

## Natural Language Processing (CS22N)

Student Name:

Babak Behkamkia

Instructor Name:

Dr. Sauleh Eetemadi

## 1. Attention exploration

a.

- i. We use softmax to calculate the alpha values. Thus, these values will be a number between 0 and 1 and their sum would be 1. This shows these values are a probability distribution. Moreover, each alpha value assigned to a word in the sequence and shows the importance of that word in that sequence of words. Therefore, alpha scores are a categorical probability distribution.
- ii.  $k_i q >> k_i q$  where i,  $j \in \{1, 2, ..., n\}$  and i!=j
- iii. In this case, the value of c would be very close to  $v_i$  because  $k_i$  q has the largest value and after applying the softmax function, the corresponding value would be 1 for it. Moreover, from the equation (1) we can understand that  $c = v_i$ .
- iv. If the key value of a word is similar to q value of the chosen word, the product of those matrices will be large and after softmax, it will put almost all the weight on the corresponding alpha.

b.

$$C = \frac{1}{2} (V_A + V_b) \Rightarrow \begin{cases} A_A = \frac{1}{2} \\ A_b = \frac{1}{2} \end{cases}$$

$$\Rightarrow \begin{cases} A_A + A_b \end{cases} \Rightarrow k;$$

$$A = B (k_A + k_b) \Rightarrow k;$$

$$\Rightarrow \begin{cases} K_A + K_b \end{cases} \Rightarrow large number$$

$$\Rightarrow \begin{cases} K_A + K_b \end{cases} \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

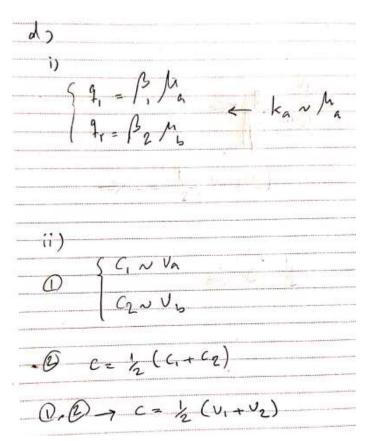
$$k_b + k_b = B \Rightarrow they are orthogonal$$

$$k_b + k_b = B \Rightarrow they are orthogonal$$

c.

E: =  $\times$  [  $\times$  |  $\times$  |

d.



## 2. Pretrained Transformer models and knowledge access

d.

Accuracy: 2%

London accuracy: 5%

f. Accuracy: 26.4%

g.

i. Accuracy: 13%

## 3. Considerations in pretrained knowledge

a. Overall, pretrained models are better than non-pretrained models because they have trained on a large dataset previously. In other words, their weights are not random and they have learned some knowledge from the previous task.

b.

- 1. Trustworthiness: If an application generates unreal information like made up birthplaces, research papers, or websites, it will reduce the users trust on the application.
- 2. Using wrong information: People may not notice the made up answer and use it in critical situations, which leads to a big problem.
- c. Obviously, the model cannot determine the birthplace of a person that it never seen it before but with providing more information about that person for the model can help it in order to find similar individuals and make a prediction based on this information. For example, being angry most of the time may be exclusive to the people of a specific country.