SEMESTER 2 FINAL ASSESSMENT 2020/21

MODULE TITLE Natural Language Processing

Duration 24 hours, Open Book

Student ID:

This paper is a WRITE-ON examination paper.

You ***must*** write your Student ID on this Page and must not write your name anywhere on the paper.

All answers should be written within the designated boxes in this examination paper and sufficient space is provided for each question.

If, for some reason, space is required to complete or correct an answer to a question, use the "Additional Space" provided on the facing page to the question. Clearly indicate which question the answer corresponds to.

No credit will be given for answers presented elsewhere and without clear indication of to what question they correspond. Blue answer books may be used for scratch – they will be discarded without being looked at.

Answer **ALL** questions**.**

There are **three** questions. Each question is expected to take about 30 minutes to answer.

Question	Mark	*Arithmetic checked*	*Double Marked*
Total:			

Question 1

Sequence processing algorithms often make use of sparse embeddings to measure vector similarity between terms and documents for NLP applications (e.g. speech to text). The choice of what constitutes a term and a document is based on the needs of the downstream NLP application, but using term-document and term-term matrices populated with occurrence frequency information is standard.

Below is a term document matrix and term co-occurrence matrix from a NLP application. Documents are defined as books and terms are words within those books.

Term document matrix	David Copperfield	A Study in Scarlet	A Tale of Two Cities	Emma	Middlemarch
sherlock	0	51	0	0	0
treat	91	8	26	40	79
enough	154	34	71	129	237
astonishment	15	9	0	6	6
antagonist	0	1	0	1	0
trifles	7	1	2	2	3

Term co-occurrence matrix	holmes	style	enough	his	your
sherlock	43	0	0	0	0
cavalier	0	1	0	1	0
and	3	0	1	34	3
money	0	0	2	1	1
purse	0	0	0	1	0

(a) Provide a worked example of manually calculating the TF-IDF score for the token "astonishment" in document "David Copperfield".

[5 marks]

5

(b) Provide a worked example of manually calculating the PPMI score for the token "sherlock" and context "holmes".

[5 marks]

5

(c) Provide a worked example of manually calculating the cosine similarity distance between document "David Copperfield" and document "A Study in Scarlet".

[5 marks]

5

TURN OVER

(d) Describe an NLP application which uses a sparse embedding, and another type of NLP application which using a dense embedding. For both examples, explain the method used to compute the embedding in detail and the main benefit of using that type of embedding.

[10 marks]

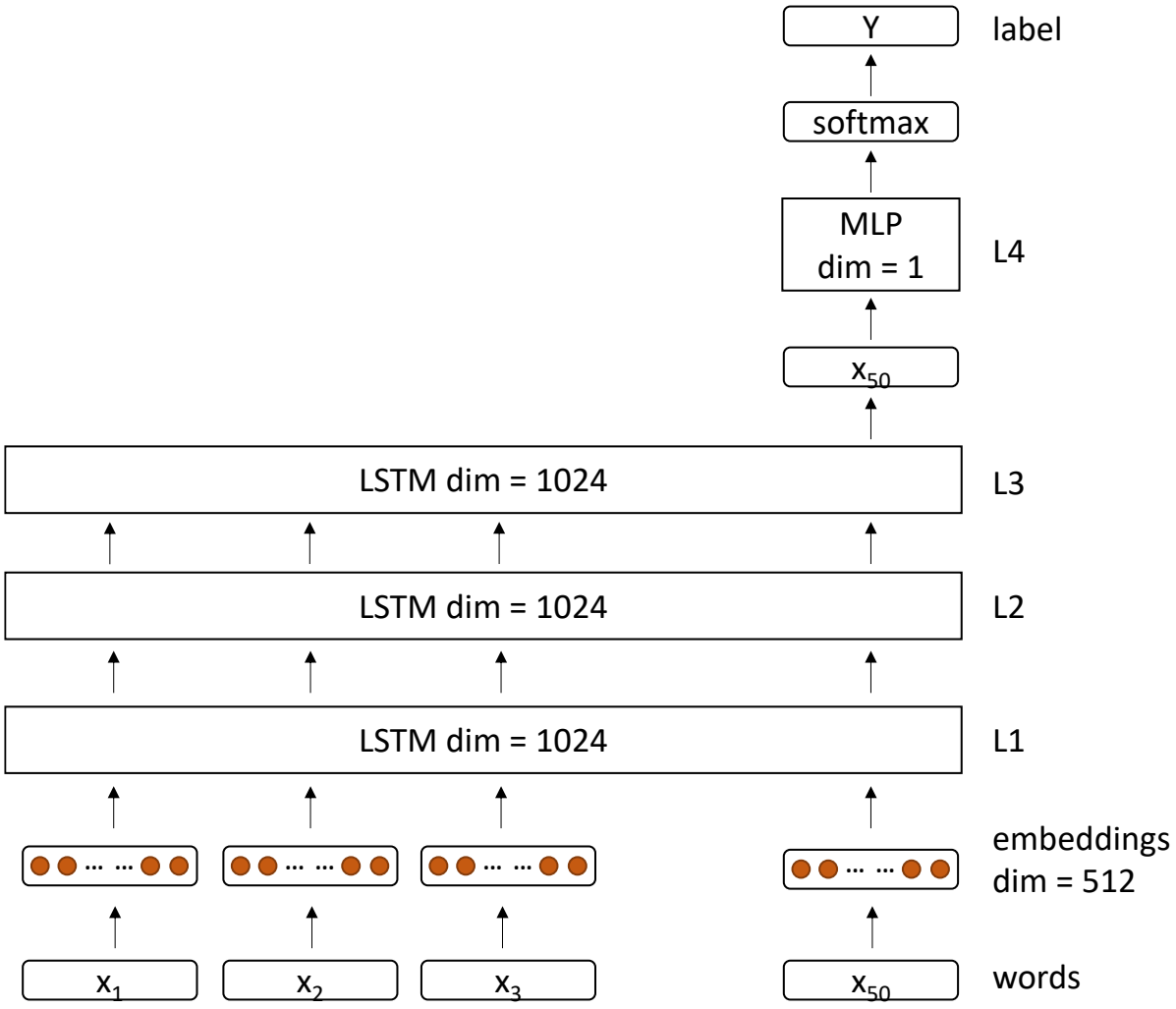
10

$$\frac{}{5} + \frac{}{5} + \frac{}{5} + \frac{}{10} = \frac{}{25}$$

Question 2

Recurrent Neural Networks (RNN) are used in many NLP applications (e.g. Machine Translation). There are many types of RNN, each with its own layered architecture and shape of input and output vectors.

Below is a RNN architecture for an example NLP application.



(a) What class of RNN application is this architecture best suited for? In the context of that class, explain the purpose layer L4.

[5 marks]

5

(b) If (1,512) is the tensor shape of the output of the embedding layer, what is the tensor shape of the output of L4? Provide an explanation of how tensor shapes change through the layers.

[5 marks]

5

(c) Explain why a cross-entropy loss function often used when training a sequence processing architecture.

[5 marks]

5

(d) In a Transformer architecture why are positional embeddings needed? Discuss how you might go about generating positional embeddings, and what the benefits/drawbacks are for different approaches.

[10 marks]

10

$$\frac{}{5} + \frac{}{5} + \frac{}{5} + \frac{}{10} = \frac{}{25}$$

TURN OVER

Question 3

The n-gram family of language models captures information about the likely use of words based on the words that precede it.

(a) N-gram data can be used to calculate probability of text sequences in a language.

i. Given the bigram and unigram occurrence frequencies from the novel Peter Pan in the table below, explain how to derive the Maximum Likelihood Estimation of the sentence `she flew away` .

Unigrams	**Frequency**	**Bigrams**	**Frequency**
she	602	in the	204
peter	400	peter pan	18
wendy	358	flew away	8
away	57	she flew	7
flew	24	away she	1
		<s> she	127
		away </s>	14

Vocabulary size	4854	**Unique bigrams**	26950
Total tokens	48077	**Sentences**	372

[5 marks]

5

ii. N-gram tables are very sparse because specific combinations of known tokens are relatively unlikely to occur in any given text. Discuss the range of techniques that can be used to compensate for missing data when estimating the probability of an unseen bigram.

[10 marks]

10

(b) Skip-grams are a variant of the standard n-gram model.

i. Calculate the set of 1-skip-2-grams from the proverb below:

```
the grass is greener on the other side of the fence
```

[2 marks]

2

ii. Explain how skip-grams and syntactic n-grams extend the concept of n-grams to improve the perplexity of a language model.

[8 marks]

8

$$\frac{}{5} + \frac{}{10} + \frac{}{2} + \frac{}{8} = \frac{}{25}$$

The last page of this style of paper will always be blank and should be completed as below.

This page is intentionally blank

END OF PAPER