UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11157 NATURAL LANGUAGE UNDERSTANDING, GENERATION, AND MACHINE TRANSLATION

Thursday 12 th May 2022

13:00 to 15:00

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **This is an OPEN BOOK examination.**

MSc Courses

Convener: A.Pieris
External Examiners: A.Cali, V.Gutierrez Basulto.

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **Attention and Transformers**

(a) Assume a self-attention layer that consists of three input words, each of which is represented as a three-dimensional word embedding. For the following input vectors, compute the output $\mathbf{y}_2$ of this layer:

$$\mathbf{x}_1 = [0, 1, 0.5] \quad \mathbf{x}_2 = [0, 1, 0] \quad \mathbf{x}_3 = [1, 0, 0]$$

Assume that the layer doesn't use any parameter matrices (Q, V, K). Give the formulae that you use for your computations. [*4 marks*]

(b) Using an example, explain what is meant by the statement "self-attention layers are permutation equivariant" (in your explanation, you can refer back to the previous question). When is permutation equivariance a desirable property in NLP? [*2 marks*]

(c) Describe how you would fine-tune BERT for the following tasks:

i. Sentiment analysis

ii. Named entity recognition

In each case, a brief answer is expected, but it should include a description of the input and output representations. [*3 marks*]

(d) *Text classification* is the task of taking a text and assigning it a label denoting the text type. Here, the texts you are dealing with are novels, and the labels you assign denote genres (crime, horror, romance, science fiction, etc.).

Your task is to design a model that performs text classification. Your training data consists of a large collections of texts with genre labels. The texts are long (in the order of 200 pages), and are subdivided into chapters, paragraphs, and sentences.

You also have a pre-trained transformer-based language model such as BERT available.

A brief answer is expected for each part, in a few sentences or less, stating the main ideas at a high level.

i. Which issue makes the use of BERT for this task difficult? [*1 mark*]

ii. Assume that you nevertheless want to use BERT to encode aspects of the texts in your training data. How would you achieve this? This should involve partitioning each text into appropriate subunits. [*2 marks*]

iii. How could you combine the BERT encodings for subunits to obtain a representation for the whole text? [*3 marks*]

iv. Once you have a text representation, how can you use it to predict genre labels? Describe an architecture and how you would train it. [*2 marks*]

2. **Translation and Multilinguality**

You are employed by a company which provides call center services. The calls are monitored manually for training and quality assurance purposes but this is very time consuming and they can only monitor 1% of the calls. You want to deploy an automated quality assurance service which can cover 100% of the calls. An important quality check is to make sure that the staff ask customers something similar to "Is there is anything more I can do for you?", before saying goodbye. Your first goal is to detect calls where the staff have asked customers this important question.

a) You decide to use a large pretrained language model, BERT, and fine tune it for the task. You can treat the task as a classification problem. How do you define the input and output data for fine tuning? Only use a sentence or two to describe them. [*2 marks*]

b) Your company decide they want to track the topic of each utterance. You decide that you will try to model this using prompts. How would you set up the prompting task? Please list the paradigm/architecture of the pre-trained model and describe the prompting function in terms of $[X]$ representing the input, and $[Z]$ the answer. Describe what $[X]$ and $[Z]$ consist of with an example. [*3 marks*]

You are wanting to deploy automated quality assurance across the 40 languages your company supports and you want to train a multilingual translation model to translate all conversations into English. Before preprocessing the data for training your translation model, you have to carefully consider what model vocabulary you are going to use.

c) What are two advantages of using byte pair encoding (BPE) dropout over BPE? [*2 marks*]

d) One of the languages you want to support is Nepali, and it is very low resource. Name two ways you can use monolingual Nepali text to improve the quality of the translation model for Nepali into English. [*2 marks*]

e) Consider the following translations:

Human Reference:	"Would you like an upgrade ?"
Hypothesis A:	"Do you want to be upgraded ?"
Hypothesis B:	"Would you like a sweetie ?"

i. The BLEU score is based on n-gram precision scores. Give the 1-gram and 2-gram precision of each hypothesis. [*3 marks*]

ii. How does the length of the hypothesis affect the BLEU score, describe this in terms of the reference $|r|$ and hypothesis $|h|$ length? [*2 marks*]

iii. What is a disadvantage of BLEU as an evaluation metric for MT output? [*1 mark*]

iv. What would be one advantage of BLEU over a trained metric like COMET? [*1 mark*]

3. **Question Answering and Semantic Parsing**

You are working on a Natural Language User Interface (NLUI) for which you need to develop different language understanding and generation modules.

a) The NLUI provides a short Natural Language (NL) description for an entity in response to a question asked by a user, e.g., *Tell me about the Louvre Museum*. The NLUI relies on a big Knowledge Graph (KG) to store factual information about entities which it then uses to generate answers to such questions.

Once the NLUI has parsed the question, it selects a sub-graph of the KG around the requested entity which is large enough to include potentially relevant information about the entity. You can think of this sub-graph as a set of triples $\{t_1, \cdots, t_m\}$ where each t_i is of the form (s_i, r_i, o_i) where s_i and o_i are subject and object entities of binary relations r_i. Below is an example triple set with two KG triples:

{(`Louvre Museum`, `director`, `JL Martinez`), (`Louvre Museum`, `location`, `Paris`)}

You have decided to use a neural sequence-to-sequence generation approach to build a model that generates NL answers from KG triples. You have a dataset of (set of triples, NL description) pairs for training.

i. How would you represent the input to the encoder of your sequence-to-sequence model (i.e., the KG triples)? Describe your representation and discuss the benefits and downsides of your solution. [*4 marks*]

ii. In order to model the content selection sub-task in Natural Language Generation (NLG), you will need to condition the output NL description on the input KG triples. Describe a mechanism for achieving this, using formulas or a drawing. Assume the encoder you have chosen in the previous question and an LSTM decoder. [*4 marks*]

iii. In this conditional generation task, which decoding strategy, beam search or top-n sampling with a large n, would be more appropriate and why? [*2 marks*]

b) In addition to generating entity descriptions from KG factual information, you want to extend the NLUI to generate short entity descriptions from long documents related to entities. You have already trained an extractive summarisation model which, given a document D_i associated with an entity and consisting of sentences $(s_1, \cdots, s_l)$, will extract a subset of $k, k < l$ summary sentences. One issue with extractive summaries is that the extracted sentences may still contain unwanted details. What could you do to render each of the extracted summary sentences shorter and more general? Do you need to collect specific training data for this (justify your answer)? [*2 marks*]

c) To make the NLUI more robust to language variation, you were told that you can take advantage of paraphrases.

i. Why would paraphrases help the NLUI's semantic parsing component? [*2 marks*]

ii. You have a dataset with (NL question, logical form) pairs which you want to extend with paraphrases. You apply different techniques to generate a set of candidate paraphrases for each question. You then want to take the top k best candidate paraphrases using a semantic-based score and taking advantage of pre-existing embeddings such as GloVe or BERT. Describe a simple approach to rank a set of candidate paraphrases for a given question. [*3 marks*]