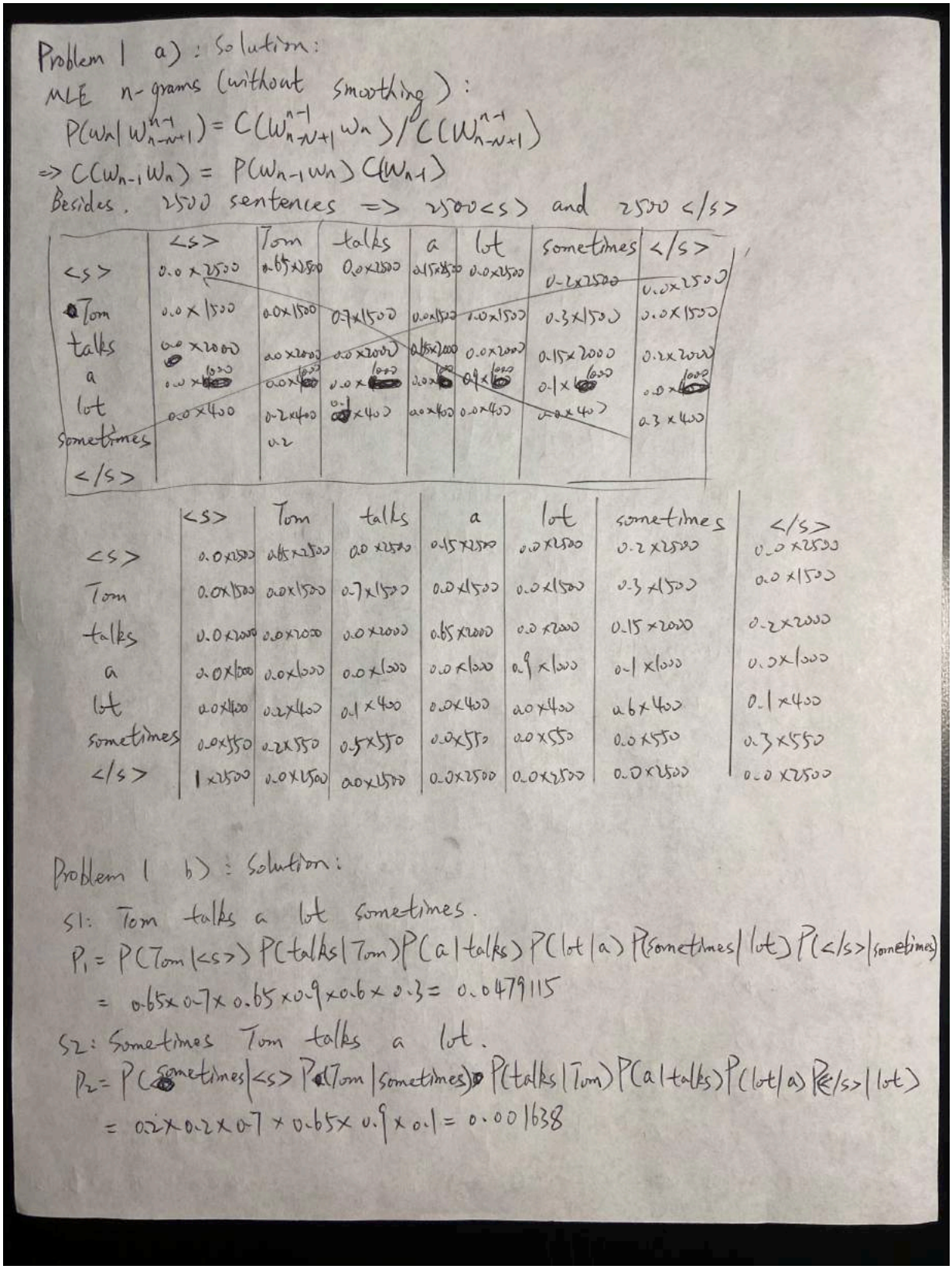
NLP Midterm



Problem 1 a) : Solution:

MLE n-grams (without smoothing):

$P(w_n | w_{n-N+1}^{n-1}) = C(w_{n-N+1}^{n-1} w_n) / C(w_{n-N+1}^{n-1})$

$\Rightarrow C(w_{n-1} w_n) = P(w_{n-1} w_n) C(w_{n-1})$

Besides, 2500 sentences $\Rightarrow$ 2500 <s> and 2500 </s>

	<s>	Tom	talks	a	lot	sometimes	</s>
<s>	0.0 x 2500	0.65 x 2500	0.0 x 2500	0.15 x 2500	0.0 x 2500	0.2 x 2500	0.0 x 2500
Tom	0.0 x 1500	0.0 x 1500	0.7 x 1500	0.0 x 1500	0.0 x 1500	0.3 x 1500	0.0 x 1500
talks	0.0 x 2000	0.0 x 2000	0.0 x 2000	0.65 x 2000	0.0 x 2000	0.15 x 2000	0.2 x 2000
a	0.0 x 1000	0.0 x 1000	0.0 x 1000	0.0 x 1000	0.9 x 1000	0.1 x 1000	0.0 x 1000
lot	0.0 x 400	0.2 x 400	0.1 x 400	0.0 x 400	0.0 x 400	0.0 x 400	0.3 x 400
sometimes		0.2					
</s>							

	<s>	Tom	talks	a	lot	sometimes	</s>
<s>	0.0 x 2500	0.65 x 2500	0.0 x 2500	0.15 x 2500	0.0 x 2500	0.2 x 2500	0.0 x 2500
Tom	0.0 x 1500	0.0 x 1500	0.7 x 1500	0.0 x 1500	0.0 x 1500	0.3 x 1500	0.0 x 1500
talks	0.0 x 2000	0.0 x 2000	0.0 x 2000	0.65 x 2000	0.0 x 2000	0.15 x 2000	0.2 x 2000
a	0.0 x 1000	0.0 x 1000	0.0 x 1000	0.0 x 1000	0.9 x 1000	0.1 x 1000	0.0 x 1000
lot	0.0 x 400	0.2 x 400	0.1 x 400	0.0 x 400	0.0 x 400	0.6 x 400	0.1 x 400
sometimes	0.0 x 550	0.2 x 550	0.5 x 550	0.0 x 550	0.0 x 550	0.0 x 550	0.3 x 550
</s>	1 x 2500	0.0 x 1500	0.0 x 1500	0.0 x 2500	0.0 x 2500	0.0 x 2500	0.0 x 2500

Problem 1 b) : Solution:

S1: Tom talks a lot sometimes.

$P_1 = P(\text{Tom} | \langle s \rangle) P(\text{talks} | \text{Tom}) P(\text{a} | \text{talks}) P(\text{lot} | \text{a}) P(\text{sometimes} | \text{lot}) P(\langle /s \rangle | \text{sometimes})$

$= 0.65 \times 0.7 \times 0.65 \times 0.9 \times 0.6 \times 0.3 = 0.0479115$

S2: Sometimes Tom talks a lot.

$P_2 = P(\text{sometimes} | \langle s \rangle) P(\text{Tom} | \text{sometimes}) P(\text{talks} | \text{Tom}) P(\text{a} | \text{talks}) P(\text{lot} | \text{a}) P(\langle /s \rangle | \text{lot})$

$= 0.2 \times 0.2 \times 0.7 \times 0.65 \times 0.9 \times 0.1 = 0.001638$

Problem 1 c): Solution:

For Bigram, $PP(w) = P(w_1 \cdots w_N)^{-\frac{1}{N}} = \left[\prod_{\text{for}} \frac{1}{P(w_i|w_{i-1})}\right]^{\frac{1}{N}}$

TEST: <s> Tom talks home a lot </s> <s> Home Tom talks </s>
w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10} w_{11} w_{12}

<s> Home talks </s>
w_{13} w_{14} w_{15} w_{16} $N=16$

merge replicated items

$$\therefore PP(w) = \left[\frac{1}{P(\text{Tom}|\langle s\rangle)} \cdot \frac{1}{P(\text{talks}|\text{Tom})} \cdot \frac{1}{P(\text{home}|\text{talks})} \cdot \frac{1}{P(\text{a}|\text{home})} \cdot \frac{1}{P(\text{lot}|\text{a})} \cdot \frac{1}{P(\langle/s\rangle|\text{lot})} \cdot \frac{1}{P(\langle s\rangle|\langle/s\rangle)} \cdot \frac{1}{P(\langle s\rangle|\langle/s\rangle)} \cdot \frac{1}{P(\text{home}|\langle s\rangle)} \cdot \frac{1}{P(\text{Tom}|\text{home})} \cdot \frac{1}{P(\text{talks}|\text{Tom})} \cdot \frac{1}{P(\langle/s\rangle|\text{talks})} \cdot \frac{1}{P(\langle s\rangle|\langle/s\rangle)} \cdot \frac{1}{P(\text{home}|\langle s\rangle)} \cdot \frac{1}{P(\text{talks}|\text{home})} \cdot \frac{1}{P(\langle/s\rangle|\text{talks})}\right]^{\frac{1}{20}}$$

We can get all no smoothing probabilities from given table. But we need to calculate laplace smoothing probabilities. Fortunately, we can merge some replicated items. Besides, $|V| = 25$

$P_{Laplace}(w_i) = \frac{(c_i+1)}{N+V}$, $P_{MLE}(w_i) = \frac{c_i}{N}$. (Use P_L short for $P_{Laplace}$ and P_M short for P_{MLE} below)

$\therefore P_L(w_i) = \frac{P_M(w_i)N+1}{N+V}$

$P_L(\text{Tom}|\langle s\rangle) = \frac{0.7\times 2500+1}{2500+25} = 0.693465$ $P_L(\text{talks}|\text{Tom}) = \frac{0.75\times 1500+1}{1500+25} = 0.738361$

$P_L(\text{home}|\text{talks}) = \frac{0.1\times 2000+1}{2000+25} = 0.099259$ $P_L(\text{a}|\text{home}) = \frac{0.1\times 15000+1}{15000+25} = 0.09900$

$P_L(\text{lot}|\text{a}) = \frac{0.3\times 1000+1}{1000+25} = 0.293658$ $P_L(\langle/s\rangle|\text{lot}) = \frac{0.05\times 400+1}{400+25} = 0.049412$

$P_L(\langle s\rangle|\langle/s\rangle) = \frac{1.0\times 2500+1}{2500+25} = 0.990495$ $P_L(\text{home}|\langle s\rangle) = \frac{0.1\times 2500+1}{2500+25} = 0.099406$

$P_L(\text{Tom}|\text{home}) = \frac{0.5\times 1500+1}{15000+25} = 0.499234$ $P_L(\text{talks}|\text{Tom}) = \frac{0.75\times 1500+1}{1500+25} = 0.738361$

$P_L(\langle/s\rangle|\text{talks}) = \frac{0.05\times 2000+1}{2000+25} = 0.049876$ $P_L(\langle s\rangle|\langle/s\rangle) = 0.990495$

$P_L(\text{home}|\langle s\rangle) = 0.099406$ $P_L(\text{talks}|\text{home}) = \frac{0\times 15000+1}{15000+25} = 0.000066$

$P(\langle/s\rangle|\text{talks}) = 0.049876$

$\therefore PP_{Laplace}(w) = [0.693465\times 0.738361\times 0.099259\times 0.09900\times 0.293658\times 0.049412\times 0.990495\cdot 0.099406\times 0.499234\times 0.738361\times 0.049876\times 0.990495\times 0.099406\times 0.000066\times 0.049876]^{\frac{1}{16}} = 6.843318$

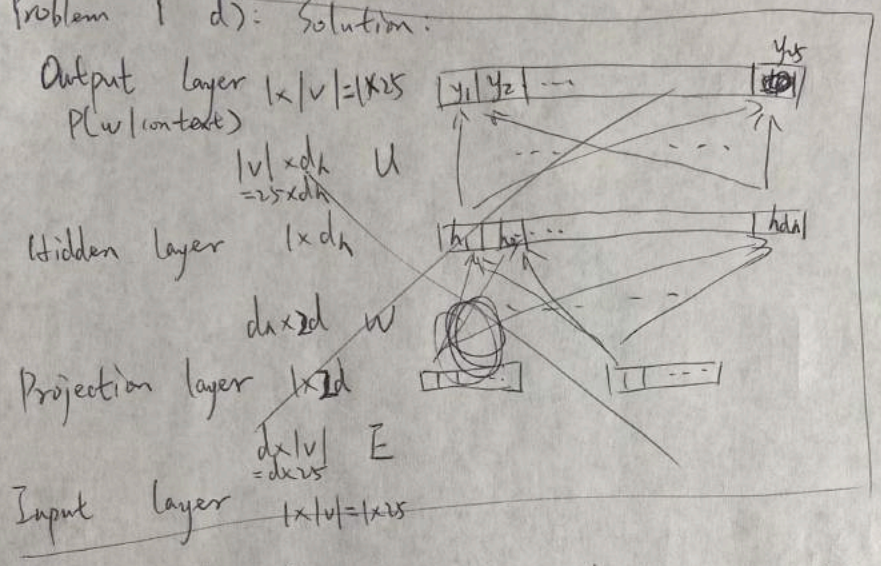
Since $P_{MLE}(\text{talks} \mid \text{home}) = 0$ [connect to Problem 1 c)]

$PP_{\text{no-smoothing}}(w) = \left[\frac{1}{0}\right]^{\frac{1}{n}} = \infty$

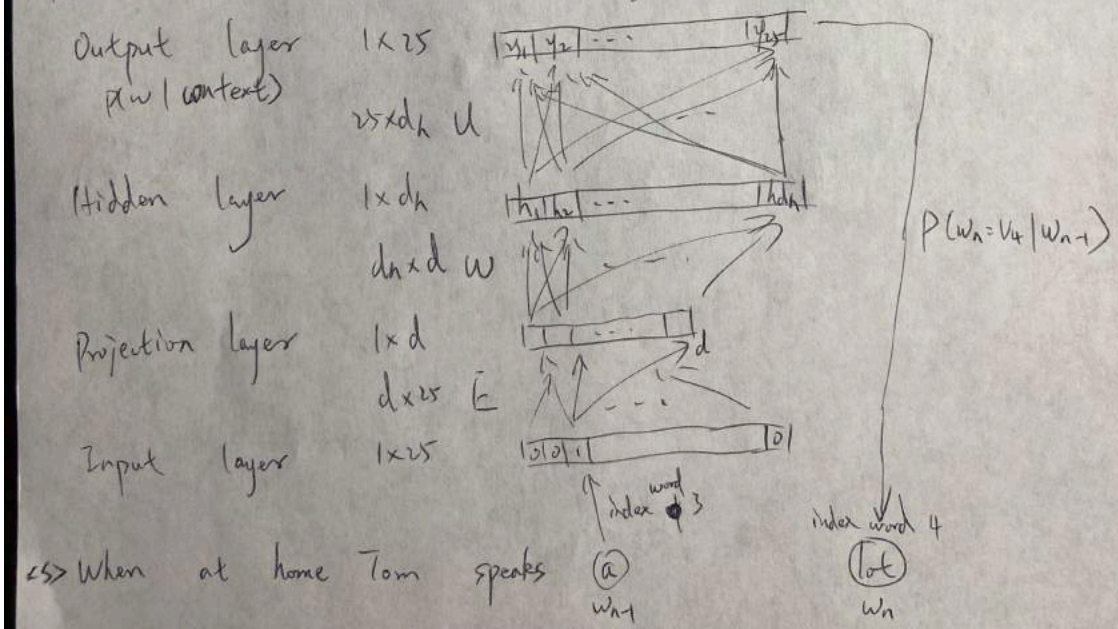
$\therefore PP_{\text{Laplace}}(w) = 6.843318$ $\quad PP_{\text{no-smoothing}}(w) = \infty$

That is why we apply Laplace smoothing for calculating perplexity

Problem 1 d): Solution:



Problem 1 d): Solution: $|v| = 25$ bigram: $N = 2 - 1 = 1$



[Connect to Problem 1 d)]

$e_i = Ex_i$

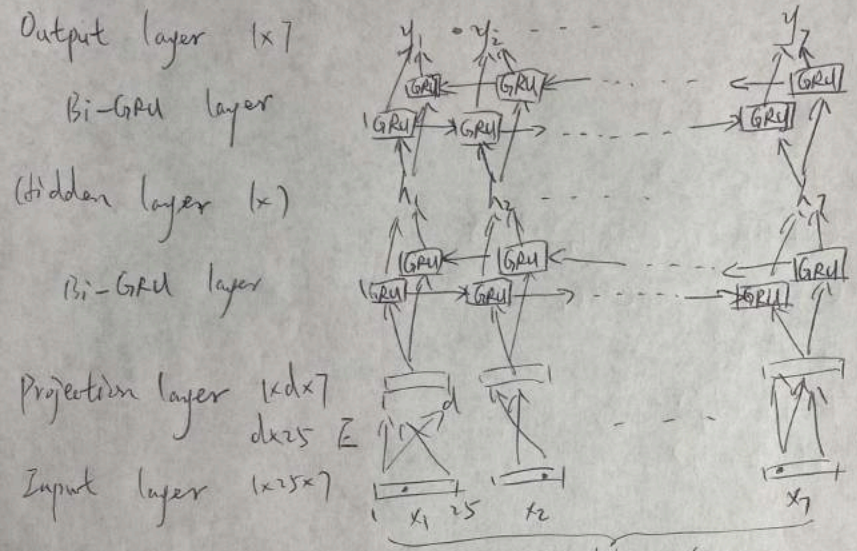
$h = \sigma(We + b)$

$z = Uh$

$y = \text{softmax}(z)$

} Equations

Problem 1 e: Solution:



Teacher Forcing:

$L_{CE}(\hat{y}^t, y^t) = -\log \hat{y}^t_{w_{t+1}}$

The correct distribution y comes from knowing the next word.

At time t, the CE loss is the negative log probability assigned to the next word in the training sequence.

[Connects to Question 1 e]

$x_i = p_c \times$ Wembedding

$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$

$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$

$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$

$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$

$h_t = w^t h_m^t + w^b h_o^b$

~~$h_{t_2} = \sigma(w h_t + t$~~

$r_{t_2} = \sigma(W_{r_2} h_{t_2} + U_r h_{t-1_2} + b_{r_2})$

$z_{t_2} = \sigma(W_{z_2} h_{t_2} + U_z h_{t-1_2} + b_{z_2})$

$\hat{h}_{t_2} = \tanh(W_{h_2} h_t + U_{h_2}(r_{t_2} \odot h_{t-1_2}) + b_{h_2})$

$h_{t_2} = (1 - z_{t_2}) \odot h_{t-1_2} + z_{t_2} \odot \hat{h}_{t_2}$

$e_t = w^t_2 h_m^t{}_2 + w^b_2 h_o^b{}_2$

$P(y = j \mid e) = \frac{\exp(w_j^T e + b_j)}{\sum_{j=1}^{k} \exp(w_j^T e + b_j)}$

Problem 2. a): Solution:

S1: Iceland/NNP volcano/NN remains/VBZ hazardous/JJ after/IN eruption/NN near/IN Reykjavik/NNP.

S2: Concerns/NNS were/VBD raised/VBN about/IN the/DT proximity/NN of/IN the/DT volcano/NN to/IN the/DT country/NN 's/POS main/JJ airport/NN. Keflavik/NNP International/NNP Airport/NNP, which/WDT is/VBZ just/RB a/DT 25/CD min/NN car/NN ride/NN from/IN the/DT peninsula/NN.

Question 2 b): Solution:

First, words are transformed into hot vectors by word embeddings and characters embeddings.
Then, with bi-directional LSTM layers. we did not compute the probability for each tag at each step. Instead, we used log-linear functions to get a global probability for the whole sequence.

Possible tag sequences : $\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|w)$

$$P(Y|X) = \frac{\exp\left(\sum_{k=1}^{k} w_k f_k(X,Y)\right)}{\sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{k} w_k f_k(X,Y')\right)}$$

current output token y_i
previous output token y_{i-1}
input string X
current position i

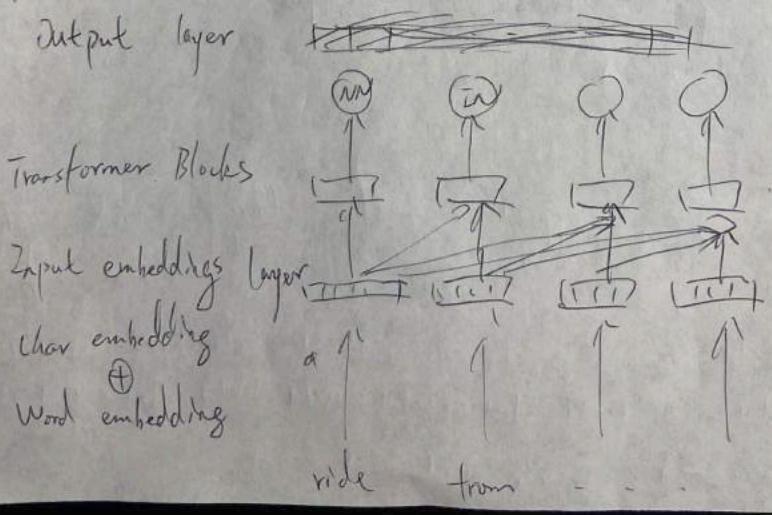
Linear Chain CRF:

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

Then, we have:

$$\hat{Y} = \underset{Y \in \mathcal{Y}^n}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{k} w_k f_k(y_{i-1}, y_i, X, i)$$

Problem 2 c): Solution:



[connect to Problem 2 c]

$Q = W^Q X$

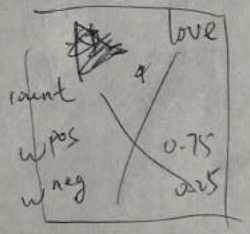
$K = W^K X$

$V = W^V X$

$\text{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Use char embedding and word embedding to firstly transform word into ~~bit vectors~~ embedding vectors for input embedding layer.

Problem 3 a): Solution



change	hate	love	opposite	suffer	thrive	~~want~~	wish
1	IIII	IIII	1	II	1	1	1

$f^+ = 0.66\times1 + 0.22\times4 + 0.75\times4 + 0.27\times1 + 0.25\times2 + 0.88\times1 + 0.69\times1 + 0.55\times1 = 7.43$

$f^- = 0.34\times1 + 0.78\times4 + 0.25\times4 + 0.63\times1 + 0.75\times2 + 0.12\times1 + 0.31\times1 + 0.45\times1 = 7.47$

~~sentiment~~ $\frac{f^+}{f^-} = 0.994645 > 0.55 = T$

$\therefore$ sentiment = +

b): Solution

	change	hate	love	opposite	suffer	thrive	~~want~~	wish
remain same	1	3	3	1	2	1	1	1
Swap value	0	1	1	0	0	0	0	0

$f^+ = 0.66\times1 + 0.22\times3 + 0.78\times1 + 0.75\times3 + 0.25\times1 + 0.27\times1 + 0.25\times2 + 0.88\times1 + 0.69\times1 + 0.55\times1 = 7.49$

$f^- = 0.34\times1 + 0.78\times3 + 0.22\times1 + 0.25\times3 + 0.75\times1 + 0.63\times1 + 0.75\times2 + 0.12\times1 + 0.31\times1 + 0.45\times1 = 7.41$

$f^+/f^- = 1.010796 > 0.55 = T$ $\therefore$ sentiment = +