NLP – homework 5 **Ghazal Zamaninejad**

Part 1

a) $c \simeq v_j \Rightarrow \sum_{i=1}^{n} v_i \alpha_i = v_j \Rightarrow \begin{cases} \alpha_i = 0 & i \neq j \\ \alpha_i = 1 & i = j \end{cases}$

$$\Rightarrow \begin{cases} i \neq j & \dfrac{\exp(k_i^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)} = 0 \Rightarrow k_i^T q = -\infty \\ \\ i = j & \dfrac{\exp(k_i^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)} = 1 \Rightarrow \exp(k_i^T q) = \sum_{j=1}^{n} \exp(k_j^T q) \Rightarrow \sum_{j \neq i} \exp(k_j^T q) = 0 \end{cases}$$

$$\Rightarrow k_i^T q = -\infty \xrightarrow[q \neq -\infty]{} k_i = -\infty \quad (i \neq j)$$

b) $c = \frac{1}{2}(v_a + v_b) \Rightarrow \sum_{i=1}^{n} v_i \alpha_i = \frac{1}{2}(v_a + v_b) \Rightarrow \begin{cases} \sum_{i=1}^{n} v_i \alpha_i = 0 & i \neq a, i \neq b \\ \alpha_a = \alpha_b = \frac{1}{2} & i = a, i = b \end{cases}$

$$\Rightarrow \frac{\exp(k_a^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)} = \frac{\exp(k_b^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)} = \frac{1}{2} \Rightarrow \exp(k_a^T q) = \exp(k_b^T q) = \frac{1}{2} \sum_{j=1}^{n} \exp(k_j^T q)$$

$$\Rightarrow k_a q \simeq k_b q \rightarrow (k_a - k_b) q = 0 \rightarrow q = \underbrace{x}_{\text{large scalar}}(k_a + k_b)$$

$$k_a \cdot k_b = 0$$

c) i) $q = x(\mu_a + \mu_b)$ same as part b

ii) assume $q = x(\mu_a + \mu_b) \Rightarrow c = \sum_{i=1}^{n} v_i \alpha_i$

$$\begin{cases} i \neq a, i \neq b & k_i q^T = 0 \\ i = a & k_a q^T = \varepsilon_a \mu_a \times x(\mu_a + \mu_b) = x \varepsilon_a (\overbrace{\mu_a \mu_a}^{1} + \overbrace{\mu_a \mu_b}^{0}) = x \varepsilon_a \\ i = b & k_b q^T = \mu_b \times x(\mu_a + \mu_b) = x(\underbrace{\mu_a \mu_b}_{0} + \overbrace{\mu_b \mu_b}^{1}) = x \end{cases}$$

$$c = \frac{\exp(x \varepsilon_a) v_a + \exp(x) v_b + 0}{\exp(x \varepsilon_a) + \exp(x)}$$

$\varepsilon_a \uparrow \rightarrow c$ will be close to v_a

$\varepsilon_a \downarrow \rightarrow c$ " v_b

d) i) $q_a = x\mu_a$, $q_b = x\mu_b$ $\quad$ x large scalar

ii) $c_1 = \dfrac{\exp(x\,\varepsilon_a)}{\exp(x\,\varepsilon_a)} v_a = v_a$

$\rightarrow c = \frac{1}{2}(c_1 + c_2) = \frac{1}{2}(v_a + v_b)$

$c_2 = \dfrac{\exp(x)}{\exp(x)} v_b = v_b$

e) i) $v_1 = q_1 = k_1 = x_1 = u_d + u_b$

$v_2 = q_2 = k_2 = x_2 = u_a$

$v_3 = q_3 = k_3 = x_3 = u_c + u_b$

$\sum_{i=1}^{3} \exp(k_i^T q_2) = \exp(u_a^T \cdot u_a) = \exp(\beta^2)$

$c_2 = \alpha_{21} v_1 + \alpha_{22} v_2 + \alpha_{23} v_3$

$$\alpha_{21} = \frac{\exp(k_1^T q_2)}{\sum_{i=1}^{3} \exp(k_1^T q_2)} = \frac{\exp((u_d + u_b)^T \cdot u_a)}{\underbrace{\exp((u_d + u_b)^T \cdot u_a)}_{0} + \underbrace{\exp(u_a^T u_a)}_{\beta^2} + \underbrace{\exp((u_c + u_b)^T \cdot u_a)}_{0}}$$

$= 0 \rightsquigarrow u_a, u_b, u_c, u_d$ mutually orthogonal

$$\alpha_{22} = \frac{\exp(k_2^T q_2)}{\sum_{i}^{3} \exp(k_i^T q_2)} = \frac{\exp(u_a^T u_a)}{\exp(\beta^2)} = \frac{\exp(\beta^2)}{\exp(\beta^2)} = 1$$

$$\alpha_{23} = \frac{\exp(k_3^T q_2)}{\sum_{i=1}^{3} \exp(k_1^T q_2)} = \frac{\exp((u_c + u_b)^T \cdot u_a)}{\exp(\beta^2)} = 0$$ $\rightsquigarrow u_a, u_b, u_c, u_d$ mutually orthognal

$c_2 = \alpha_{21} v_1 + \alpha_{22} v_2 + \alpha_{23} v_3 = 0 + v_2 + 0 = v_2 = u_a$

ii) $v_i = V x_i \ , \ k_i = K x_i \ , q_i = Q x_i$

$x_1 = u_d + u_b \ , \ x_2 = u_a \ , \ x_3 = u_c + u_b$

$c_2 = u_b \ , \ c_1 = u_b - u_c$

$$\left.\begin{aligned} v_1 &= u_b = V x_1 \\ v_3 &= u_b - u_c = V x_3 \end{aligned}\right\} \Rightarrow V = \frac{1}{\beta^2} u_b u_b^T - \frac{1}{\beta^2} u_c u_c^T$$

$$v_1 = \frac{1}{\beta^2} u_b u_b^T u_d + \frac{1}{\beta^2} u_b u_b^T u_b - \frac{1}{\beta^2} u_c u_c^T u_d - \frac{1}{\beta^2} u_c u_c^T u_b = \frac{1}{\beta^2} u_b u_b^T u_b$$

$$= \frac{\beta^2}{\beta^2} u_b = u_b$$

$v_3 = u_b - u_c \ , \ v_2 = 0$

$$K = I \Rightarrow \begin{cases} k_1 = K x_1 = u_d + u_b \\ k_2 = K x_2 = u_a \\ k_3 = K x_3 = u_c + u_b \end{cases}$$

$$\alpha_{11} = \frac{\exp(k_1^T q_1)}{\exp(k_1^T q_1) + \exp(k_2^T q_1) + \exp(k_3^T q_1)}$$

$$\alpha_{12} = \frac{\exp(k_2^T q_1)}{\exp(k_1^T q_1) + \exp(k_2^T q_1) + \exp(k_3^T q_1)}$$

$$\alpha_{13} = \frac{\exp(k_3^T q_1)}{\exp(k_1^T q_1) + \exp(k_2^T q_1) + \exp(k_3^T q_1)}$$

$c_1 = u_b = v_1 \ , \ c_2 = u_b - u_c = v_3 \qquad \Rightarrow Q = \frac{1}{\beta^2} u_d u_d^T + \frac{1}{\beta^2} u_c u_a^T$

$q_1 = Q x_1 = u_d \ , \ q_2 = Q x_2 = u_c \ , \ q_3 = Q x_3 = u_c$

$k_1 = K x_1 = u_d + u_b \ , \ k_2 = K x_2 = u_a \ , \ k_3 = K x_3 = u_c + u_b$

$v_1 = b = V x_1 \ , \ v_3 = u_b - u_c = V x_2 \ , \ v_2 = 0$

$$c_1 = \frac{\exp(k_1^T q_1)}{\exp(k_1^T q_1) + \exp(k_2^T q_1) + \exp(k_3^T q_1)} v_1 + 0 + \frac{\exp(k_3^T q_1)}{\exp(k_1^T q_1) + \exp(k_2^T q_1) + \exp(k_3^T q_1)} v_3$$

$$= \frac{\exp(\beta^2)}{\exp(\beta^2) + 2} v_1 + \frac{1}{\exp(\beta^2) + 2} v_3 \simeq v_1$$

$$c_2 = \frac{\exp(k_1^T q_2)}{\exp(k_1^T q_2) + \exp(k_2^T q_2) + \exp(k_3^T q_2)} v_1 + 0 + \frac{\exp(k_3^T q_2)}{\exp(k_1^T q_2) + \exp(k_2^T q_2) + \exp(k_3^T q_2)} v_3 =$$

$$= \frac{1}{\exp(\beta^2) + 2} v_1 + \frac{\exp(\beta^2)}{\exp(\beta^2) + 2} v_3 \simeq v_3$$

Part 2

d)

```
data has 418352 characters, 256 unique.
number of parameters: 3323392
500it [00:43, 11.36it/s]
Correct: 5.0 out of 500.0: 1.0%
```

London prediction evaluation: Correct 25.0 of 500.0, Accuracy: 5.0%

f)

```
data has 418352 characters, 256 unique.
number of parameters: 3323392
500it [00:43, 11.43it/s]
Correct: 153.0 out of 500.0: 30.599999999999998%
```

g)

i)

```
data has 418352 characters, 256 unique.
number of parameters: 3076988
500it [00:54,  9.11it/s]
Correct: 34.0 out of 500.0: 6.800000000000001%
```

ii)

The synthesizer cannot evaluate the relevance between all pairs of words and can't understand contextual information. Adding multiple layers, cause it works better.

Part 3

a) When we pretrain the model, it learns facts about world and the training process takes much time. It uses span corruption which helps the model to learn general facts.

b) It causes in unreliable outputs and misleading users. Moreover, its outputs might be immoral.

c) It might choose the nearest person's name in embedding space to the unknown persons, and output his birth place which is completely wrong in real world. Using attention might improve its understanding the meaning of the word in context, which helps to improve the performance of the model.