \; h_i^{(t-1)} \cdot \mathbf{e}_j \tag{20}$$

Focusing on the t-th appearence of H:

$$\left. \frac{\partial}{\partial H_{i,j}} \right|_{(t)} \left({h^{(t-1)}}^T H \right) = h_i^{(t-1)} \cdot \mathbf{e}_j \tag{21}$$

And in a similar fashion to (17)

$$\boxed{\left. \frac{\partial J^{(t)}}{\partial H} \right|_{(t)} (\theta) = h^{(t-1)} \otimes \left(\left(\hat{y}^{(t)} - y^{(t)} \right)^T U^T \operatorname{diag}\left[h^{(t)} \right] \operatorname{diag}\left[\vec{1} - h^{(t)} \right] \right)^T} \tag{22}$$

2.a.v $\frac{\partial J^{(t)}}{\partial h^{(t-1)}}(\theta)$

In a similar fashion to (6), we apply the chain rule:

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}}(\theta) = \frac{\partial J^{(t)}}{\partial \hat{y}_k^{(t)}} \frac{\partial \hat{y}_k^{(t)}}{\partial o_l^{(t)}} \frac{\partial o_l^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \tag{23}$$

where again only $\frac{\partial h^{(t)}}{\partial h^{(t-1)}}$ is yet to be computed, and is given by:

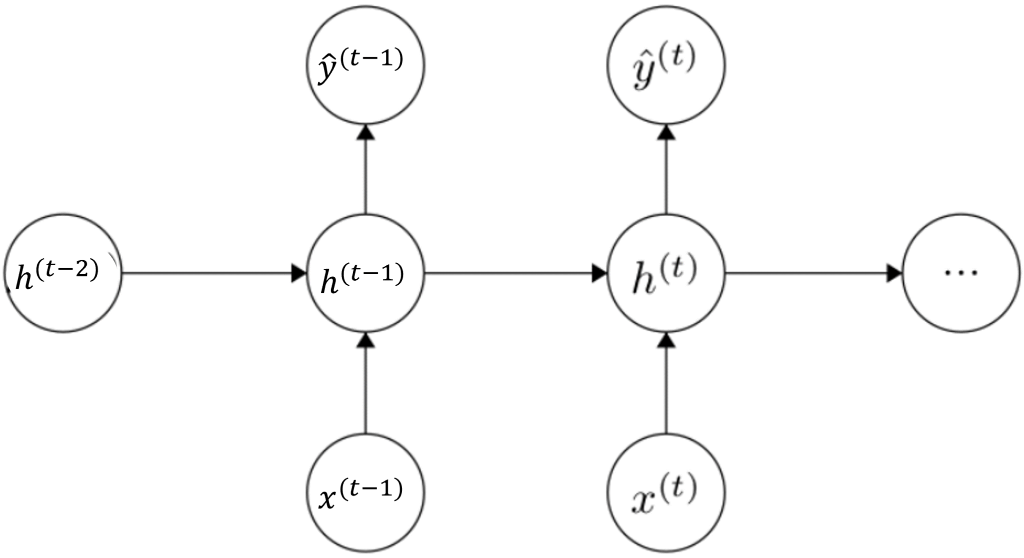
$$\frac{\partial h^{(t)}}{\partial h^{(t-1)}} = H^T \tag{24}$$

Plugging this into (23) yields:

$$\boxed{\frac{\partial J^{(t)}}{\partial h^{(t-1)}}(\theta) = \left(\hat{y}^{(t)} - y^{(t)} \right)^T U^T \operatorname{diag}\left[h^{(t)} \right] \operatorname{diag}\left[\vec{1} - h^{(t)} \right] H^T} \tag{25}$$

2.b

We now backpropagate 1 step through time, expressing out answer using the terms requested in part (a).



2.b.i $\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}}(\theta)$

Looking at equation (8) shows us that:

$$\begin{aligned}\frac{\partial h^{(t)}}{\partial L_{i,j}} &= \frac{\partial \sigma}{\partial a^{(t)}}\left(\frac{\partial}{\partial L_{i,j}}\left(L_{x^{(t)}}I\right) + \frac{\partial}{\partial L_{i,j}}\left({h^{(t-1)}}^{T}H\right)\right) \\ &= \left.\frac{\partial h^{(t)}}{\partial L_{i,j}}\right|_{(t)} + \frac{\partial h^{(t)}}{\partial h^{(t-1)}}\cdot\frac{\partial h^{(t-1)}}{\partial L_{i,j}}\end{aligned} \tag{26}$$

We can unroll one timestamp further into the past $(t-2)$ by applying the chain rule once again:

$$\frac{\partial h^{(t-1)}}{\partial L_{i,j}} = \left.\frac{\partial h^{(t-1)}}{\partial L_{i,j}}\right|_{(t-1)} + \frac{\partial h^{(t-1)}}{\partial h^{(t-2)}}\cdot\frac{\partial h^{(t-2)}}{\partial L_{i,j}} \tag{27}$$

Plugging this expression back into (26) gives:

$$\begin{aligned}\frac{\partial h^{(t)}}{\partial L_{i,j}} &= \left.\frac{\partial h^{(t)}}{\partial L_{i,j}}\right|_{(t)} + \frac{\partial h^{(t)}}{\partial h^{(t-1)}}\left(\left.\frac{\partial h^{(t-1)}}{\partial L_{i,j}}\right|_{(t-1)} + \frac{\partial h^{(t-1)}}{\partial h^{(t-2)}}\cdot\frac{\partial h^{(t-2)}}{\partial L_{i,j}}\right) \\ &= \left.\frac{\partial h^{(t)}}{\partial L_{i,j}}\right|_{(t)} + \frac{\partial h^{(t)}}{\partial h^{(t-1)}}\cdot\left.\frac{\partial h^{(t-1)}}{\partial L_{i,j}}\right|_{(t-1)} + \frac{\partial h^{(t)}}{\partial h^{(t-2)}}\cdot\frac{\partial h^{(t-2)}}{\partial L_{i,j}} \\ &= \sum_{l=t-1}^{t}\frac{\partial h^{(t)}}{\partial h^{(l)}}\cdot\left.\frac{\partial h^{(l)}}{\partial L_{i,j}}\right|_{(l)} + \frac{\partial h^{(t)}}{\partial h^{(t-2)}}\cdot\frac{\partial h^{(t-2)}}{\partial L_{i,j}}\end{aligned} \tag{28}$$

Which using our notation yields:

$$\begin{aligned}\frac{\partial J^{(t)}}{\partial L_{i,j}}(\theta) &= \sum_{l=t-1}^{t}\frac{\partial J^{(t)}}{\partial h^{(l)}}\cdot\left.\frac{\partial h^{(l)}}{\partial L_{i,j}}\right|_{(l)} + \frac{\partial J^{(t)}}{\partial h^{(t-2)}}\cdot\frac{\partial h^{(t-2)}}{\partial L_{i,j}} \\ &= \sum_{l=t-1}^{t}\frac{\partial J^{(t)}}{\partial h^{(l)}}\cdot\left.\frac{\partial h^{(l)}}{\partial L_{i,j}}\right|_{(l)} + \frac{\partial J^{(t-2)}}{\partial L_{i,j}}\end{aligned} \tag{29}$$

Substituting in for $\partial L_{x^{(t-1)}}$:

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}}(\theta) = \sum_{l=t-1}^{t} \frac{\partial J^{(t)}}{\partial h^{(l)}} \cdot \left.\frac{\partial h^{(l)}}{\partial L_{x^{(t-1)}}}\right|_{(l)} + \frac{\partial J^{(t-2)}}{\partial L_{x^{(t-1)}}} \tag{30}$$

where $\left.\frac{\partial h^{(l)}}{\partial L_{x^{(t-1)}}}\right|_{(l)} = 0$ whenever $x^{(l)} \neq x^{(t-1)}$. For simplicity, again assuming $x^{(t_1)} \neq x^{(t_2)}$ for any $t_1 \neq t_2$ gives:

$$\boxed{\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}}(\theta) = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \left.\frac{\partial h^{(t-1)}}{\partial L_{x^{(t-1)}}}\right|_{(t-1)}} \tag{31}$$

Or explicitly:

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}}(\theta) = \left(\hat{y}^{(t)} - y^{(t)}\right)^T U^T \operatorname{diag}\left[h^{(t)}\right] \operatorname{diag}\left[\vec{1} - h^{(t)}\right] H^T \operatorname{diag}\left[h^{(t-1)}\right] \operatorname{diag}\left[\vec{1} - h^{(t-1)}\right] I^T \tag{32}$$

2.b.ii $\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)}$

By applying the same derivations such as in (19):

$$\begin{aligned} \left.\frac{\partial J^{(t)}}{\partial H_{i,j}}\right|_{(t-1)} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \left.\frac{\partial h^{(t-1)}}{\partial H_{i,j}}\right|_{(t-1)} \\ &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \operatorname{diag}\left[h^{(t-1)}\right] \operatorname{diag}\left[\vec{1} - h^{(t-1)}\right] h_i^{(t-2)} \cdot \mathbf{e}_j \end{aligned} \tag{33}$$

which is reconized as the outer product:

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)} = h^{(t-2)} \otimes \left(\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \operatorname{diag}\left[h^{(t-1)}\right] \operatorname{diag}\left[\vec{1} - h^{(t-1)}\right]\right)^T} \tag{34}$$

Or explicitly:

$$\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)} = h^{(t-2)} \otimes \left(\left(\hat{y}^{(t)} - y^{(t)}\right)^T U^T \operatorname{diag}\left[h^{(t)}\right] \operatorname{diag}\left[\vec{1} - h^{(t)}\right] H^T \operatorname{diag}\left[h^{(t-1)}\right] \operatorname{diag}\left[\vec{1} - h^{(t-1)}\right]\right)^T \tag{35}$$

2.b.iii $\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t-1)}$

In a similar fashion to the previous derivation:

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t-1)} = e^{(t-1)} \otimes \left(\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \operatorname{diag}\left[h^{(t-1)}\right] \operatorname{diag}\left[\vec{1} - h^{(t-1)}\right]\right)^T} \tag{36}$$

2.b.iv $\left.\frac{\partial J^{(t)}}{\partial b_1}\right|_{(t-1)}$

$$\begin{aligned}\left.\frac{\partial J^{(t)}}{\partial b_1}\right|_{(t-1)} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}}\cdot\left.\frac{\partial h^{(t-1)}}{\partial b_1}\right|_{(t-1)} \\ &= \boxed{\frac{\partial J^{(t)}}{\partial h^{(t-1)}}\cdot \operatorname{diag}\left[h^{(t-1)}\right]\operatorname{diag}\left[\vec{1}-h^{(t-1)}\right]}\end{aligned} \tag{37}$$

3 Generating Shakespeare Using a Character-level Language Model

3.a

The advantage of character level model stems from the fact that two words can be very similar, for example ”walk” and ”walked”. With a word level model we don’t leverage this similarity and ignore this information which may lead to poorer performance. Character level model can infer that the character sequence ”walk” and the character sequence ”walked” are similar. If we treat words with capital letters as different words, e.g. ”Walk” and ”walk” then it’s even more confusing to word level model. Another advantage of character level models is that they’re more expressive - many times we want to convey in text sounds that aren’t necessarily words, for example we convey thinking in text as ”hmm” or we can convey drunkenness as giberrish of characters, this is hard (or impossible) to do with word level model. The disadvantage of character level model is that it increases the length of the sequence. If we have a document of 100 words, and the average length of a word is for example 4, then with a word level model the sequence length is 100 and with a character level model the length is 400. Plus, with character level model there’s an added difficulty - the model needs to learn what sequences of characters constitute as legitimate words. This is not a problem with word level model that get this information ”for free”.

4 Perplexity

4.a

By using basic power laws, we can rewrite the given expression as follows:

$$\begin{aligned}&2^{-\frac{1}{M}\sum log_2 p(s_i|s_1,...,s_{i-1})} = \\ &= (2^{\sum_i log_2 p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\ &= (\prod_i 2^{log_2 p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\ &= (\prod_i p(s_i|s_1,...,s_{i-1}))^{-\frac{1}{M}} = \\ &= (\prod_i e^{ln p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\ &= e^{-\frac{1}{M}\sum ln p(s_i|s_1,...,s_{i-1})}\end{aligned}$$

4.b

Model/Dataset	Shakespeare	Wikipedia
RNN	6.857	15.361
Bi-Gram	86.552	75.8611

4.c

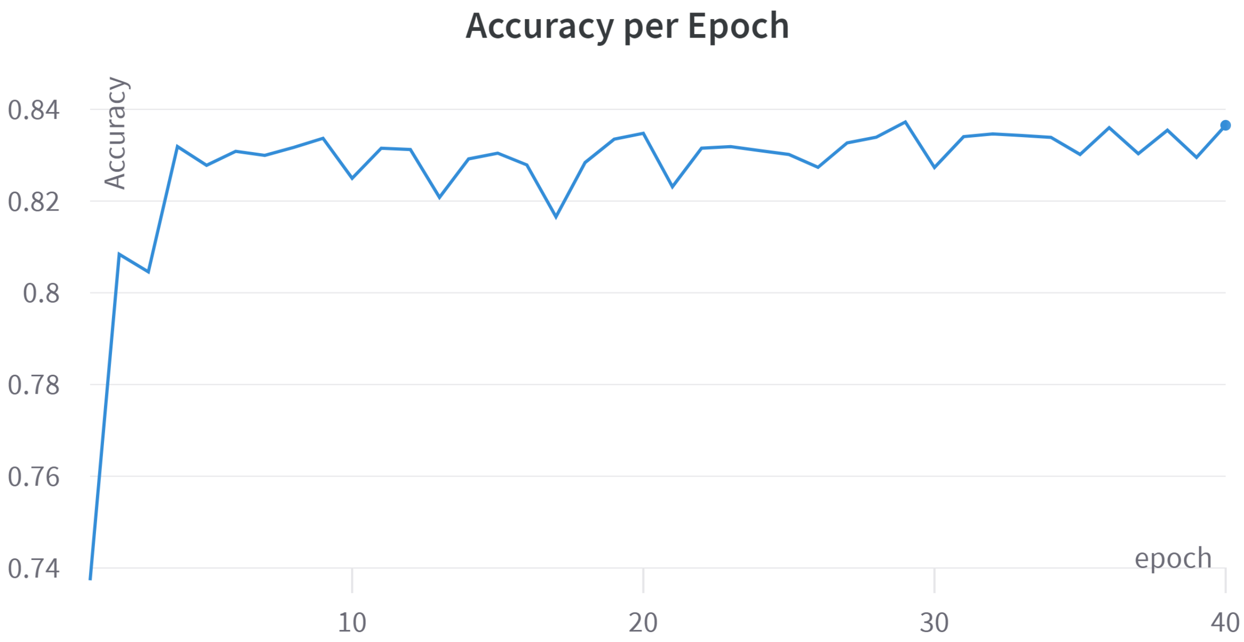
There are two insights that we conclude from the results. First, the RNN seems to be better at predicting the next token than the Bi-Gram model for both of the datasets, as we can see from its much lower perplexity values. Second, the perplexity of the RNN is lower for the Shakespeare dataset than for the Wikipedia dataset. The reason that the RNN is much better than the Bi-Gram model is, in my opinion, that the RNN is charachter level model. That means that in each step it has to predict the next token out of 26 possible tokens (the 26 letters in the English alphabet). On the other hand, the Bi-Gram model is a word level model, and therefore it has to predict the next token out of a much larger vocabulary. Hence, it's not a "fair game", in that regard. Not only the Bi-Gram model a word level token but examining the code we can see that its vocabulary is limited only to 2000 words.

Another reason that may contribute to the superiority of the RNN is that the RNN is not a bigram model, i.e each character prediction depends (theoretically, at least) on all the previous characters, while the Bi-Gram model only depends on the previous character. We've seen in class (and understand intuitively) that the markovian assumption that underlies the bigram model does not hold in reality.

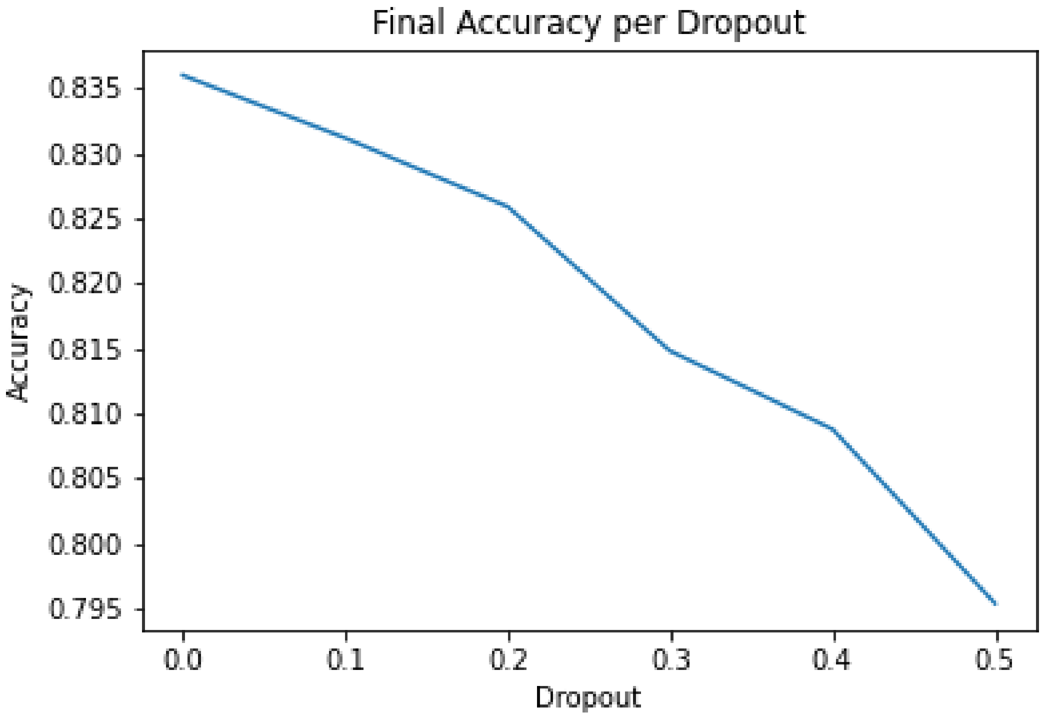
Finally, the reason that the RNN is better for Shakespeare than for Wikipedia is that the RNN was trained on Shakespeare, and therefore it is more familiar with the Shakespeare style of writing.

5 Deep Averaging Networks

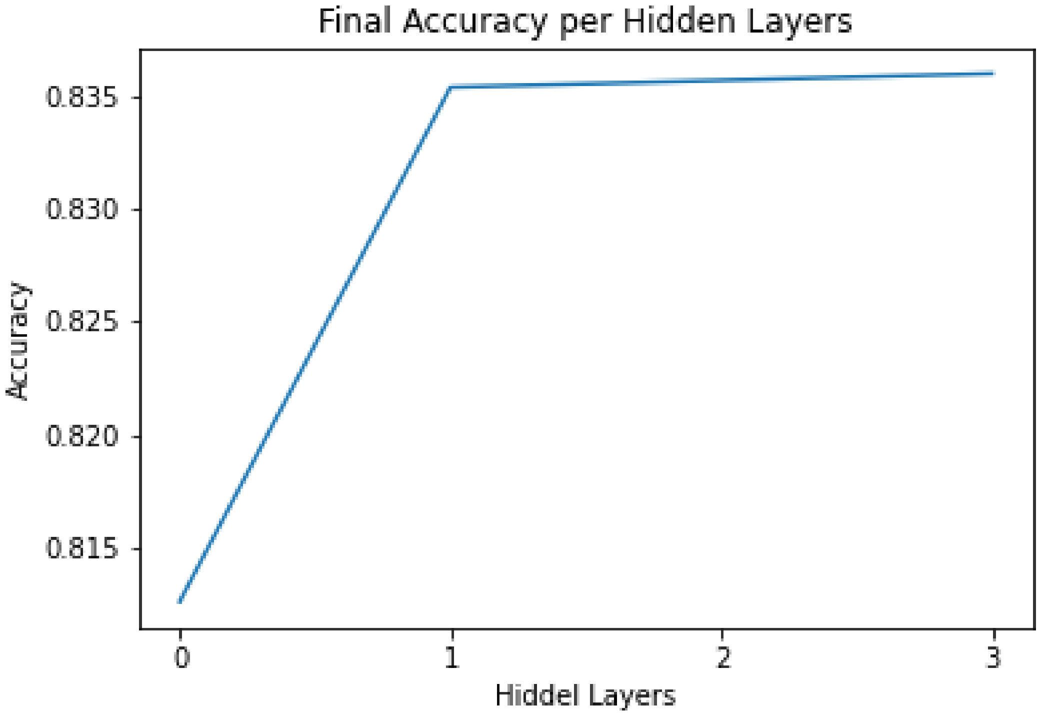
5.a



5.b

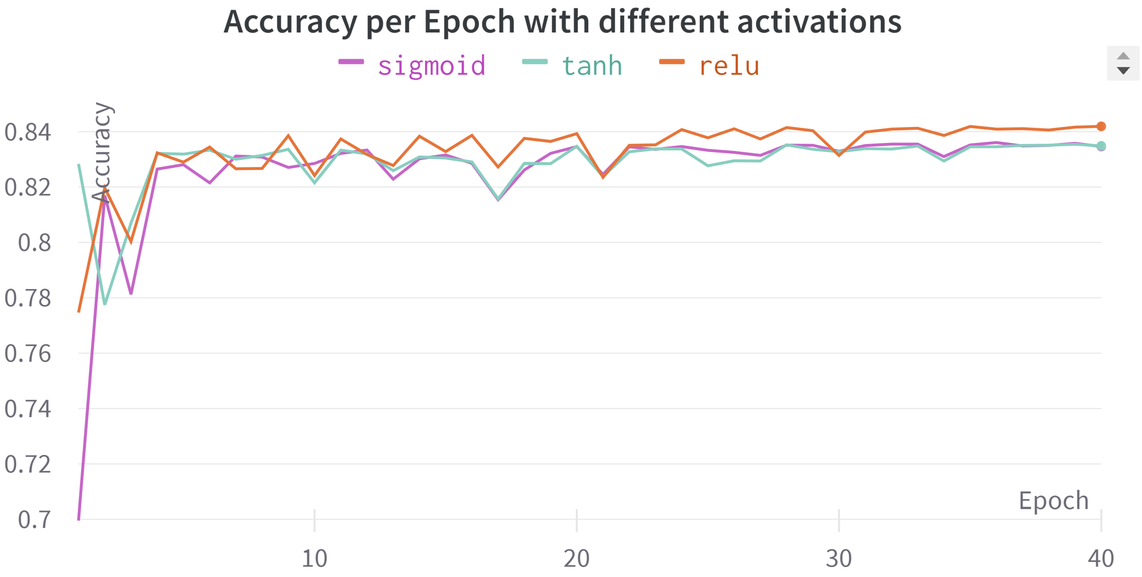


5.c



Overall, adding linear layers yields a slight improvement but in any case it wasn't significant.

5.d



We learned that adding non-linearity to the training can significantly improve the accuracy of the model.

5.e

Example 1: This is the latest entry in the long series of films with the French agent, O.S.S. 117 (the French answer to James Bond). The series was launched in the early 1950's, and spawned at least eight films (none of which was ever released in the U.S.). 'O.S.S.117:Cairo,Nest Of Spies' is a breezy little comedy that should not...repeat NOT, be taken too seriously. Our protagonist finds himself in the middle of a spy chase in Egypt (with Morroco doing stand in for Egypt) to find out about a long lost friend. What follows is the standard James Bond/Inspector Cloussou kind of antics. Although our man is something of an overt xenophobe,sexist,homophobe, it's treated as pure farce (as I said, don't take it too seriously). Although there is a bit of rough language & cartoon violence, it's basically okay for older kids (ages 12 & up). As previously stated in the subject line, just sit back,pass the popcorn & just enjoy.

Explanation: The sentence "breezy little comedy that should not...repeat NOT, be taken too seriously" has negative connotation, it can potentialy make the model classify this example as negative.

Example 2: I was expecting "Born to Kill" to be an exciting, high-tension film noir. Instead, it's got two good action set-pieces (one at the beginning and one at the end) and some marvelously atmospheric cinematography by Robert de Grasse (usually a "glamor" cameraman and a surprising credit for a noir), but the rest of the film is pretty boring. Lawrence Tierney goes through his psycho kick but it's a strictly by-the-numbers performance, mechanically churning out what the audience expected from him after "Dillinger" (an overrated movie but at least better than this). Claire Trevor's character is too stupid and unmotivated to have any audience appeal, and the action (such as it is) stays so resolutely inside that damned house in San Francisco the film becomes claustrophobic instead of genuinely thrilling. It's one of those movies in which the supporting players – notably Elisha Cook, Jr. (whose character's homoerotic itch for Tierney's is one of the few subtleties in an otherwise pretty obvious script) and Isabel Jewell – out-act the leads. It also doesn't help that, nearly half a century after Alfred Hitchcock and Anthony Perkins revolutionized the depiction of psycho killers on the screen in "Psycho," Tierney's is so gross and obvious he might as well have "PSYCHO" tattooed on his forehead. Also, there's no indication in the film as it stands as to why the source novel was called "Deadlier than the Male" – but perhaps James Gunn made the female characters stronger and more interesting than they are in the film. "Born to Kill" is a real disappointment from Robert Wise, who already had some quality movies under his belt and would go on to a stellar career.

Explanation: Although the review is obviously negative, it contains some positive descriptions that might cause a false classification. "exciting", "noir", "good action set-piece", "marvelously atmospheric cinematography", are all examples for the positive words contained in this review.

Example 3: It seems everyone wants to jump on the bandwagon and say "Maha Go Go Go"....The word is MACHA........Like "Mach".....Pronounced maa - ka"...¡br /¿¡br /¿I grew up with this series in the early 70's here in LA on the late and VERY lamented channel 56...Before that there was Tetsuwan Atomu (Astro Boy), dating from 1963 on ol' KHJ TV. Astro Boy was the first TV example of anime we got here in the states...I was into anime as a kid and followed it until the late 80's when, by then it'd become a series of badly animated "talking heads", a phenomenon which has only gotten worse. 'Nuff said.¡br /¿¡br /¿As for "Speed Racer", I really enjoyed the basics there, the POV shots, the cinematic aspects of live action skillfully adopted to animation...That was fairly typical of most Japanese anime back then...Graphics graphics graphics! Take note sometime how obviously the series was inspired by Stanley Kramer's film "Grand Prix" (1966), especially the redone American credits....¡br /¿¡br /¿Oh yeah, I have the original comics from which the series is based, so I know of which I speak.¡br /¿¡br /¿What were we doing animation-wise besides crap like Johnny Quest?.....Th' same ol' stuff we'd been doin' since the 20's....Ho-hum!¡br /¿¡br /¿I guess the real problem I had/have with the way anime was/is shown on American TV is the hatchet job done on the scripts, credits, etc to "sanitize" them for American audiences...I won't go into other programs as we're talking' Speed here.¡br /¿¡br /¿Look at clowns like Peter Fernandez as one of the culprits here, as he was 99% responsible for the re-writes of the series...Not to mention the voice of Speed, Racer X and others...Between him and the goofs at Trans/Lux (Think Felix the Cat and the Mighty Hercules - oy vey!) they took a slick, very sophisticated show and dropped it down to the level of Sesame Street. Think "Cruncher Bloch", The "Forthebird Company", "Skull Duggery"...If I go on I'll puke.¡br /¿¡br /¿This series dates from 40-odd years ago but I, at the time, was keen enough to feel insulted by the dumbing down of this and other Japanese programs...I mean it's obvious when someone's getting' killed but they either remove it or gloss it over........Pleeeeeze!¡br /¿¡br /¿Good show - originally. Sadly all the more recent incarnations of the series have that CRAPPY "made in Korea" look, not to mention being nauseatingly "pc" in content. Even the Japanese outsource their animation now..¡br /¿¡br /¿Try watchin' the original Japanese opening on YouTube sometime...It sends chills up my spine.....If only......Oh well. Robert

Explanation: The review is too longs and contains a lot of information regarding unrelated anime films. Combination of long text together with many walk-arounds before getting the point can confuse our model.

Example 4: Recap: Since the warrior queen Gedren raised and slaughtered most of Sonja's family, she has trained in the art of sword fighting. Now, Gedren has taken a very powerful talisman, that threatens to destroy the world if not destroyed, killing Sonja's sister in the process. Now Sonja is out for revenge, and to save the world. Along the way, she meets the very Conan-like (but not Conan, no!) Kalidor, the child-prince Tarn and his bodyguard Falcon. At first Sonja declines all help, but is later forced to accept it, and together they go to save the world.¡br /¿¡br /¿Comments: When you watch a movie like this, and you think that it is the story that the is the best element in this movie, the movie is in big trouble. Because 1) a movie like this should draw its strength upon good swordfights and effects, and 2) the story is really, really bad. It is simple, and uncomplicated and really offers nothing in way of character development or even suspense. It is predictable and boring, and the obvious couple, Sonja and Kalidor, has no chemistry at all. And the kid is just annoying. And most of the scenes is drawn out so long that they become boring. Though the movie is not very long, it has not material enough to fill its time. And so back to point 1). The fighting is slow, uneventful and really bad. It clearly shows most fighters clearly blocking the opponents strokes far ahead of the opponent has even begun to strike. In my honest opinion, I believe most kids, fighting with sticks, creates more exciting fights playing knights than this movie did. All in all, this is a really bad spin-off, that should be avoided by all who liked the Conan-movies.¡br /¿¡br /¿2/10

Explanation: While the review itself is negative, it begins with a long recap of the storyline that contains positive words that can confuse the model. For example: "art of sword fighting", "very powerful". In addition, the sentence "the story is the best element" that is written later, might confuse the model.

Example 5: Not as bad as some people say...This is a unofficial Bond movie and a remake of "Thunderball", written by Kevin McClory (co- producer in "Thunderball"). Well, the cast is very very interesting, Maria Brandauer is a great Bond- villain, Kim Basinger and Barbara Carrera are just like the "original" Bond- girls, plus Rowan Atkinson and a truly great Edward Fox, who looks really refreshing in the "M" role. In fact, the whole movie is refreshing and gives some new impulses. Sean Connery does it once more confident and charming, except that he looks a little bit too old. But alright, he is the original Bond and it was great to see him once more in this role. The locations are also typical- Bahamas, France, etc. The only thing that really fails is the music score, the song "Never say never again" is O.K., but the theme song is just missing. All in one, a nice try to make a difference from the comic and silly Roger Moore movies like "Moonraker". Only if there was another story, "Thunderball" was a excellent movie and really did not needed a remake

Explanation: The review itself is ambiguous and shows some both negative and positive aspects of the film.

6 Right-to-left vs left-to-right Estimation

6.a

Let's denote the left-to-right count-based bi-gram model as L and the right-to-left count-based bi-gram model as R. The probability estimates for each model are as follows:

$L(w|v) = \frac{C_L(w,v)}{C(v)}$ - Probability of w given v in the left-to-right model.

$R(w|v) = \frac{C_R(w,v)}{C(v)}$ - Probability of w given v in the right-to-left model.

Where $C_L(w,v)$ and $C_R(w,v)$ is counting the bi-gram sequence from left/right respectively.

Notice that by definition $C_L(w,v) = C_R(v,w)$

Also:

$L(x_i|x_{i-1}) = \frac{C_L(x_i,x_{i-1})}{C(x_{i-1})} = \frac{C_R(x_{i-1},x_i)}{C(x_{i-1})} = \frac{R(x_{i-1}|x_i)}{C(x_{i-1})}C(x_i)$

Estimating the bi-gram from left to right for a corpus of size M gives:

$P(x_0x_1...x_n) = P(x_0)\Pi_{i=1,...,n}L(x_i|x_{i-1}) = \frac{C(x_0)}{M}\prod_{i=1}^{n} L(x_i|x_{i-1}) = \frac{C(x_0)}{M}\prod_{i=1}^{n} \frac{R(x_{i-1}|x_i)}{C(x_{i-1})}C(x_i)$

Notice that we got a telescopic product which after cancelling elements equals:

$\frac{C(x_0)}{M}\frac{C(x_n)}{C(x_0)}\prod_{i=1}^{n} R(x_{i-1}|x_i) = \frac{C(x_n)}{M}\prod_{i=1}^{n} R(x_{i-1}|x_i)$

This is exactly estimating the bi-gram model with right to left count, which is what we wanted to prove.