UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11157 NATURAL LANGUAGE UNDERSTANDING, GENERATION, AND MACHINE TRANSLATION

Tuesday 4 th May 2021

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: W.Knottenbelt, M.Dunlop, E.Vasilaki.

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **Neural Networks**

   (a) Suppose you have a neural network that is overfitting to the training data. Name **two possible ways** to fix this situation. [*2 marks*]

   (b) You are training a neural network and watching the loss function on the training set over epochs. Rather than decreasing, it seems to fluctuate around where it started. What is one change you could make to your training procedure that could fix this? [*2 marks*]

   (c) Assume you have to chose either an RNN or a CNN model. Say which one you would prefer for each of the following tasks, and give a short explanation (not more than one sentence per task).

      i. document classification
      ii. part of speech tagging
      iii. image caption generation
      iv. spam detection

      [*4 marks*]

   (d) List three problems associated with conditional language models when applied to tasks such as summarization or data-to-text generation. [*3 marks*]

2. **Machine Translation**

   You are working for a company with a large call centre in the UK and they want to support customers all around the globe by using machine translation. They notice that there are many translation errors with pronouns referring to the wrong person, gender or number that cause serious difficulties for the support staff. You are brought in to solve the problem.

Customer: I bought a gizmo 2000.
Customer: It is not working.
Customer: Can you help me?
Staff: Does it have batteries?
Customer: No, they did not come in the box.
Staff: I am afraid it won't work without them.

   Table 1: An example conversation. The customer does not speak English and the translation model must translate their utterances into English, and the staff utterances into the other language. Pronouns are underlined and are often translated incorrectly.

   (a) The translation model is an encoder-decoder model which uses a simple RNN on the source and target. One of the team claims that because RNNs model an unlimited history, if you translate the entire context of the conversation for each utterance up to that utterance, the model will be able to guess which pronouns to use. Name two reasons why this might not be the case. [*2 marks*]

(b) You change the model to include all the dialogue history and you know that LSTMs can model long-distance context better, so you try swapping the simple RNN for an LSTM. What are potential disadvantages of the LSTM over the RNN? [*2 marks*]

(c) You notice that there are still problems with pronouns and you decide that you will use attention to solve these problems. Immediately the speed of your translation system slows down drastically for the longer conversations. Can you explain why this happens? [*2 marks*]

(d) In the interest of efficiency, you consider alternative solutions that do not use attention and do not include the context of previous sentences. What linguistic information could you provide to the encoder which might help the decoder predict the correct pronouns? How could you include it in the model? [*3 marks*]

(e) If your training corpus had the following words and frequencies, what would the vocabulary be if you used byte pair encoding (BPE) and set the number of BPE merges to 5? Please do not include end of word symbols in your answer. [*2 marks*]

Types	Token counts
it	4
them	7
her	2
the	10

Table 2: Words in training corpus with their frequencies.

(f) Your new contextual machine translation model does not report better BLEU scores and you notice that although pronouns are key to the quality of the translations, they make up a tiny percentage of the number of words in the target language. How could you design an automatic evaluation metric which captures the model's performance on pronouns? [*3 marks*]

3. **Parsing**

In the lectures, we discussed the unsupervised parsing model proposed by Cao et al. in *Unsupervised Parsing via Constituency Tests*. The key idea in this paper is to determine whether a span of words is a constituent by combining syntactic transformations with a language model that judges if the output of a transformation is grammatical. Figure 1 lists the transformations employed by Cao et al.

(a) Apply the **transformations** in Figure 1 to following sentence:

Name	Applied to "A [B] C"	Example
Clefting	it {is, was} B that A C	*it {is, was} the london market that by midday , was in full retreat*
Coordination	A B and B C	*by midday , the london market and the london market was in full retreat*
Substitution	A {it, ones, did so} C	*by midday , {it, ones, did so} was in full retreat*
Front Movement	B , A C	*the london market , by midday , was in full retreat*
End Movement	A C B	*by midday , was in full retreat the london market*

Figure 1: Transformations used for constituency tests by Cao et al.

(1) The American space agency [is expected to release] new photos from its Perseverance rover shortly.

Here, the brackets indicate the span whose constituency you should test. Give the output of each transformation for the sentence in (1). [*3 marks*]

(b) Now judge the output of each transformation in the previous example as grammatical or not. Is the bracketed span as constituent or not? [*2 marks*]

(c) A major component of the Cao et al.'s approach is a **grammaticality model.** It takes as its input a sentence and outputs a score between 0 and 1, indicating whether the sentence is grammatical. Let's assume you want to train such a model by fine-tuning BERT, and you have a small corpus of sentences annotated with grammaticality judgments as your training data. Describe what the input to BERT would need to look like, and what changes you need to make to the BERT architecture (if any) to fine-tune on this dataset. [*3 marks*]

(d) Cao et al. train and test their model only for English. However, low resource languages are often cited as a motivation for unsupervised parsing. Let's assume you want see if the Cao et al. model works for Malay, an Austronesian language with 290 million speakers. Describe what you would need to do to achieve this. You can assume you can harvest a substantial amount of Malay text from the web. [*3 marks*]

4. **Recognizing Textual Entailment**

The ability to determine the semantic relationship between two sentences is an integral part of machines that understand and reason with natural language. Recognizing textual entailment (RTE) is the task of determining whether two natural language sentences are (a) contradicting each other, (b) not related, or whether (c) the first sentence (called premise) entails the second sentence (called hypothesis). Examples of sentence pairs and their relations are given below:

Premise	**Relation**	**Hypothesis**
A couple walk hand in hand down a street.	Entailment	A couple is walking together.
A couple walk hand in hand down a street.	Contradiction	A couple is sitting on a bench.
A couple walk hand in hand down a street.	Neutral	The couple is married.

(a) You first want to tackle this problem by building a standard classifier (e.g., Naive Bayes). What types of (symbolic) features would you use? List **three features** you would use and provide a justification for coosing them. [*3 marks*]

(b) Suddenly, you realize you could have used neural networks for this task! A colleague has also told you about *word embeddings* like word2vec and BERT. Using these pre-existing embeddings, describe how you would represent the premise and hypothesis in neural space? Also, you don't want to use Naive Bayes any longer, but rather a neural classifier. Describe your choice of neural classifier and the input it would expect. [*4 marks*]

(c) Another colleague tells you about a dataset she recently discovered which contains annotations for two million premise-hypothesis pairs and their relations. Now you want to use LSTMs to represent the premise and hypothesis. How would you encode the hypothesis so that it is conditioned on the premise? Justify your answer and use formulas or a drawing to illustrate your model. [*4 marks*]

(d) Can you improve your model by incorporating an attention mechanism? Why would this make sense for your task? Describe which changes are necessary to your model. [*3 marks*]