# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## End-Spring Semester Examination 2022-23

**Date of Examination: 20/04/2023** **Session: AN** **Duration: 3 hrs.** **Full Marks: 70**

**Subject No.: _CS60075______** **Subject: Natural Language Processing**

**Department/Center/School: CSE**

**Specific charts, graph paper, log book etc., required**

**Special Instructions (if any):**

---

[*Note:* For this question paper, use base 10 for any log computation]

**All of the questions are compulsory. There are no clarifications. In case of any doubts, please make a valid assumption, write that explicitly and proceed. All the subparts corresponding to a single problem need to be answered together in the answer script.**

1. Consider below your dictionary containing 4 words, along with their frequencies in the corpus. Suppose you are using BPE to create a vocabulary. (a) What will be the initial vocabulary? (b) What will be the 3 character n-grams that you will add to your vocabulary next, in that order?

| Word | Frequency |
|---|---|
| enter | 3 |
| entry | 5 |
| tenth | 6 |
| cant | 2 |

(c) Suppose that you decide to use wordPiece instead. What would be the likelihood with the initial vocabulary and the vocabulary after adding these 3 character n-grams? You may not compute the final numerical answer, just leave in terms of the numerical expression. [1+2+2 marks]

2. Suppose, you give the following input to your transformer encoder: {thinking, machines} The input embeddings for these two words are **[0,1,1,1,1,0] and [1,1,0,-1,-1,1]**, respectively. For the first attention head in your first encoder, the query, key and value matrices just take the 2 dimensions from the input each. Thus, the first 2 dimensions define the query vector, next 2 the key vector, and the final 2 the value vector. (a) What are the weights of the query, key and value transformation matrices? (b) What will be the self-attention output for the word 'machines' for this attention head. You are using the scaled dot vector for self-attention. [2+3 marks]

3. Suppose you wish to train an open-domain dialog model that provides a response by (a) first retrieving from a set of template responses (retriever), and (b) then modifying the retrieved response

based on the current context (generator). Assume that you have a large enough training data consisting of dialogs between the users, and you also have a list S of template responses (~500). However, no ground truth about the selected template response is available. What would be your strategy during training time, so that you can train both the retriever and the generator?

To be concrete, assume you have [U1, S1, U2] as the context and S2 be the system response (available during training time). [3+2 marks]

4. Suppose you have a document collection with 4 documents. The set of words present in the document are shown below. Now, you receive a query, "whale boat water". What will be the score for each of these 4 documents as per tf-idf weighting scheme? Use the simplified scoring scheme discussed in the class. No stemming. [5 marks]

| |
|---|
| Doc 1: whale, sea, sea, whale, boat, boat, boat, boat, boat |
| Doc 2: whales, sea, sea, water |
| Doc 3: whale, water, water, whale, whale |
| Doc 4: whales, whales, whales |

5. Suppose you are pretraining a BERT model with 8 layers, 768-dim hidden states, 8 attention heads, and a sub-word vocabulary of size 40k. Also, your feed-forward hidden layer is of dimension 3072. What will be the number of parameters of the model? You can ignore the bias terms, and other parameters used corresponding to the final loss computation from the final encoder representation. The BERT model can take at most 512 tokens in the input. If you instead wish to pretrain the GPT model with the same configuration but allow at most 1024 tokens in the input, what will be the number of parameters? [5 marks]

6. Suppose you trained a 2-layer (on top of word embeddings) forward LM as well as a 2-layer backward LM for ELMo. Assume that the 3-dimensional embedding for a word *w* be [0.2 0.4 -0.5], and the contextual representation for the first and second layer be [0.2 0.4 0.7] and [0.3 0.6 0.5] (forward LM), and [0.2 -0.4 0.7] and [0.3 0.6 -0.5] (backward LM), respectively. Assume that the softmax normalized mixture-model weights for the embedding, 1st and 2nd layers are 0.2, 0.3 and 0.5, respectively (for your particular task). What would be the ELMo representation that you can use for the word *w*? What can you say about the task? [4 marks]

7. Suppose you are using BERT for reading comprehension based question answering. Suppose your input paragraph is, "You can ignore the bias terms", and assume that each individual word is part of the vocabulary and the underlined span is the ground truth answer. For the sake of simplicity, assume that you are working with 2-dimensional hidden states. Let your start and end vectors be [1,-1] and [-1,1], respectively, and the final embedding for the words in the input sentence be [-1,-1], [1,1], [1,2], [2,1], [1,-2] and [2,-1], respectively. (a) If this was a sentence in the training, what would be the corresponding loss for this sentence? (b) If this was a sentence during inference, how many spans would you have to consider? (c) Assume that it is given that the span can have at most 2 words. What will be the output of the model at inference time? [3+1+3 marks]

8. ***Answer the following questions*** [16 marks]

a. Suppose you are using BERT-base for topic classification of documents with 5 classes. How many task-specific parameters will you need to use? On the other hand, what if you were using BERT-base for language identification in code-switched dataset, how many task-specific parameters will need to be used? Assume that there are two languages, English and Hindi, using the same Roman script, and each word belongs to exactly one of these languages. [2 marks]

b. Suppose you are pretraining BART model and your sentence is "Thank you for inviting me". What would be the input-output for the.model during pre-training if you use token deletion for the token "you" and "inviting"? You can show this pictorially by taking the required blocks and naming these. [2 marks]

c. Suppose that for named entity recognition, you encounter a sentence as follows

[$_{\text{LOC}}$ **Mt. Sanitas** ] is in [$_{\text{LOC}}$ **Sunshine Canyon**]

Now, the tokenizer gives you the following output for this sentence:

```
'Mt', '.', 'San', '##itas', 'is', 'in', 'Sunshine', 'Canyon'
```

How would you present the sentence for training based on the BIO tagging scheme? [2 marks]

d. For reading comprehension, suppose a particular question has 3 reference spans as answers by three annotators: "bias", "ignore the bias terms" and "bias terms". What will be the evaluation scores for a model generated output, "ignore the bias". [2 marks]

e. Suppose you are using BLEU score to evaluate three candidate dialog generations: A. "rained last night", B. "it rained", C. "it rained yesterday" for the ground truth, "it rained last night". What can you say about the BLEU scores for A, B and C if you only use unigrams (provide the answer in increasing order of BLEU scores). [2 marks]

f. Suppose you have access to GPT-2 parameters while GPT-3 is only available via an API. You want to teach these models to convert a decimal number into binary. Explain briefly how you will do this, clearly mentioning if it is the training step / fine-tuning step, etc. and what will be the input-output. You can make use of 3 examples to teach and the model should give output for a new input. [2+2 marks]

g. Suppose you are using BERTScore to evaluate a candidate dialog generation, "it rained last night" for a ground truth, "it poured yesterday". After you pass these sentences through BERT to obtain the contextual representations, the cosine similarity between the representations is given below.

| | it | poured | yesterday |
|---|---|---|---|
| it | 0.8 | 0.1 | 0.05 |
| rained | 0.2 | 0.9 | 0.1 |
| last | 0.1 | 0.3 | 0.7 |
| night | 0.1 | 0.2 | 0.6 |

Assume that IDF for the words, {it, rained, last, night, poured, yesterday} be {1.1, 2.8, 2.1, 1.9, 4.5, 2.7}, respectively. What will be the BERTScore? [2 marks]

9. ***Answer the following questions.*** Be very brief and clear in your answers / reasoning. [18 marks]
   a. During encoder-decoder training for RNNs, teacher forcing is typically used. However, during inference, the last generated token(s) might be incorrect, leading to a completely unfamiliar situation for the decoder. What can be a remedy during training time to avoid this unfamiliar situation? [2 marks]
   b. Suppose you want to use Bi-LSTM with Glove embedding for named entity recognition (NER). However, some of the named entity labels do not have enough labeled data but these come with natural language descriptions. What would be the strategy to train your model so that it performs well for this NER task? [3 marks]
   c. In an RNN-based encoder-decoder framework for machine translation, it is typically seen that the decoder gets the context mainly from the last token in the encoder, which is used for encoding the entire input. Suppose, you were not using the encoder-decoder attention, what can be a remedy so that the decoder generates the initial token correctly, and can rely on the language model (decoder) to generate the rest of the sequence? [2 marks]
   d. Suppose you are using RNN encoder-decoder to perform abstractive summarization for a new domain, where most of the tokens are not covered in the general domain embedding (Word2Vec, Glove). You can either train a char-LSTM model to learn embedding for OOV tokens, or use a subword vocabulary using BPE. Which of these should be the correct choice, and why? [2 marks]
   e. While BERT makes use of a bidirectional language model with masked language modeling (MLM), ELMo makes use of a forward and a backward language model. Would it make sense to exchange their language modeling ideas? [3 marks]
   f. What is the objective function used for Skip-gram with negative sampling? Write down the expression, and define each of the terms. Why is this preferred over Softmax? [2+1 marks]
   g. Can you describe the connection between word2Vec and Glove objectives? Give a brief intuitive explanation, followed by describing the connection mathematically. [1+2 marks]