SEMESTER 2 EXAMINATIONS 2021-2022

COMP3225 Natural Language Processing

Duration 120 mins (2 hours)

This paper contains 5 questions.

Answer **only THREE** questions in this paper**.**

Answer **ALL** questions from section A.

Answer **only ONE** question from section B.

Only University approved calculators may be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct ‘Word to Word’ translation dictionary AND it contains no notes, additions or annotations.

8 page examination paper

SECTION A

Answer ALL questions

Question A1

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 from an 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	50	0	0	0
treat	90	8	20	40	70
enough	150	30	70	120	230
astonishment	15	10	0	6	6
antagonist	0	1	0	1	0
trifles	7	1	2	2	3

Below is the formula for normalized dot product of vectors v and w, which is commonly used to evaluate vector similarity.

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Question continues on following page

tfidf，点积

(A1.a) Provide a worked example of manually calculating the normalized dot product of the term-document vectors for "David Copperfield" and "A Study in Scarlet".

[5 marks]

(A1.b) Write down the formula for computing Positive Pointwise Mutual Information (PPMI) and describe what PPMI measures. Explain Laplace smoothing and how you would apply it to PPMI.

公式，描述衡量指标（ppt8），拉普拉斯平滑书中第六章91页

[10 marks]

(A1.c) Describe an NLP application which could use a Word2Vec dense embedding. How would you train the Word2Vec embeddings and what is the main advantage of using dense embeddings over sparse embeddings for this application? Give two examples of popular pretrained dense word embeddings used in NLP applications today (other than Word2Vec), and for these word embeddings describe the type of semantic properties that can be discovered within a corpus.

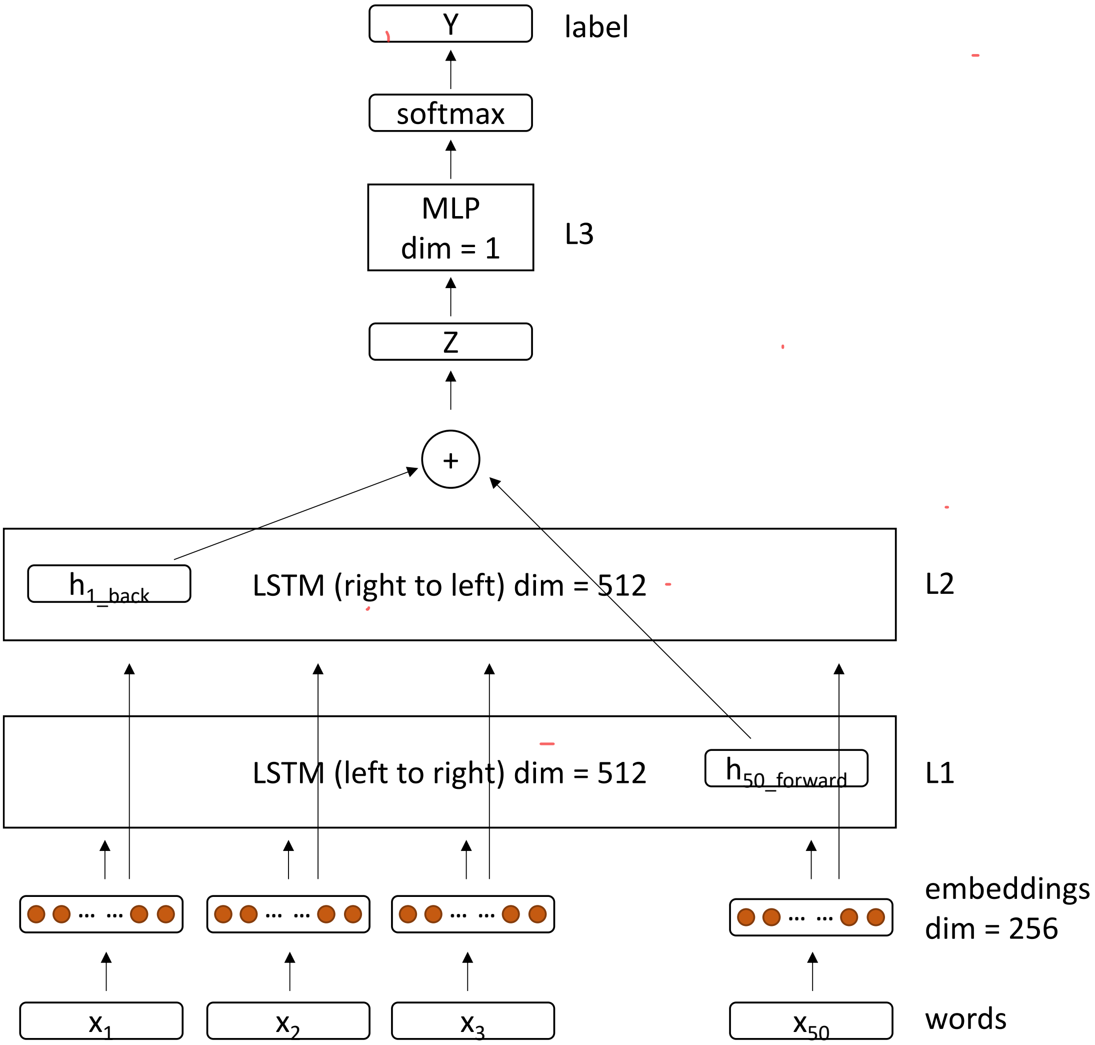
[15 marks]

1.word2vec的训练流程（点积，距离，两种dataset）。2.稀疏和密集的优缺点（ppt p3一句话）3.流行语料库描述

Question A2

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.



Question continues on following page

1.ppt和书p150。PPT中（predict next word）。2. 没答案，（flatten，因为输出是一个label）3.concat 512or（2,256）？张量可能要自学

ppt（RNN） p8

a) Provide a worked example of how a sentence with L words can be encoded as a one-hot vector, using a vocabulary of V words, for input to an RNN.

[5 marks]

(A2.b) What class of RNN application is this RNN architecture best suited for? In the context of that class, explain the purpose of layer L3. If (1,256) is the tensor shape of the output of the embedding layer, what is the tensor shape of Z? Provide a detailed explanation of how tensor shapes change through the layers.

[10 marks]

(A2.c) For an encoder-decoder RNN model describe how the final encoder hidden layer state is used during training. Explain how it changes when max or mean pooling is applied and describe how teacher forcing could be useful for this model. How would you use cross-entropy loss to compute the total loss for the decoder? Attention mechanisms can add value to encoder decoder models, describe in detail how they work and name two types of popular attention mechanism.

[15 marks]

池化层没讲，2. ppt RNN p9. 3. ppt12

SECTION B

Answer *ONE* out of *THREE* questions

Question B1

The University of Southampton wishes to develop an AI question-answering (QA) system to use for student support. It is hoped that it will provide specific answers to student questions based on the comprehensive descriptions of university processes, bureaucracy and regulations that are already available from its internal and external web sites.

1.ppt20 p15。2. ppt20 p16

(B1.a) Describe how question answering evaluations are calculated using mean reciprocal rank and F1, detailing how “ranking”, “precision” and “recall” are interpreted.

[5 marks]

ppt20

(B1.b) Describe what kind of QA system is outlined by this use case with a justification for your choice. Explain the components that feature in this kind of QA system.

[10 marks]

ppt20 p14,但是不完整，书上没有这个数据库，但是有类似的解释在p387

(B1.c) A QA system can be constructed to draw information from a knowledge database such as DBPedia. Provide an architecture diagram for such a model and explain its inputs and outputs. Describe the different types of Knowledge-based QA models available and the different techniques that underpin them.

[15 marks]

Question B2

N-grams are a statistical language model that assigns probabilities to words based on the word(s) immediately preceding them.

ppt05 p11

(B2.a) Given the bigram and unigram occurrence frequencies from the news articles in the table below, explain how to derive the Maximum Likelihood Estimation of the phrase *“the Dorset police officer”.*

Unigrams	Freq	Bigrams	Freq	Trigrams	Freq
government	130	in the	384	in the Dorset	8
Dorset	34	Dorset police	11	the Dorset police	5
police	221	police officer	10	Dorset police officer	6
officer	49	the Dorset	7	in Dorset the	14
the	5651				

Vocabulary size	11269	**Unique bigrams**	56713
Total tokens	92585	**Unique trigrams**	82583
Sentences	4236		

[5 marks]

(B2.b) Perplexity is a way of measuring the predictive power of a specific language model against a specific piece of text, by using the probability that the model gives to that text. Explain the steps needed to calculate the perplexity of a bigram model, using as an example the test sequence the *Dorset police officer* and relevant data from part (a) of this question.

ppt04 p11

[10 marks]

(B2.c) Language models which use 3-, 4-, 5-grams or more suffer from sparsity of training data, because of the combinatorial possibilities of human language. Discuss the range of techniques that can be used to compensate for missing data, when encountering word combinations that were not present in the training set.

[15 marks]

ppt05 一次activity讨论过（女老师）

TURN OVER

Question B3

Information extraction has been a foundation of NLP for decades and allows actionable knowledge to be extracted from free text corpora. Types of information extraction include named entity recognition, relation extraction, temporal extraction, event extraction and semantic role labelling.

ppt07 p6

(B3.a) Provide a worked example of BIO tagging in the context of a named entity recognition application. What is BIOES tagging and how does it differ from BIO tagging?

[5 marks]

ppt17,提了一句
bert在RE也有用

(B3.b) Describe using an architecture diagram how a pre-trained BERT word embedding can be used with a Transformer-based relation extraction (RE) model, and how exactly the input and output is encoded. Describe how named entity recognition (NER) can enhance training performance of this model.

[10 marks]

(B3.c) Provide a worked example of how PropBank argument and modifier annotations could be embedded to train a neural SRL model. How would you change this to use FrameNet? Discuss the differences in approach between neural SRL models and feature-based SRL, highlighting some advantages and disadvantages of each.

[15 marks]

ppt19

END OF PAPER