CSE 556: Natural Language Processing (NLP) — Mid-Sem

Date: 25th Feb 2025 **Duration:** 60 mins. **Max Marks: 40**

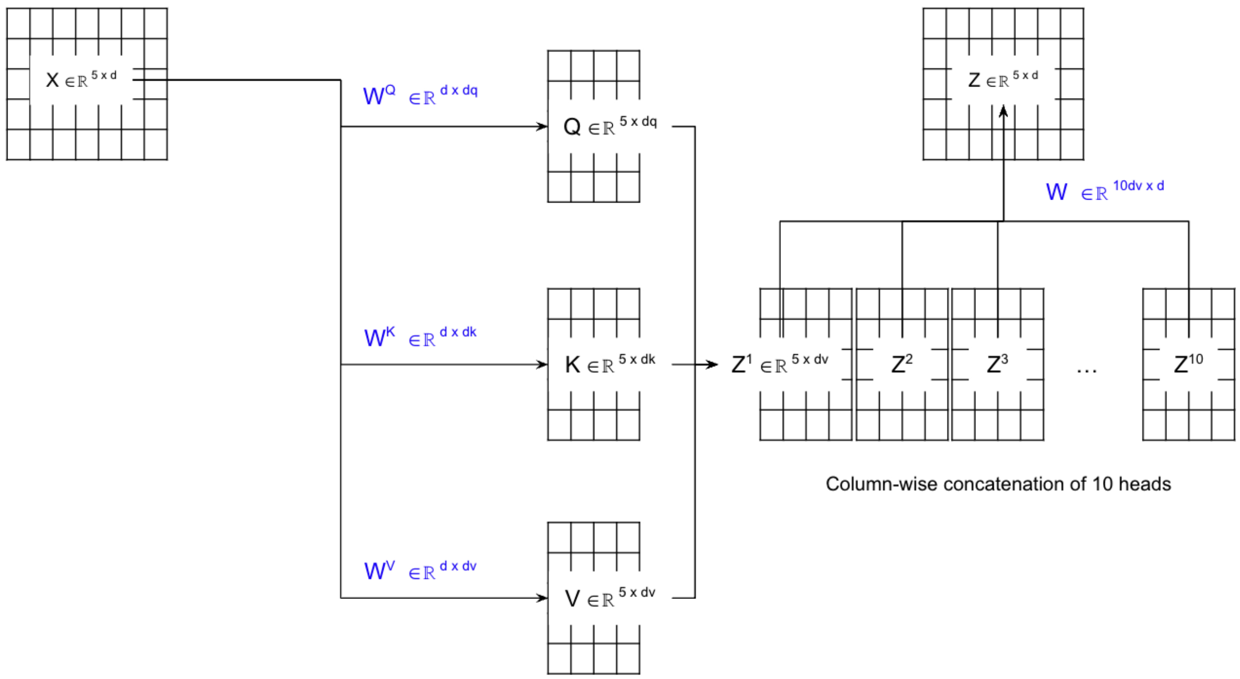
1. State the objective functions of BERT, GPT, and BART models. **[1+1+4]**
Note: We are not necessarily looking for equations.
BERT: MLM and NSP [0.5 + 0.5]
GPT: LM and Classification loss [0.5 + 0.5]
BART: Token Masking, Token Deletion, Sentence Permutation, Document Rotation, and Text Infilling.
Note for BART:
    - Any four out of five. 1 mark each.
    - If definitions are not present, exact matching should be done. Else, go by the definition instead of topic.

   For all cases, we will accept the answer (textual, graphical, equation, etc.) as long as it conveys the correct answer.

2. Given an input matrix, $X \in \mathbb{R}^{5 \times d}$, and dimensions of query, key, and value as d_q, d_k, and d_v, report the dimension of project matrix W^Q, W^K, W^V, and W in a 10-headed self attention block. **[4]**

   $W^Q \in \mathbb{R}^{d \times dq}$
   $W^K \in \mathbb{R}^{d \times dk}$
   $W^V \in \mathbb{R}^{d \times dv}$
   $W \in \mathbb{R}^{10dv \times d}$ 1 mark each

   Explanation (We are not expecting this):


$X \in \mathbb{R}^{5 \times d}$
$W^Q \in \mathbb{R}^{d \times dq}$
$Q \in \mathbb{R}^{5 \times dq}$
$W^K \in \mathbb{R}^{d \times dk}$
$K \in \mathbb{R}^{5 \times dk}$
$W^V \in \mathbb{R}^{d \times dv}$
$V \in \mathbb{R}^{5 \times dv}$
$Z^1 \in \mathbb{R}^{5 \times dv}$
Z^2
Z^3
…
Z^{10}
$Z \in \mathbb{R}^{5 \times d}$
$W \in \mathbb{R}^{10dv \times d}$
Column-wise concatenation of 10 heads

3. Suppose you are designing an architecture for a multimodal setup (e.g., text and image). Which of the following projections is/are not appropriate and which is/are appropriate? Justify for each case. No marks without justification. **[5*2]**
    a. Text: Query and Key; Image: Value
    b. Text: Query and Value; Image: Key
    c. Text: Value and Key; Image: Query
    d. Text: Query; Image: Key and Value
    e. Text: Query, Key, and Value; Image: Query, Key, and Value

    Key and Value should always be in pair. For a query, we compute the relevance of value by computing similarity with its corresponding key.
    Softmax(Q.K/ root(d_K)). V

4. Differentiate between LM and MLM as tasks. **[4]**

    LM → Generation task. Can't use future tokens. Only unidirectional RNNs can be used.

    MLM → Classification task. Can use future tokens. Bi-directional RNNs are possible.

    Note: If someone has mentioned at least two differences, give full marks. Else, deduct some marks.

5. For an information extraction task, if we are certain that no two entities can appear back-2-back, what would be a better encoding scheme? Justify your answer. **[2]**

    BIO: if we can have back-2-back entities ⇒ Ternary classification

IO: If there is guarantee that back-2-back entities are not possible. ⇒ Binary classification (less challenging than 3-class classification)

Note: We may accept other answers as well, if they are appropriate.

6. Discuss Markov assumption in the formulation of HMM. Why do we need it? **[2]**

Probability of a state (s_t) depends on the previous few states instead of the entire sequence of states. For Markov assumption of size 2, it depends on the previous state only.

$$P(s_t \mid s_1 s_2 \ldots s_{t-1}, O) \Rightarrow P(s_t \mid s_{t-1}, O)$$

7. Either one of the two: **[4]**
   a. State the equation of absolute discounting smoothing technique. Recall that it has two components, why do we need the second component?
   Check slides for the equation.
   For the case when discount is more than the count, the numerator of the first component becomes zero. To account for the same, we add unigram probability as interpolation.
   OR
   b. State the equation of continuation probability in Kneser-Ney smoothing technique. Discuss why we need it.
   Check slides for the equation.
   Higher frequency words should not dominate the probability computation if they are not an appropriate continuation of the previous token. Continuation probability assigns higher mass to the word which is an appropriate continuation despite having lesser unigram count.

8. State the generic principle of advanced smoothing techniques (e.g., Good Turing, Witten-Bell, Kneser-Ney, etc.). **[2]**

To distribute some probability mass (m_u) to all unseen (ngram-with-zero-count) samples, take it (m_u) from the probability mass (m_s) of samples seen just once.

Note: Answer should be on a similar line.

9. Describe morphological analysis and synthesis using an example. Which one of the two is easy and which one is challenging, and why? **[4]**

Mapping the surface-level form to the lexical-level form

○ Analysis: Surface-level (cats) ⇒ Lexical-level (cat + N + PL)

○ Synthesis: Lexical-level (cat + N + PL) ⇒ Surface-level (cats)

Note: 1 mark for each definition. Deduct 0.5 for no example.

Synthesis is easier than analysis (1 mark):

Caught ⇒ Catch + V + Past (verb 2nd form) **OR**
Catch + V + Past-Part (verb 3rd form) (**Ambiguity**)

Catch + V + Past ⇒ caught (Always)

Note: 1 mark for why?

10. Which linguistic concept in the NLP hierarchy can you associate to the following sentence? "*The use of shin bone is to locate furniture in a dark room.*" **[2]**

Pragmatics.

Note: If someone answered Pragmatics and Discourse, we will assign full marks. However, please note that Pragmatics and Discourse are different.