Assignment 2:

1. Part 1

1.1 Problem 1:

i)

t, m, h	3, 7
m, h, a	0, 2
h, a, d	1, 5
a, d, c	7, 4

ii)

∅, t, m	0, 3
m, h, a	0, 2
a, d, c	7, 4

iii) Apply to (i)

2 - max pooling 3, 7
7, 4

Apply to (ii)

2-max pooling 0, 3
7, 4

2.

- It is useful because it allow model to handle new words and increase generalization of model.

3.

i) The idea of attention flow is that the attention should flow in both ways - from the context to the question and from the question to context

ii)

Context-to-question task is to generate an answer based on context

=> Context-to question attention is to attend to parts of context that relate to the question by computing attention weight for words in the context

iii)

ii)

Context-to-question attention pay attention to different parts of the context based on the question to generate answer

iii)

Question-to-context attention pay attention to different part of the question based on the context to generate answer

4.

- Because single-headed attention can only focus on one set of relationship between different parts while multi-headed attention can capture all relevent information

=> prefer mult-headed attention

5.

i) if the beam size k is too small. The algorithm may not explore all possible choices

=> may miss out on some good results

ii) if the beam size k is too large. The algorithm can explore many good choices and have better result.

⇒ Cost more to compute and increase search space and may cause over-fitting

6.

i) ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is a set of metrics used to evaluate the quality of a summary or generated text by comparing it to one or more references text.

ii) BLEU stands for Bilingual Evaluation Understudy. It is a metric used to evaluate the quality of machine translation systems by comparing the generated translation to one or more reference translations.

iii) The difference:

BLEU: focuses on precision: how much the words in the candidate model outputs appear in the human reference.

ROUGE: focuses on recall: how much the word in the human references appear in the candidate model outputs.

7.

i) Coreference resolution is identify all mentions that refer to to the same real world entity

ii) * Cluster 1

♡ P = 7/9
R = 7/8

⬠ P = 2/9
R = 2/6

* Cluster 2

♡ P = 1/5
R = 1/8

⬠ P = 4/5
R = 4/6

$$P = \left(7.\frac{7}{9} + 2.\frac{2}{9} + 1.\frac{1}{5} + 4.\frac{4}{5}\right)/14 = 0,663$$

$$R = \left(7.\frac{7}{8} + 2.\frac{2}{6} + 1.\frac{1}{8} + 4.\frac{4}{6}\right)/14 = 0,685$$

8.

For x_1:

$$v_1 = V.(u_d + u_\ell) = u_d + u_\ell = x_1$$
$$k_1 = K(u_d + u_\ell) = u_d + u_\ell = x_1$$
$$q_1 = Q(u_d + u_\ell) = u_d + u_\ell = x_1$$

For x_2:

$$v_2 = V.u_a = u_a = x_2$$
$$k_2 = K.u_a = u_a = x_2$$
$$q_2 = Q.u_a = u_a = x_2$$

For x_3:

$$v_3 = V.(u_c + u_\ell) = u_c + u_\ell = x_3$$
$$k_3 = K.(u_c + u_\ell) = u_c + u_\ell = x_3$$
$$q_3 = Q.(u_c + u_\ell) = u_c + u_\ell = x_3$$

$$\alpha_{21} = \frac{\exp(k_1^T . q_2)}{\sum_{\ell=1}^{3} \exp(k_\ell^T . q_2)} = \frac{\exp(k_1^T . q_2)}{\exp(k_1^T . q_2) + \exp(k_2^T . q_2) + \exp(k_3^T . q_2)}$$

$$= \frac{\exp((u_d + u_\ell)^T . u_a)}{\exp((u_d + u_\ell)^T . u_a) + \exp(u_a^T . u_a) + \exp((u_c + u_\ell)^T . u_a)}$$

$$= \frac{\exp(u_d^T . u_a + u_\ell^T . u_a)}{\exp(u_d^T . u_a + u_\ell^T . u_a) + \exp(u_a^T . u_a) + \exp(u_c^T . u_a + u_\ell^T . u_a)}$$

u_a, u_b, u_c, u_d are mutually orthogonal

$$\Rightarrow \alpha_{21} = \frac{\exp(0+0)}{\exp(0+0) + \exp(\|u_a\|^2) + \exp(0+0)}$$

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

$$a_{22} = \frac{\exp(k_2^T \cdot q_2)}{\sum_{l=1}^{3} \exp(k_l^T \cdot q_2)} = \frac{\exp(u_a^T \cdot u_a)}{2 + 0 \cdot \exp(\beta^2)}$$

$$= \frac{\exp(\beta^2)}{2 + \exp(\beta^2)}$$

$$a_{32} = \frac{\exp(k_3^T \cdot q_2)}{\sum_{l=1}^{3} \exp(k_l^T \cdot q_2)} = \frac{\exp(0+0)}{2 + \exp(\beta^2)} = \frac{1}{2 + \exp(\beta^2)}$$

Result:

$$c_2 = \sum_{j=1}^{3} a_{2j} \cdot v_{1j} = a_{21} \cdot v_1 + a_{22} \cdot v_2 + a_{23} \cdot v_3$$

$$= \frac{1}{2 + \exp(\beta^2)} \cdot (u_d + u_b) + \frac{\exp(\beta^2)}{2 + \exp(\beta^2)} \cdot u_a + \frac{1}{2 + \exp(\beta^2)} \cdot (u_c + u_b)$$

$$= \frac{\exp(\beta^2)}{2 + \exp(\beta^2)} \cdot u_a + \frac{2}{2 + \exp(\beta^2)} u_b + \frac{1}{2 + \exp(\beta^2)} \cdot (u_c + u_d)$$

When add u_d or u_c to $x_2 \Rightarrow x_2 = u_a + u_d$

$$\Rightarrow \sum_{l=1} \exp(k_l^T \cdot q_2) = \exp(k_1^T \cdot q_2) + \exp(k_2^T \cdot q_2) + \exp(k_3^T \cdot q_2)$$

$$= \exp((u_d + u_b)^T \cdot (u_a + u_d)) + \exp(\beta^2) + \exp((u_c + u_b)^T \cdot (u_a + u_d))$$

$$= \exp(u_d^T \cdot u_d) + \exp(\beta^2) + \exp(0)$$

$$2 \exp(\beta^2) + 1$$

$\Rightarrow a_{21} =$

- Since adding u_d or u_c to x_2 would not result in zero products we won't have c_2 to be approximate u_b