IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2019

MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the
Associateship of the City and Guilds of London Institute

PAPER C490H

NATURAL LANGUAGE PROCESSING

Monday 18th March 2019, 10:00
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators not required

1 a In prediction-based distributional models of meaning such as skip-gram in word2vec, a technique called "negative sampling" can be used for training.

i) Describe how this technique works.

ii) What is the main advantage of this technique?

iii) Explain the intuition behind the loss function used when training with negative sampling.

b Consider the architecture of a network for a skip-gram model with 1 hidden layer of 500 dimensions and a vocabulary of 30,000 words. Assuming a softmax output layer, with training done using cross-entropy as loss function, what will be the size of the output layer?

c Given the following training corpus:

```
Mary plays the piano
John got tickets for the play
We saw Mary playing games
John games the system
```

i) Construct a bigram language model for this corpus. Mention all pre-processing techniques you applied to the text before building the language model.

ii) What is the perplexity of your bigram language model for **each** of the following test corpora (where each sentence is a test corpus)? Note: you do not need to solve the equation, just indicate it as part of your answer.

```
John plays games
We saw Mary playing games
```

iii) Discuss a strategy to mitigate issues with zero-count bigrams.

d Consider the following training corpus:

```
Mary/PROPN plays/VERB the/DET piano/NOUN
John/PROPN got/VERB tickets/NOUN for/PREP the/DET play/NOUN
We/PRON saw/VERB Mary/PROPN playing/VERB games/NOUN
John/PROPN games/VERB the/DET system/NOUN
```

i) Create a Hidden Markov Model POS tagger, showing both tables in your answer.

ii) Apply the Hidden Markov Model above to assign the most likely sequence of tags to the following test sentence: `John saw Mary gaming the piano`.

The four parts carry, respectively, 30%, 10%, 35%, and 25% of the marks.

2 a Consider a Convolutional Neural Network for the task of sentiment analysis with these hyperparameters: word embedding dimensionality = 50, 1d convolution layer with 50 filters, window size = 3 and ReLU activation, a max-pooling layer and an output layer. Given the following pre-processed training corpus:

```
excellent !
you are doing very well .
i can 't say job well-done .
```

and their respective labels: `[1, 1, 0]`

i) Provide the filter size and output shapes of each layer of the network for the second example from the training corpus.

ii) Explain how predictions are computed given the output of the network.

iii) Given the provided window size: is padding necessary for any of the training examples? Provide an explanation.

iv) Consider the following test example: `you had your job done`. Which additional pre-processing could be applied to the training corpus to decrease the number of unknown words for this test sample?

b Consider the following two types of training data for the binary task of cancer detection:

(a) Clinical notes in English with manually annotated medical terms. A training example with annotated terms in bold is: `She has completed nine cycles of` **`chemotherapy`**`, has` **`metastatic disease`** `and chronic daily` **`headaches.`**

(b) Clinical notes in German (no term annotations provided). A training example is: `Er hat zwei Knochenmetastasen im Rücken und bekommt keine Chemo`. Note that part of `Knochenmetastasen` resembles the English word `metastastic` (it literally means `bone metastasis` written in one word), and `Chemo` resembles the English word `chemotherapy`.

i) Consider counts of annotated terms as pre-defined features. Knowing that not all the terms are equally important for the diagnosis (for example, `headaches` are less important than `chemotherapy`), choose one type of classifier (from the ones seen during the course) that would be the suitable for data of type (a) and explain why.

ii) Assume that training data of type (a) is used to complement training data of type (b). Considering the linguistic proximity of English and German, explain which pre-processing and classifier type (from the ones seen during the course) would be suitable for data of type (b).

The two parts carry, respectively, 55% and 45% of the marks.

3 a i) Describe a Recurrent Neural Network (RNN) language model and provide the basic equation for updating the hidden states.

ii) Assume you have access to pre-trained word embeddings. Indicate how you would use them to initialise an RNN. Would this be beneficial? Explain why.

iii) Based on the transformer sequence models you saw in the course, how would you introduce self-attention to an RNN language model? What properties will a self-attention-based RNN exploit that are different from the standard RNN?

b i) Describe in brief the role of attention in neural machine translation.

ii) Describe a simple form of attention that uses a target query vector $q \in \mathcal{R}^d$ to attend over hidden states $\{h_1, \cdots, h_t\} \in \mathcal{R}^d$.

c i) Name a popular loss function used in an RNN language model.

ii) Briefly describe a metric to evaluate an RNN language model.

The three parts carry, respectively, 50%, 35%, and 15% of the marks.