Natural Language Processing

Hadi Sheikhi 97521369

1)

a) Suppose we want to c = v_j :

$$c = v_j \rightarrow for\ \ every\ \ \ i \neq j, we\ have\ \alpha_i = 0 \rightarrow \frac{exp(k_i^T q)}{\sum_{t=1}^{n} exp\ (k_t^T q)} = 0$$

And this means for every $i \neq j$we should have $k_i^T q = (-\infty)$then we have:

$$\alpha_j = \frac{exp\left(k_j^T q\right)}{exp\left(k_j^T q\right)} = 1$$

now by replacing values of αin formula we have:

$$c = \sum_{i=1}^{n} v_i\, \alpha_i = v_j \alpha_j = v_j$$

Now if we want to do so, we can simply set the k vectors to $-\infty$ except k_j.

b) Suppose:

$$q = \ C(k_a + k_b)$$

If we choose q like that and by considering C as a very large scalar we can write:

$$\alpha_i = \frac{\exp(q^T . k_i)}{\sum_{j=1}^{n} \exp(q^T . k_j)} = \frac{\exp(q^T . k_i)}{n - 2 + 2\exp(C)} \xrightarrow{C \gg n} \alpha_i = \frac{\exp(q^T . k_i)}{2\exp(C)}$$

$$c = \sum_{i=1}^{n} v_i \alpha_i \approx \sum_{i=1}^{n} v_i \frac{\exp(q^T . k_i)}{2\exp(C)} \approx \frac{1}{2\exp(C)} v_1 + \frac{1}{2\exp(C)} v_2 + \cdots + \frac{\exp(C)}{2\exp(C)} v_a + \frac{\exp(C)}{2\exp(C)} v_b + \cdots + \frac{1}{2\exp(C)} v_n \approx 0 + \frac{1}{2} v_a + \frac{1}{2} v_b$$

c)

$$k_i \sim N(\mu_i, \Sigma_i)$$

i) Suppose $\Sigma_i = \alpha I$.

$$k_i \sim N(\mu_i, \alpha I) \rightarrow k_i \approx u_i\ because\ E(k_i) = \ \mu_i$$

Now same as part b, suppose:

$$q = \ C(\mu_a + \mu_b)$$

Proof is just like part b.

ii) Suppose $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^T)$.

$$k_a \sim N\left(\mu_a, \alpha I + \frac{1}{2}(\mu_a \mu_a^T)\right) \rightarrow k_a \approx \epsilon_a \mu_a, \epsilon_a \sim N\left(1, \frac{1}{2}\right)$$

When we consider $q = C(\mu_a + \mu_b)$:

$$for\ i \neq a, b;\ q^T.k_i = 0$$
$$q^T.k_a = \epsilon_a C u_a^T.u_a = \epsilon_a C$$
$$q^T.k_b = C u_b^T.u_b = C$$
$$c = \sum_{i=1}^{n} v_i \alpha_i \approx \frac{\exp(\epsilon_a C)}{\exp(\epsilon_a C) + \exp(C)} v_a + \frac{\exp(C)}{\exp(\epsilon_a C) + \exp(C)} v_b$$

For larger values of ϵ_a, c is closer than to v_a and vice versa. Note that $\epsilon_a \sim N\left(1, \frac{1}{2}\right)$.

d)

i) Consider

$$q_1 = C_1 \mu_a$$
$$q_2 = C_2 \mu_b$$

where C_1 and C_2 are very large positive scalars.

ii) Same as part ‘ii’ of part c (!), if we wrote the equations we get:

$$c_1 \approx \frac{\exp(\epsilon_a C)}{\exp(\epsilon_a C)} v_a \approx v_a$$
$$c_2 \approx \frac{\exp(C)}{\exp(C)} v_b \approx v_b$$
$$c = \frac{1}{2}(c_1 + c_2) = \frac{1}{2}(v_a + v_b)$$

e)

i)

$$q_1 = k_1 = v_1 = x_1 = u_d + u_b$$
$$q_2 = k_2 = v_2 = x_2 = u_a$$
$$q_3 = k_3 = v_3 = x_3 = u_c + u_b$$

$$\alpha_{21} = \frac{\exp(q_2^T.k_1)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)}$$
$$= \frac{\exp(u_a^T.u_d + u_a^T.u_b)}{\exp(u_a^T.u_d + u_a^T.u_b) + \exp(u_a^T.u_a) + \exp(u_a^T.u_c + u_a^T.u_b)}$$
$$= \frac{1}{2 + \exp(\beta^2)} \approx 0$$
$$\alpha_{22} = \frac{\exp(q_2^T.k_2)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)} = \frac{\exp(\beta^2)}{2 + \exp(\beta^2)} \approx 1$$
$$\alpha_{23} = \frac{\exp(q_2^T.k_3)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)} = \frac{1}{2 + \exp(\beta^2)} \approx 0$$

$$c_2 = \frac{u_d + u_b}{2 + \exp(\beta^2)} + \frac{\exp(\beta^2)\, u_a}{2 + \exp(\beta^2)} + \frac{u_c + u_b}{2 + \exp(\beta^2)} \approx u_a$$

Now suppose we add u_c to x_2. In calculating values of α_2:

$$\alpha_{21} = \frac{\exp(q_2^T.k_1)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)}$$
$$= \frac{\exp(u_a^T.u_d + u_a^T.u_b + u_c^T.u_d + u_c^T.u_b) = 1}{1 + \exp(u_a^T.u_a + u_a^T.u_c + u_c^T.u_a + u_c^T.u_c) + \exp(u_a^T.u_c + u_a^T.u_b + u_c^T.u_c + u_c^T.u_b)}$$
$$= \frac{1}{1 + \exp^2(\beta^2) + \exp(\beta^2)} \approx 0$$
$$\alpha_{22} = \frac{\exp(q_2^T.k_2)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)} = \frac{\exp^2(\beta^2)}{1 + \exp^2(\beta^2) + \exp(\beta^2)} \approx 1$$
$$\alpha_{23} = \frac{\exp(q_2^T.k_3)}{\exp(q_2^T.k_1) + \exp(q_2^T.k_2) + \exp(q_2^T.k_3)} = \frac{\exp(\beta^2)}{1 + \exp^2(\beta^2) + \exp(\beta^2)} \approx 0$$

And now calculating c_2.

$$c_2 \approx x_2 = u_a + u_c$$

As we can see we cannot approximate u_b, because u_b is in x_1 and x_2 and α_{23} and α_{21} are always 0 by adding either u_c or u_d.

ii)

First we find V such that:

$$v_1 = u_b = Vx_1$$
$$v_3 = u_b - u_c = Vx_2$$

If we write:

$$V = \frac{1}{\beta^2} u_b u_b^T - \frac{1}{\beta^2} u_c u_c^T$$

Then

$$v_1 = \frac{1}{\beta^2} u_b u_b^T u_d + \frac{1}{\beta^2} u_b u_b^T u_b - \frac{1}{\beta^2} u_c u_c^T u_d - \frac{1}{\beta^2} u_c u_c^T u_b = \frac{1}{\beta^2} u_b u_b^T u_b = \frac{\left|\left|u_b\right|\right|^2}{\beta^2} u_b$$
$$= u_b$$
$$v_3 = \frac{1}{\beta^2} u_b u_b^T u_c + \frac{1}{\beta^2} u_b u_b^T u_b - \frac{1}{\beta^2} u_c u_c^T u_c - \frac{1}{\beta^2} u_c u_c^T u_b = \frac{1}{\beta^2} u_b u_b^T u_b - \frac{1}{\beta^2} u_c u_c^T u_c$$
$$= \frac{\left|\left|u_b\right|\right|^2}{\beta^2} u_b - \frac{\left|\left|u_c\right|\right|^2}{\beta^2} u_c = u_b - u_c$$
$$v_2 = 0$$

We suppose $K = I$.

$$k_1 = Kx_1 = u_d + u_b$$
$$k_2 = Kx_2 = u_a$$
$$k_3 = Kx_3 = u_c + u_b$$

Now let's compute c_1

$$c_1 = \sum_{i=1}^{n} v_i \alpha_{1i} = v_1 \alpha_{11} + v_2 \alpha_{12} + v_3 \alpha_{13}$$
$$\alpha_{11} = \frac{\exp(q_1^T k_1)}{\exp(q_1^T k_1) + \exp(q_1^T k_2) + \exp(q_1^T k_3)}$$
$$\alpha_{12} = \frac{\exp(q_1^T k_2)}{\exp(q_1^T k_1) + \exp(q_1^T k_2) + \exp(q_1^T k_3)}$$

Hadi Sheikhi 97521369

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

We want to $c_1 \approx u_b \approx v_1$ so $\exp(q_1^T k_1)$ should be big enough.
Same for $c_2 \approx u_b - u_c \approx v_3$ so $\exp(q_2^T k_3)$ should be big enough.

If we consider:

$$Q = \frac{1}{\beta^2} u_d u_d^T + \frac{1}{\beta^2} u_c u_a^T$$

$$q_1 = Qx_1 = u_d, q_2 = Qx_2 = u_c, q_3 = Qx_3 = u_c$$
$$k_1 = Kx_1 = u_d + u_b, k_2 = Kx_2 = u_a, k_3 = Kx_3 = u_c + u_b$$
$$v_1 = u_b = Vx_1, v_3 = u_b - u_c = Vx_2, v_2 = 0$$

Now let's validate the answer:

$$\begin{aligned} c_1 = v_1 & \frac{\exp(q_1^T k_1)}{\exp(q_1^T k_1) + \exp(q_1^T k_2) + \exp(q_1^T k_3)} + 0 \\ & + v_3 \frac{\exp(q_1^T k_3)}{\exp(q_1^T k_1) + \exp(q_1^T k_2) + \exp(q_1^T k_3)} \\ & = v_1 \frac{\exp(\beta^2)}{\exp(\beta^2) + 2} + v_3 \frac{1}{\exp(\beta^2) + 2} \approx v_1 \end{aligned}$$

$$\begin{aligned} c_2 = v_1 & \frac{\exp(q_2^T k_1)}{\exp(q_2^T k_1) + \exp(q_2^T k_2) + \exp(q_2^T k_3)} + 0 \\ & + v_3 \frac{\exp(q_2^T k_3)}{\exp(q_2^T k_1) + \exp(q_2^T k_2) + \exp(q_2^T k_3)} \\ & = v_1 \frac{1}{\exp(\beta^2) + 2} + v_3 \frac{\exp(\beta^2)}{\exp(\beta^2) + 2} \approx v_3 \end{aligned}$$

So far so good!

2)

- d) Correct: 4.0 out of 500.0: 0.8%
  London prediction evaluation: Correct 25.0 of 500.0, Accuracy: 5.0%
- f) Correct: 124.0 out of 500.0: 24.8%
- g) Correct: 78.0 out of 500.0: 15.6%
  - ii) Synthesizer can't understand the contextual information but if we do it on multilayer maybe it works better.

3)

- a) When we pretrain model with span corruption, we give more information to model about the actual context of our corpus and it causes more accuracy.
- b) –
- c) By applying attention, we can understand words in contexts. Now it may didn't even see that name but we can understand the meaning in context which learned before.