#2.

(a) $J = -\log P(O=o|C=c) = -\sum_{w \in vocab} \cancel{y_w \log P(O=o|C=c)}$

$= 0 - \log P(O=o|C=c) = -\sum_{w \in vocab, w \neq o} y_w \log P(O=w|C=c) - y_o \log P(O=o|C=c)$

$= -\sum_{w \in vocab} y_w \log P(O=w|C=c) = -\sum_{w \in vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

(b) $\frac{\partial J}{\partial v_c} = \frac{\partial \log\left(\frac{\exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)}\right)}{\partial v_c} = -\frac{\sum \exp(u_w^T v_c)}{\exp(u_o^T v_c)} \frac{\partial \left(\frac{\exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)}\right)}{\partial v_c}$

$\frac{\partial J}{\partial v_c} = \frac{\partial -\log \hat{y}}{\partial v_c} = -\frac{1}{\hat{y}_o} \frac{\partial \hat{y}}{\partial v_c} = -\frac{1}{\hat{y}_o} \frac{u_o \exp(u_o^T v_c) \sum \exp(u_w^T v_c) - \exp(u_o^T v_c) \sum u_w \exp(u_w^T v_c)}{\{\sum \exp(u_w^T v_c)\}^2}$

$= -\frac{1}{\hat{y}_o}\left(u_o \hat{y}_o - \hat{y}_o \frac{\sum u_w \exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)}\right) = \left(\frac{\sum u_w \exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)} - u_o\right)$

$= \frac{(u_1 \exp(u_1^T v_c) + \cdots + u_w \exp(u_w^T v_c))}{\sum \exp(u_w^T v_c)} - Uy = u_1 \hat{y}_1 + \cdots + u_w \hat{y}_w - Uy$

$= U(\hat{y} - y)$

$n \times 1$: $y = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $d \times n$: $U = \begin{bmatrix} | & | & \boxed{o} & | & | \end{bmatrix}$

$Uy = u_o$

$\hat{y} = \begin{bmatrix} P(O=o_1|C=c) \\ \vdots \\ P(O=o_w|C=c) \end{bmatrix}$, $d \times 1$: $v_c = \begin{bmatrix} \\ \end{bmatrix}$

(c) $\frac{\partial J}{\partial u_o} = \frac{\partial(-\log \exp(u_o^T v_c) + \log \sum \exp(u_w^T v_c))}{\partial u_o} = -v_c + \frac{1}{\sum \exp(u_w^T v_c)} \frac{\partial (\exp u_1^T v_c + \cdots + \exp u_w^T v_c)}{\partial u_o}$

$= -v_c + \frac{v_c \exp u_o^T v_c}{\sum \exp(u_w^T v_c)} = v_c(\hat{y}_o - y_o)$

$\frac{\partial J}{\partial u_{w \neq o}} = \frac{v_c \exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)} = v_c \hat{y}_w$

(d) $\partial U = [\partial u_1 | \cdots | \partial u_w]$

(e) $\sigma'(x) = \frac{e^x(e^x+1) - e^x \cdot e^x}{(e^x+1)^2} = \frac{e^x}{e^x+1} \cdot \frac{1}{e^x+1} = \sigma(1-\sigma)$

(f) $\frac{\partial J}{\partial v_c} = -\frac{1}{\cancel{\sigma(u_o^T v_c)}} \cdot \cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c)) \cdot u_o - \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c)) \cdot (-u_k)$

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

$\frac{\partial J}{\partial u_o} = -v_c(1-\sigma(u_o^T v_c))$

$\frac{\partial J}{\partial u_k} = v_c(1-\sigma(-u_k^T v_c))$

no exp + less word vocab

#4. 1-(g) set attention score to $-\inf$ for the padded part. padding implementation shouldn't affect the calculation.

(h) 12.29

(i) I. dot product attention is computationally cheap and forces encoder and decoder to have similar embedding spaces. Multiplicative attention has learnable parameters.

II. Additive attention has more freedom of embedding but is more expensive.

2-(a) polysynthetic language has lots of affixes, so a word means a sentence.

(b) It needs to contain less information.

(c) transfer; insights gained through one can be applied to other. It is effective of learning generalization.

(d) I. crown of daisies is rare, so her hair got attended high. ~~probably not enough samples.~~ Change attention mechanism not to capture several times.

II. didn't catch sexual bias. provide more examples.

III. Don't know phrase Littlefish. put Littlefish in training data.

(e) I. "Well", said Charlotte. It captured dialogue form. It contains name of person.

II. The Abraham said unto him, Moses and the prophets; that he may be with me. It fails to understand the end-part meaning of a sentence.

(f) I. $P_{c_1,1} = \frac{0+1+1+1+0}{5} = \frac{3}{5}$ $\quad P_{c_1,2} = \frac{0+1+1+0}{4} = \frac{1}{2}$

$BP_{c_1} = 1$ $\quad BLEU_{c_1} = \exp\left(\frac{1}{2}\log 3/5 + \frac{1}{2}\log 1/2\right) = 0.5477$

$P_{c_2,1} = \frac{1+1+0+1+1}{5} = \frac{4}{5}$ $\quad P_{c_2,2} = \frac{1+0+0+1}{4} = \frac{1}{2}$

$BP_{c_2} = 1$ $\quad BLEU_{c_2} = \exp\left(\frac{1}{2}\log 4/5 + \frac{1}{2}\log 1/2\right) = 0.6325$

Second. c_2 sounds like to be better translation.

II. $P_{c_1,1} = \frac{0+1+1+1+0}{5} = \frac{3}{5}$ $\quad P_{c_1,2} = \frac{0+1+1+0}{4} = \frac{1}{2}$ $\quad BP_{c_1} = \exp(1-\frac{6}{5}) = 0.8187$

$BLEU_{c_1} = 0.8187 \exp\left(\frac{1}{2}\log 3/5 + \frac{1}{2}\log 1/2\right) = 0.4484$

$P_{c_2,1} = \frac{1+1+0+0+0}{5} = \frac{2}{5}$ $\quad P_{c_2,2} = \frac{1+0+0+0}{4} = \frac{1}{4}$ $\quad BP_{c_2} = 0.8187$

$BLEU_{c_2} = 0.8187 \exp\left(\frac{1}{2}\log 2/5 + \frac{1}{2}\log 1/4\right) = 0.2589$

c_1 receives. No.

III. better translation can get low BLEU score due to lack of exact overlap of n-grams.

IV. adv) quantitive, simple

disadv) only count of overlaps, no semantic score.

#5. 1. (a) suppose $\exists j \in \{1, \cdots n\}$ s.t $c = v_j$

That means, $\alpha_j = 1$, α_i for $\forall i \in \{1, \cdots n\} \setminus \{j\} = 0$

$\Rightarrow$ $\exp(k_i^T q)$ for $\forall i \in \{1, \cdots n\} \setminus \{j\} = 0$, $\exp(k_j^T q) = 1$

$k_i = -\inf$, $q = 1$, $k_j = 0$ $\Rightarrow k_j^T q = 0$

(b) $\alpha_a = \alpha_b = 1/2$, $\alpha_{etc} = 0$ $\Rightarrow \exp(k_1^T q) = \exp(k_2^T q) \gg 1$, $\exp(k_{etc}^T q) = 1$

Let $q = \cancel{\alpha}(k_a + k_b)$ $\alpha_a = \dfrac{\exp(k_a^T (k_a + k_b))}{\exp(k_a^T (k_a + k_b)) + \exp(k_b^T (k_a + k_b))} \simeq \dfrac{\exp \alpha}{2 \exp \alpha + n - 2} - \dfrac{1}{2}$

Let $q = (k_a + k_b) \times \log n \; \Sigma k_{etc} \times \log \frac{1}{n-2}$ $\alpha_a = \dfrac{n}{2n + \frac{n-2}{n-2}} = \dfrac{n}{2n+1} \simeq \dfrac{1}{2}$

(c) I. same. $E(k_a^T(\log n (k_a + k_b) + \Sigma k_{etc} \times \log \frac{1}{n-2})) = E[\log n] + E[\log n k_a^T k_b]$

$+ E[\Sigma k_{etc} \times \log \frac{1}{n-2}] = \log n$

II. since k_a becomes larger, c will tilt to a magnitude.

(d) I. $q_1 = \log n \, \mu_a + \Sigma \mu_{etc} \times \log \frac{1}{n-1}$ $q_2 = \log n \, \mu_b + \Sigma \mu_{etc} \times \log \frac{1}{n-1}$

II. scale becomes equal at α_a so c remains steady.

(e) I. $c_2 = \sum_{j=1}^{n} \alpha_{2j} v_j$ $\alpha_{21} = \dfrac{1}{\exp((u_d + u_b)^T u_a) + \exp(u_a^T u_a) + \exp((u_c + u_b)^T u_a)}$

$\simeq \alpha_{22} v_2 = \frac{1}{e^{\beta^2}} u_a$ $\quad 1 \quad e^{\beta^2} \quad 1$

No. α_{21} or α_{23} won't get affected.

II. $V(u_d + u_b) = u_b$, $V(u_c + u_b) = u_b - u_c$ $\qquad k_1^T q_2 = \beta^4$, $k_2^T q_2 = 0$, $k_3^T q_2 = 0$

$V = \frac{1}{\beta^2}(u_b u_b^T - u_c u_c^T)$ $\qquad k_1^T q_1 = 0$, $k_2^T q_1 = 0$, $k_3^T q_1 = 2\beta^4$

$v_1 = u_b$ $\qquad c_2 = \alpha_{21} u_b + \alpha_{23}(u_b - u_c) = u_b$ $\qquad (u_d + u_b)^T K^T Q u_a = \beta^2$

$v_2 = 0$ $\qquad c_1 = \alpha_{11} u_b + \alpha_{13}(u_b - u_c) = u_b - u_c$ $\qquad (u_c + u_b)^T K^T Q (u_d + u_b) = \beta^2$

$v_3 = u_b - u_c$

$Q = u_d u_a^T + u_c (u_d + u_b)^T$ $\qquad K = (u_d u_d^T + u_c u_c^T$

$Q_1 = -2\beta^2 u_c$ $Q_2 = u_d \beta^2$ $Q_3 = \beta^2 u_c$ $\qquad k_1 = u_d \beta^2$ $k_2 = 0$ $k_3 = u_c \beta^2$

2. (g) TT. It cannot learn relationship between Xs.

3. (a) by the charcorruption dataset model, it can learn more information and knowledge for deep epochs.

Also, it can be injected more general knowledge and finetune for specific target.

(b) A person cannot no whether the model is precise or lucky.

It can lead to distrust of the system.

(c) It will see relevant person such as similar job, age, address, etc.

However, it will never guarantee the answer hence resulting less reliability.