Assignment 2

1 Question 2

1.1 a

y is a one–hot vector, thus there exists a k s.t $y_j = 0$ for $j \neq k$ and $y_k = 1$.

$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\sum_{i=1}^{|V|} y_i^{(t)} \log(\hat{y_i}^{(t)}) = -\log(\hat{y_k}^{(t)})$$

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial U}$$

$$\frac{\partial J}{\partial \hat{y}^{(t)}} = -\frac{\frac{\partial \hat{y_k}}{\partial \hat{y}^{(t)}} \hat{y_k}}{\hat{y_k}^{(t)}} = -\frac{1}{\hat{y_k}^{(t)}} = -\frac{y^{(t)}}{\hat{y}^{(t)}}$$

$$\hat{y_i}^{(t)} = softmax\left(h^{(t)}U + b_2\right)_i = \frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}$$

$$\frac{\partial \hat{y_i}^{(t)}}{\partial U} = \frac{h^{(t)} e^{(h^{(t)}U+b_2)_i} \cdot \sum_j e^{(h^{(t)}U+b_2)_j} - e^{(h^{(t)}U+b_2)_i} \cdot h^{(t)} e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= \frac{h^{(t)} e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \frac{e^{(h^{(t)}U+b_2)_i} \cdot h^{(t)} e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= h^{(t)} \left(\frac{e^{(h^{(t)}U_i+b_2)}}{\sum_j e^{(h^{(t)}U_j+b_2)}} - \frac{\left(e^{(h^{(t)}U+b_2)_i}\right)^2}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} \right) = h^{(t)} \left(\hat{y_i}^{(t)} - \left(\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} \right)^2 \right) = \hat{y_i}^{(t)} h^{(t)} \left(1 - \hat{y_i}^{(t)}\right)$$

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial U} = -\hat{y}^{(t)} \frac{y^{(t)}}{\hat{y}^{(t)}} h^{(t)} \left(1 - \hat{y}^{(t)}\right) = y^{(t)} h^{(t)} \left(\hat{y}^{(t)} - 1\right)$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial \mathfrak{e}^{(t)}} \frac{\partial \mathfrak{e}^{(t)}}{\partial L_{x^{(t)}}}$$

$$\frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} = -\frac{y^{(t)}}{\hat{y}^{(t)}}$$

$$\hat{y_i}^{(t)} = softmax\left(h^{(t)}U + b_2\right)_i = \frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}$$

$$\frac{\partial \hat{y_i}^{(t)}}{\partial h_i^{(t)}} = \frac{U \cdot e^{(h^{(t)}U+b_2)_i} \cdot \sum_j e^{(h^{(t)}U+b_2)_j} - U \cdot e^{(h^{(t)}U+b_2)_i} \cdot e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= \frac{U \cdot e^{(h^{(t)}U+b_2)_i} \cdot \left(\sum_j e^{(h^{(t)}U+b_2)_j} - e^{(h^{(t)}U+b_2)_i}\right)}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= U \cdot \left[\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \frac{\left(e^{(h^{(t)}U+b_2)_i}\right)^2}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2}\right] =$$

$$= U \cdot \left[\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \left(\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}\right)^2\right] =$$

$$= U \cdot \left[\hat{y}_i^{(t)} - \left(\hat{y}_i^{(t)}\right)^2\right]$$

$$h^{(t)} = sigmoid\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right) = \frac{1}{1 + e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}}$$

$$\frac{\partial h^{(t)}}{\partial \mathfrak{e}^{(t)}} = \frac{I \cdot e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}}{\left(1 + e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}\right)^2}$$

$$\mathfrak{e}^{(t)} = x^{(t)}L$$

L is the embedding matrix, $L_{x^{(t)}}$ is the row of L corresponding to the current input word

$$\frac{\partial \mathfrak{e}^{(t)}}{\partial L_{x^{(t)}}} = x^{(t)} \cdot L_{x^{(t)}}$$

2 Question 3

2.1 a

One advantage of a character-based language model over a word-based language model could be learning more delicate rules of language and, which enables a more detailed analysis of linguistic patterns such as rare words and different spelling.

A word-based language model, on the other hand, can reduce the number of possibilities for sequences it has to generate as well as the size of the vocabulary it must consider, thus reducing computational complexity compared to character-based models.

2.2 b

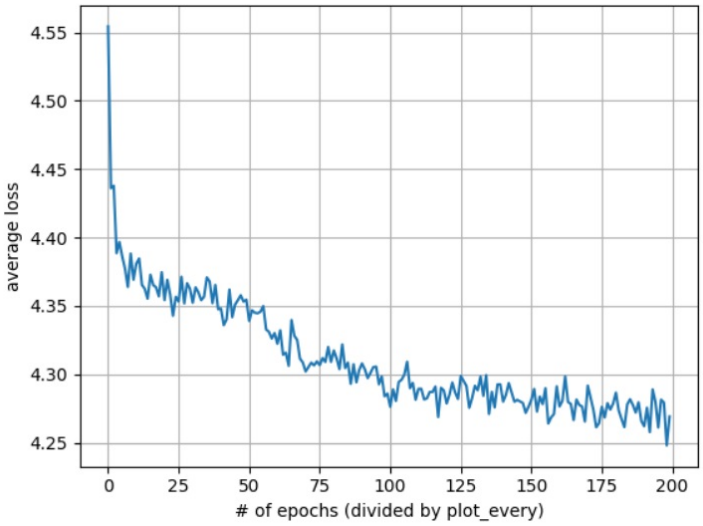

Figure :1 Training Loss

3 Question 4

3.1 a

$$2^{-\frac{1}{M}\sum_{i=1}^{M}\log_2 p(s_i|s_1,\ldots,s_{i-1})} \underbrace{=}_{base\ change} 2^{-\frac{1}{M}\sum_{i=1}^{M}\frac{\log_e p(s_i|s_1,\ldots,s_{i-1})}{\log_e 2}} = 2^{-\frac{1}{M}\sum_{i=1}^{M}\frac{\ln p(s_i|s_1,\ldots,s_{i-1})}{\ln 2}} =$$

$$= 2^{-\frac{1}{M\cdot\ln 2}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})} =$$

$$\underbrace{=}_{powers\ rules} \left(2^{\log_2 e}\right)^{\left(-\frac{1}{M}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})\right)} = e^{-\frac{1}{M}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})}$$

4 Question 6

$$P(x_0x_1\ldots x_n) = P(x_n)P(x_{n-1}|x_n)\cdot\ldots\cdot P(x_0|x_1) =$$

$$\underbrace{=}_{bayes\ rule} P(x_n)\frac{P(x_n|x_{n-1})P(x_{n-1})}{P(x_n)}\cdot\frac{P(x_{n-1}|x_{n-2})P(x_{n-2})}{P(x_{n-1})}\cdot\ldots\cdot\frac{P(x_1|x_0)P(x_0)}{P(x_1)} =$$

$$= P(x_n|x_{n-1})\cdot P(x_{n-1}|x_{n-2})\cdot\ldots\cdot P(x_1|x_0)P(x_0) =$$

$$\underbrace{=}_{Commutative\ property} P(x_0)P(x_1|x_0)\ldots P(x_n|x_{n-1})$$