Name: ______________________________________ Roll: _____________________________

CSE556: NLP - Exam-sem

Marks: 50

Duration: 2 hours **Date:** 4-May-2024

1. During parsing, what are the conditions to conclude whether an input is a syntactically valid input or not? [**No partial marking.**] **[2]**

   In a parse,

   - **All leaf nodes must be terminal symbols representing tokens of the input.**
   - **All tokens of the input must be present in the parse-tree.**

2. Discuss the purpose of negative sampling and subsampling in the word2vec model. **[2]**

   Negative sampling: It helps to reduce the complexity of the system by calculating the unrelatedness of a word with a fixed/small set of non-contextual words.

   Subsampling: It helps to minimize the effect of highly frequent but low semantic contextual words.

3. Define continued pretraining. How is it different from pretraining and fine-tuning? **[2]**

   Pretraining:

   - **Take some large-scale generic unlabelled data**
   - **Initialize the model parameters randomly**
   - **Pre-train a model using MLM or other objective functions**
   - **Save weights of the model.**

   Continued pretraining:

   - **Take medium-size domain-specific unlabelled data**
   - **Initialize the model parameters using the saved weights after pre-training**
   - **Continue pre-training using MLM or other objective functions**
   - **Save updated model**

   Fine-tuning:

   - **Take labelled data for a task**
   - **Initialize the model parameters using the saved weights after pre-training or continued pretraining**
   - **Optimized the model using task-specific loss functions.**
   - **Save and evaluate the model.**

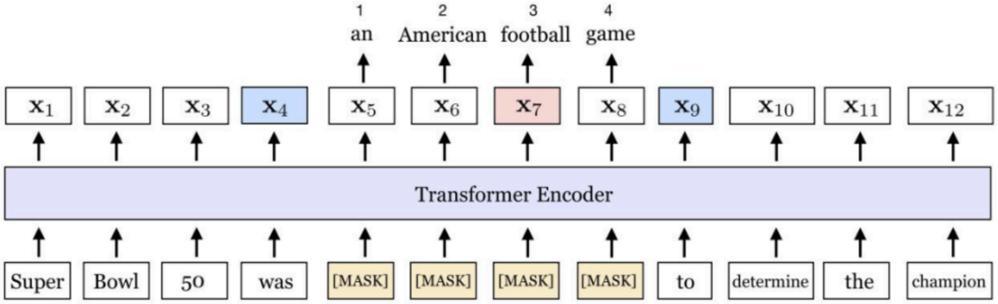
[No marks, if someone has not defined continued pre-training appropriately]

4. What is PHQ-9 and what is its objective? **[2]**

PHQ-9 (PATIENT HEALTH QUESTIONNAIRE) is a set of 9 questions used for initial / self diagnosis of depression symptoms.

5. Name and define the two objectives of SpanBERT. Equations are not mandatory**, but welcomed**. **[4]**

- Mask a sequence of tokens representing a span for prediction.



- Two objectives: MLM and SBO
  - Span-boundary objective (SBO) learns to predict the entire masked span from the observed tokens at its boundary.

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$

$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

Two objectives:

- **MLM**
- **Span-boundary objective (SBO) learns to predict the entire masked span from the observed tokens at its boundary.**

[1 mark for MLM; 1 mark for mentioning Span-boundary objective (SBO); and 2 marks for defining SBO.]

6. Clearly explain the process for computing the macro-F1 and weighted F1 scores for the following example. [You must have computed the macro-F1 and weighted-F1 scores in assignments 2 and 4, respectively.] **[6]**
[Note: We are not expecting definitions, so marks will not be assigned for it.]

		Predicted		**Support %**
		Positive	**Negative**	
Actual	**Positive**	**a**	**b**	**s1 = (a+b) / (a+b+c+d)**
	Negative	**c**	**d**	**s2 = (c+d) / (a+b+c+d)**

Classes are positive and negative.

For positive: P1: a / (a+c) R1: a / (a+b) F1: 2.P1.R1 / (P1+R1)	For negative: P2: d / (d+b) R2: d / (d+c) F2: 2.P2.R2 / (P2+R2)
Weighted: s1 * F1 + s2 * F2	
Macro: (F1 + F2) / 2	

7. How does Pointer Generator Network (PGN) work? Mention only key points in 2-3 points. [**Architecture or Equations are *not necessary*, but welcomed**] **[6]**

   - **Learn attention distribution over input and vocabulary distribution in a standard over decoding unit.**
   - **Learn a binary classifier, P-gen, which acts as a switch.**
   - **Combine vocabulary and attention distribution according to P-gen, for the final distribution.**

8. Mention different output perspectives for a summarization task. What are the advantages and disadvantages of one over the other? **[6]**

   Output perspectives:

   - **Extractive: [1 mark]**
     - **[+] Relatively easier task – Binary classification for each sentence [1 mark]**
     - **[-] Selected sentences in the summary may not be coherent with each other [1 mark]**
   - **Abstractive: [1 mark]**
     - **[-] Relatively difficult task – Generation problem [1 mark]**
     - **[+] Summaries are highly coherent (fluency/smoothness). [1 mark]**

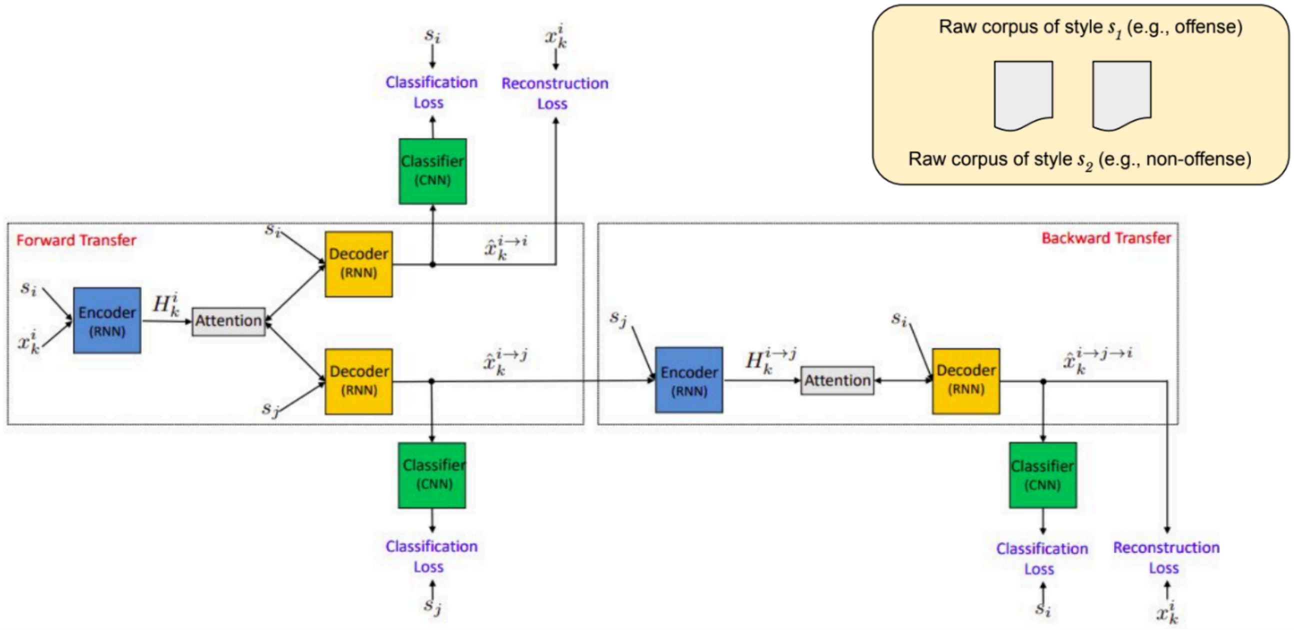
9. Assume you are given two random corpus from two different languages, e.g., D1 ∈ L1 and D2 ∈ L2, where L1 and L2 are two languages. Can you suggest a strategy (e.g., architecture, approach, etc.) to develop a MT model for translation from L1 → L2? You can not assume additional resources. **[10]**

   Random corpus means they are not parallel corpus.

   One of the possible solutions would be to do unsupervised style-transfer. Instead of offensive and non-offensive, you can have D1 and D2.

Answers who just mentioned the name of the approach (e.g., rule-based) without providing the complete process (e.g., without mentioning rules) will not get any marks.

Other solutions may also be possible and we will evaluate them based on its merits.



10. Assume that you were asked to design and develop a chat-bot for online banking. In order to do so, you need to first finalize all possible intents that might be required to serve the customer. Mention at least 5 such intents that should broadly cover typical banking operations. Further, for each intent, mention all slots that are absolutely necessary for serving the particular intent. **[10]**

Following are some of the examples. Other Intents and slots are also possible but they have to be reasonable and meaningful.

Intents	**Slots**
Balance-enquiry	Account No, Account Type
Transfer amount	ToAccount, FromAccount, Amount, Mode, Time-and-date for scheduling, Remarks
Request update [mobile, address, branch, etc.]	Account No, Account Type, Current detail, New detail, Reason for update, Proof for updated [OTP, id for address, etc.],
Generate account statement	Account No, Account Type, Duration_to, Duration-From, Format.
Add beneficiary	Beneficiary account number, Beneficiary Name, Beneficiary bank, Beneficiary branch, IFSC, Transfer Limit.

-- *End* --