THE UNIVERSITY OF WARWICK

LEVEL 7 Open Book Assessment [2 hours]

Department of Computer Science

CS918 Natural Language Processing

Instructions
1. Read all instructions carefully and read through the entire paper at least once before you start writing.
2. There are **6** questions. You should **attempt 4 questions**. You should not submit answers to more than the required number of questions.
3. All questions will carry the same number of marks unless otherwise stated.
4. You should handwrite your answers either with paper and pen or using an electronic device with a stylus (unless you have special arrangements for exams which allow the use of a computer). Start each question on a new page and clearly mark each page with the page number, your student id and the question number. Handwritten notes must be scanned or photographed and all individual solutions should (if you possibly can) be collated into a single PDF with pages in the correct order. You must upload two files to the AEP: your PDF of solutions and a completed cover sheet. You must click **FINISH ASSESSMENT** to complete the submission process. After you have done so you will not be able to upload anything further.
5. Please ensure that all your handwritten answers are written legibly, preferably in dark blue or black ink. If you use a pencil ensure that it is not too faint to be captured by a scan or photograph.
6. Please check the legibility of your final submission before uploading. It is your responsibility to ensure that your work can be read.
7. You are allowed to access module materials, notes, resources, references and the internet during the assessment.
8. You should not try to communicate with any other candidate during the assessment period or seek assistance from anyone else in completing your answers. The Computer Science Department expects the conduct of all students taking this assessment to conform to the stated requirements. Measures will be in operation to check for possible misconduct. These will include the use of similarity detection tools and the right to require live interviews with selected students following the assessment.
9. By starting this assessment you are declaring yourself fit to undertake it. You are expected to make a reasonable attempt at the assessment by answering the questions in the paper.

(continued)

Please note that:

- You must have completed and uploaded your assessment before the 24 hour assessment window closes.
- You have an additional 45 minutes beyond the stated duration of this assessment to allow for downloading and uploading the assessment, your files and technical delays.
- For further details you should refer to the AEP documentation.

Use the AEP to seek advice immediately if during the assessment period:

- you cannot access the online assessment;
- you believe you have been given access to the wrong online assessment;

Please note that technical support is only available between 9AM and 5PM (BST).

Invigilator support will be also be available (via the AEP) between 9AM and 5PM (BST).

Notify Dcs.exams@warwick.ac.uk as soon as possible if you cannot complete your assessment because

- you lose your internet connection;
- your device fails;
- you become unwell and are unable to continue;
- you are affected by circumstances beyond your control (e.g. fire alarm).

Please note that this is for notification purposes, it is not a help line.

Your assessment starts below.

---

1. (a) Write **regular expressions** for the following languages.

   i. The set of all upper case alphabetic strings ending in a `N`; [2]

   ii. All strings that have both the word "`happy`" and the word "`family`" in them; [3]

   iii. All strings that start at the beginning of the line with an integer and that end at the end of the line with a word. [2]

   (b) Describe *lemmatisation* and *stemming*. For the following three words, provide their lemmas and stems: [8]
   `having`, `flies`, `are`

   (c) State the difference between *homonymy* and *polysemy* and give an example of each. [6]

   (d) Describe two types of word similarity measurement algorithms. No equation is required. [4]

2. This question is concerned with **Language Modelling**.

   (a) What is N-gram? [3]

   (b) What is smoothing and why do we use it in language modelling? [4]

   (c) How does Laplace smoothing work and what is the problem of using it for language models? [4]

   (d) For a text corpus containing a total of 10,000 word tokens and 7,000 unique words, we have the following counts of unigrams and bigrams:

n-**gram**	rainy	grass	day	rainy day	rainy grass
Count	50	370	620	15	0

   i. Estimate the probabilities $P(day)$ and $P(day|rainy)$ using Maximum Likelihood estimation. [4]

   ii. Estimate the bigram probability $P(grass|rainy)$ using Maximum Likelihood estimation and add-k smoothing with $k = 0.1$. [4]

   iii. Give the formula for calculating the perplexity of a text $(w_1, \cdots, w_N)$ in a bigram model where N is the total number of bigrams. How is perplexity used to evaluate the quality of a language model? [6]

(continued)

3. (a) What are word embeddings and give two reasons why word embeddings are useful. [6]

   (b) Give the one-hot representation of the following two sentences. Assume they are the only sentences in your corpus. No linguistic preprocessing is required. [4]

   ```
   the man saw a cat.
   the cat caught a bird.
   ```

   (c) The evaluation of a text classifier produced the following confusion matrix.

Predicted Class	**Actual Class**		
	A	B	C
A	58	6	1
B	5	11	2
C	0	7	63

   Based on this confusion matrix, compute the following values:

   i. Macro-averaged F1. [4]

   ii. Micro-averaged F1. [4]

   (d) Consider the following Probabilistic Context-Free Grammar (PCFG): [7]

Production rule	**Probability**
S→VP	1.0
VP→Verb NP	0.7
VP→Verb NP PP	0.3
NP→NP PP	0.3
NP→Det Noun	0.7
PP→Prep Noun	1.0
Det→the	0.1
Verb→Cut | Ask | Find	0.1
Prep→with | in	0.1
Noun→envelope | grandma | scissors | men | suits	0.1

   Draw the top-ranked parse tree for the sentence below by applying the given PCFG. Does the result seem reasonable to you? Why or why not?

   ```
   Cut the envelope with scissors.
   ```

4. (a) Describe the main differences between Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM) and the advantage of using MEMM over HMM. [6]

   (b) What is the label bias problem suffered by MEMM? How does Conditional Random Fields address this problem? [4]

   (c) Consider a trigram HMM tagger with:

   - The set $\mathcal{K}$ of possible tags equal to $\{D, N, V\}$
   - The set $\mathcal{V}$ of possible words equal to $\{the, dog, barks\}$
   - the following parameters

$$\begin{array}{ll} P(D|\text{START}) = 1 & P(the|D) = 1 \\ P(N|D) = 1 & P(dog|N) = 0.4 \\ P(V|D,N) = 1 & P(barks|N) = 0.6 \\ P(\text{STOP}|N,V) = 1 & P(dog|V) = 0.1 \\ & P(barks|V) = 0.9 \end{array}$$

   with all other parameter values equal to 0. START and STOP are two special tokens indicating the start and end of a sequence.

   i. Write down an equation for the total probability of a pair of word sequence $\{x_1, \cdots, x_n\}$ and its corresponding tag sequence $\{y_1, \cdots, y_n\}$ using the trigram HMM tagger. Assume $y_0 = \text{START}$ and $y_{n+1} = \text{STOP}$. [5]

   ii. Write down the set of all pairs of sequences $\{x_1, \cdots, x_n\}$ and $\{y_0, y_1, \cdots, y_n, y_{n+1}\}$ such that the following properties hold: [10]
   - $P(x_1, \cdots, x_n, y_0, y_1, \cdots, y_n, y_{n+1}) > 0$
   - $x_i \in \mathcal{V}$ for all $i \in \{1, 2, \cdots, n\}$
   - $y_i \in \mathcal{K}$ for all $i \in \{1, 2, \cdots, n\}$, $y_0 = \text{START}$ and $y_{n+1} = \text{STOP}$

(continued)

5. A social network company runs a platform that enables users to write short messages to each other, attach photographs, and embed links to other websites. There are some concerns that this platform is being used by extremist groups to organise violent demonstrations. As an NLP expert, you have been asked to develop a system that will identify postings containing threats of violence. In this context, answer the following questions:

   (a) What assumptions do you need to make about this task? [4]

   (b) What are the set of features that you will extract from each post? What are the two most important? [6]

   (c) How would you represent each of the features extracted from social media posts? What pre-processing steps would you consider? [8]

   (d) How would you train a classifier for this application? Describe how you build your training set, which classifier you intend to train, how to evaluate the classifier performance and whether there are any factors you need to consider for training the classifier. [7]

---

6. (a) Briefly describe the **gradient descent** method for learning a neural net. Accompany your explanation with a diagram. Your description should include only one equation. Explain all the variables in the equation, and provide a short, intuitive explanation of gradient descent in plain English and why it works. Do not talk about backpropagation or the computation of gradients in your answer. [5]

   (b) Consider a neural net used for binary classification in Figure 1. The output unit has a sigmoid activation function, $\hat{y} = \sigma(z) = \frac{1}{(1+e^{-z})}$, where z is the total input to the unit. Given an input, x, for the neural net, let y be the target output, and let $\mathcal{L}(y, \hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$ be the cross entropy error. [15]

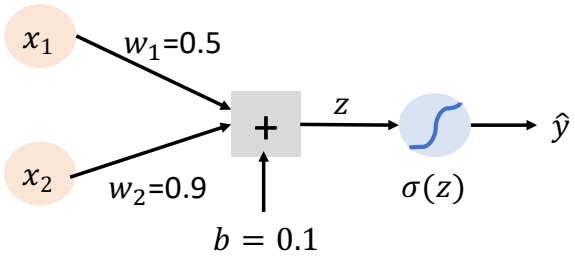

Figure 1: Simple neural network for classification.

Use the backpropagation algorithm to calculate the adjusted weights after the training instance $(x_1 = 1, x_2 = 2, y = 0)$, assuming the predicted output is $\hat{y} = 0.9$ and the learning rate is 0.001. The initial weights and bias values are: $w_1 = 0.5, w_2 = 0.9, b = 0.1$. (**Hint:** Write down the loss function, then write down the gradients for the weights (w_j) and bias b of the neural net, and apply the chain rule.)

   (c) Describe the problems of exploding and vanishing gradients in **Recurrent Neural Networks**. What are the typical solutions to address these problems? [5]

(End)