IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2020

MSc in Computing Science (Specialist)
MSc in Artifical Intelligence
MSc in Advanced Computing
MEng Honours Degrees in Computing Part IV
MEng Honours Degree in Mathematics and Computer Science Part IV
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 C490

NATURAL LANGUAGE PROCESSING

Friday 20th March 2020, 14:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 a Explain the difference between the original formulation of the skip-gram model trained using a softmax output layer and the skip-gram approximation using the negative sampling technique.

b Given the following training corpus:

```
Mary plays the piano.
John got tickets for the play.
Play the game!
```

i) To build an n-gram language model, which processing techniques would you apply and why?

ii) Construct a bigram language model for this corpus, showing any resulting tables.

iii) What is the perplexity of your bigram language model for the following test sentences? Note: you do not need to solve the equation, just indicate it as part of your answer.

```
John plays games!
```

c Given the following probabilistic grammar and lexicon:

Lexicon:

```
N -> John (0.4)
N -> Mary (0.2)
N -> piano (0.4)
V -> saw (0.5)
V -> playing (0.5)
D -> the (1.0)
```

Grammar:

```
S -> NP VP (0.5)
S -> N VP (0.5)
VP -> V N (0.4)
VP -> V NP (0.6)
NP -> D N (0.5)
NP -> N VP (0.5)
```

i) Provide the CKY parse matrix for the following sentence `John saw Mary playing the piano`.

ii) Extend the grammar and lexicon with rules such that it is able to cover the following sentence (do not worry about giving or updating the probabilities, just show the new rules):

```
John and Mary saw Jane playing games.
```

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

2 a Consider a neural network (any type) for a binary classification task where the goal is to categorise sentences as having positive or negative sentiment.

i) Which modifications to the prediction computation would you need to do in order to predict a real-valued sentiment score in the range of -1 (most negative) to 1 (most positive) (i.e. not binary labels)?

ii) What would a possible loss function for this model be?

iii) Mention two possible evaluation metrics for the predictions generated by the model and explain what they are measuring.

b Consider a convolutional neural network for the task of sentiment analysis with these hyperparameters: word embedding dimensionality = 50, two 2d convolution layers with 10 filters each and window sizes 2 and 5 respectively, ReLU activation, a max-pooling layer, a concatenation layer and an output layer. Given the following training corpus:

```
The movie makes me feel happy!
I was bored.
A nice US comedy.
```

and their respective labels (1 – positive sentiment; 0 – negative sentiment): `[1, 0, 1]`
And the following test examples:

```
This movie made us happier.
It was very boring.
```

i) Which pre-processing could be applied to the training corpus to decrease the number of unknown words for the given test examples?

ii) Which option – lowercasing or truecasing – would you find more effective in this case and why?

iii) Given the provided window size, is padding necessary for any of the training examples? Why?

iv) Provide the filter sizes and output shapes of each layer of the network for the third example from the training corpus.

v) How does the ReLU activation modify its input values? What is the advantage of using this activation?

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

3 a i) Consider the probability of word w_k given the history i.e. $P(w_k|w_1^{k-1})$. Describe how the second term w_1^{k-1} is represented by an n-gram feed-forward language model (LM) such as the one covered in the lecture. Compare this to a recurrent LM in terms of the markov assumption.

ii) Describe how the back-propagation through time (BPTT) works in the context of a recurrent neural LM and compare it to the classical back-propagation algorithm.

b Consider that we have a recurrent encoder-decoder NMT **without** attention mechanism. Let us denote by $E()$ and $D()$ the encoder and the decoder layer, respectively. The dimensions of source and target language embeddings are set to 600. The encoder and the decoder are both unidirectional RNNs with hidden dimensions set to 400 and 500, respectively. The choice of RNN variant does not matter here. You are given a source sentence $X = \{x_1, \ldots, x_S\}$ and a target sentence $Y = \{y_1, \ldots, y_T\}$.

i) Assume that $P()$ denotes the output probabilities from the decoder of your NMT. Give the equation of the cross-entropy loss for the ground-truth Y sequence.

ii) Assume that we encode the sentence X with the encoder E and obtain H with $H = E(X)$. Describe the size/shape of H and the dimensionality of each of its elements.

iii) How would your answer to question **(ii)** change if the encoder is now **bidirectional**?

c Consider that we are given the encoder-decoder NMT from the question **(3b)** and we already encoded the sentence $X = \{x_1, \ldots, x_S\}$ into H by using the encoder.

i) Describe the purpose of adding an attention mechanism to this architecture.

ii) Assume that the hidden state of the decoder at timestep t is d_t. Explain (via text or equations) how the weighted context c_t is computed by the "dot attention" method using d_t and the encoder states H.

The three parts carry, respectively, 30%, 40%, and 30% of the marks.

4 a Assume processing a large corpus to extract word representations.

i) Discuss how context is taken into account by the following approaches: word2vec skip-gram model (original formulation), n-gram language model, recurrent neural net language model, and transformer-based embeddings.

ii) Which of these representations is more powerful and why?

b In a convolutional neural network used for text classification, describe the purpose of the convolution and pooling layers, as well as the purpose of the dropout mechanism.

c Consider the following hidden-markov model (HMM) for part-of-speech tagging:

	mary	play	the	piano	john	get	ticket	for	we	see	game	system
PROPN (4)	0.50				0.50							
VERB (5)		0.40				0.20				0.20	0.20	
DET (3)			1.00									
NOUN (4)				0.25			0.25				0.25	0.25
PREP (1)								1.00				
PRON (1)									1.00			

	PROPN	VERB	DET	NOUN	PREP	PRON	</s>
<s> (4)	0.75					0.25	
PROPN (4)		1.00					
VERB (5)	0.25		0.40	0.40			
DET (3)				1.00			
NOUN (4)					0.25		0.75
PREP (1)			1.00				
PRON (1)		1.00					

Apply the Viterbi model to part-of-speech tag the following sentence. As part of your answer, draw the matrix indicating the probabilities for each state, as well as give the final tags for each word: `John saw Mary gaming the piano`.

d Explain two differences between a transformer-based neural machine translation architecture and an attentive recurrent neural machine translation architecture.

The four parts carry, respectively, 30%, 25%, 25%, and 20% of the marks.