Section 2

eq 1)
$$P(O=o \mid C=c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in W} \exp(u_w^T v_c)}$$

a)
Let $x \in \mathbb{R}^n$ be an input vector and $c \in \mathbb{R}$ a constant

We will show

$$\forall\, 1 \le i \le n \qquad \frac{\exp(x_i)}{\sum_j \exp(x_j)} = \frac{\exp(x_i + c)}{\sum_j \exp(x_j + c)}$$

$$\Updownarrow$$

$$\frac{\exp(x_i)}{\exp(x_i + c)} = \frac{\sum_{j \in W} \exp(x_j)}{\sum_{j \in W} \exp(x_j + c)}$$

$$\Updownarrow$$

$$\exp(-c) = \sum_{j \in W} \frac{\exp(x_j)}{\sum_{k \in W} \exp(x_k + c)} = \sum_{j \in W} \frac{\exp(x_j)}{\sum_{k \in W} \exp(x_j)\exp(c)}$$

$$= \sum_{j \in W} \frac{\exp(x_j)}{\exp(c) \sum_{k \in W} \exp(x_k)} = \frac{1}{\exp(c)} \sum_{j \in W} \frac{\exp(x_j)}{\exp(x_j)}$$

$$= \frac{1}{\exp(c)} = \exp(-c)$$

b) eq 2): $J_{NS}(v_c, o, U) = -\log P(O=o \mid C=c)$

we will show that $-\sum_{w \in W} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

$$-\sum_{w \in W} y_w \log(\hat{y}_w) \underset{\substack{\uparrow \\ \forall w \neq o:\ y_w = 0 \\ w = o:\ y_w = 1}}{=} -y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$$

c)

$$J_{NS} = -\log(\hat{y}_o) = -\log(P(O=o|C=c))$$

$$= -\log\left(\frac{\exp(u_o^T v_o)}{\sum_{w\in W} \exp(u_w^T v_c)}\right) = -u_o^T v_c + \log\left(\sum_{w\in W} \exp(u_w^T v_c)\right)$$

$$\frac{\partial(J_{NS}(v_c,o,U))_i}{\partial(v_c)_i} = -(u_o^T)_i + \sum_{w\in W}\left(\frac{(u_w^T)_i \exp((u_w^T)_i (v_c)_i)}{\sum_{\bar{w}\in W} \exp((u_{\bar{w}}^T)_i (v_c)_i)}\right)$$

$$= (-u_o^T)_i + \sum_{w\in W}(\hat{y}_w u_w)_i \Rightarrow \frac{\partial J_{NS}}{\partial v_c} = \sum_{w\in W} \hat{y}_w u_w - u_o^T$$

Notice! $u_o^T = Uy$, $\sum \hat{y}_w u_w = U\cdot\hat{y}$

$$\Downarrow$$

$$\frac{\partial J_{NS}}{\partial v_c} = U(\hat{y}-y)$$

d) Assumming $o=w$:

$$\frac{\partial J_{NS}}{\partial u_w} = -v_c + \frac{v_c \exp(u_w^T v_c)}{\sum_{\bar{w}\in W} \exp(u_{\bar{w}} v_c)} = -v_c + v_c \hat{y}_w = v_c(\hat{y}_w - 1)$$

Assumming $o \neq w$:

$$\frac{\partial J_{NS}}{\partial u_w} = 0 + \frac{v_c \exp(u_w^T v_c)}{\sum_{\bar{w}\in W} u_{\bar{w}} v_c)} = v_c \cdot \hat{y}_w$$

e)

$$\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{\exp(x)+1}$$

$$\frac{\partial \sigma}{\partial x} = \frac{0-(-1\exp(-x))}{1+2\exp(-x)+\exp(-2x)} = \exp(-x)\left(1+\exp(-x)\right)^{-2}$$

$$= \frac{\exp(-x)}{(\exp(-x)+1)(\exp(-x)+1)} = \sigma(x)\ \frac{\exp(-x)}{\exp(-x)+1}$$

$$= \sigma(x)\left(\frac{\exp(-x)+1}{\exp(-x)+1} - \frac{1}{\exp(-x)+1}\right) = \sigma(x)\left(1-\sigma(x)\right)$$

f)

$$J_{Neg}(V_c, o, U) = -\log(\sigma(u_o^T V_c)) - \sum_{k=1}^{k} \log(\sigma(-u_k^T V_c))$$

$$\frac{\partial J_{Neg}}{\partial V_c} = -\frac{\sigma(u_o^T V_c)(1-\sigma(u_o^T V_c))}{\sigma(u_o^T V_c)} - \sum_{k=1}^{k} \frac{\sigma(-u_k^T V_c)(1-\sigma(-u_k^T V_c))}{\sigma(-u_k^T V_c)}$$

$$= (1-\sigma(u_o^T V_c)) - k + \sum_{k=1}^{k} \sigma(-u_k^T V_c)$$

$$\frac{\partial J_{Neg}}{\partial u_o} = 1-\sigma(u_o^T V_c)$$

$$\frac{\partial J_{Neg}}{\partial u_k} = -\sigma(-u_k^T V_c)$$

Those derivatives are much more efficient because computing the derivatives of J_{NS} requires summing over $|W|$ term while here we require only k terms. Usually $|W| >> k$ so this computation is more efficient

g) $c = w_t$

$$J_{sg}(v_c, w_{t-m}, \dots, w_{t+m}, U) = \sum_{\substack{-m \le j \le m \\ j \neq 0}} J(v_c, \overset{\text{"o"}}{\overset{\parallel}{w_{t+j}}}, U)$$

i) $$\frac{\partial J_{sg}}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

ii) $$\frac{\partial J_{sg}}{\partial v_c} = \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

iii) $$\frac{\partial J_{sg}}{\partial v_w} = 0$$

$v_w \neq v_c$

Section 4

a) $$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m\le j\le m \\ j\neq 0}} P_\theta(W_{t+j} \mid W_t)$$

$$J(\theta) = \log L(\theta) = \sum_{t=1}^{T} \sum_{\substack{-m\le j\le m \\ j\neq 0}} \log P_\theta(W_{t+j} \mid W_t)$$

We will show that if $\theta^* = \arg\max_\theta L(\theta)$ then $P_{\theta^*}(o|c) = \dfrac{\#(c,o)}{\sum_{o'} \#(c,o')}$

Maximizing $L(\theta)$ is equivelant to maximizing $J(\theta)$

Since we are calculating the probability for every center word independently we can look at a specific center word c and the relevant function is

$$\sum_o \log(P_\theta(o|c))\#(c,o)$$

Let us denote $x_i = P_\theta(o_i|c)$, $k_i = \#(c,o_i)$

We will also add the constraints: $\sum x_i = 1$

The lagrangian is $L(x,\lambda) = \sum_i (\log x_i) k_i - \lambda \sum_j x_j + \lambda$

$$\frac{\partial L}{\partial x_i} = \frac{k_i}{x_i} - \lambda \overset{!}{=} 0 \Rightarrow x_i = \frac{k_i}{\lambda}$$

we demand

$$\sum x_i = 1 \Rightarrow \sum \frac{k_i}{\lambda} = 1 \Leftrightarrow \frac{1}{\lambda}\sum k_i = 1 \Leftrightarrow \sum k_i = \lambda$$

We get $x_i = \dfrac{k_i}{\lambda} = \dfrac{k_i}{\sum k_i} = \dfrac{\#(c,o_i)}{\sum_j \#(c,o_j)}$

b) Let the Vocabulary be $V = \{a, b\}$

and the Corpus $C = \{\text{"aa"}, \text{"ba"}\}$

finally we will assume the mapping is $M: V \to \mathbb{R}$

if $P(o|c) = \dfrac{\exp(M(o)\cdot M(c))}{\sum_{\tilde{o}\in V} \exp(M(\tilde{o})M(c))}$

the optimal solution is $\dfrac{\#(C,o)}{\sum_{\tilde{o}\in V} \#(C,\tilde{o})}$

We will denote by X_v the scalar matching $v\in V$

(o,c)	optimal	$P(o\|c)$
(b,a)	0.5	$\frac{\exp(X_a\cdot X_b)}{\exp(X_a X_b)+\exp(X_a^2)}$
(a,a)	$\frac{2}{3}$	$\frac{\exp(X_a\cdot X_a)}{\sum \exp(\cdot\,\cdot)}$
(a,b)	1	$\frac{\exp(X_a X_b)}{\sum(\cdot\;\cdot)}$
(b,b)	0	$\frac{\exp(X_b^2)}{\sum(\cdot,\cdot)}$

if p is equal to the optimal solution we get that

$$\frac{\exp(X_b^2)}{\exp(X_a\cdot X_b)+\exp(X_b^2)} = 0 \iff \exp(X_b^2) = 0$$

there is no scalar assignment that satisfies $\exp(X^2 b) = 0$!

Q5- Paraphrase Detection

The model here is:

$$p\left(\text{the pair is paraphrase} \mid x_1, x_2\right) = \sigma(\text{relu}(x_1)^T \text{relu}(x_2))$$

where $\text{relu}(x) = \max(0, x)$

(a) Because of using the Relu function the dot product of $\text{relu}(x_1)^T \text{relu}(x_2)$ will always be greater or equal to 0. And $\sigma(z) \geq 0.5$ where $z \geq 0$, therefore the model will always classify the data as positive paraphrase.

Because the positive example ratio is 25% then we expect maximal accuracy of 0.25

(b) All we have to do to fix this is to drop the relu function so:

$$p\left(\text{the pair is paraphrase} \mid x_1, x_2\right) = \sigma(x_1^T x_2)$$

In that the dot product can be also negative and the model will be able to classify negative examples as well.