The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE **Spring** Semester **21-22**

NATURAL LANGUAGE PROCESSING **1 hour 30 minutes**

Answer TWO questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

THIS PAGE IS BLANK

1. a) What is a language model? Describe the data required to train a language model, and what a trained model is expected to return. [20%]

   b) What is the equation for the probability of a sentence? How is this probability approximated in an n-gram language model? Explain the equation and the approximation terms.

   [30%]

   c) What is add-1 smoothing? Why is it important for language modelling? Describe using equations how add-1 smoothing is applied to the bigram language model. [20%]

   d) Language models can be evaluated intrinsically and extrinsically. Discuss the advantages and disadvantages for each approach and describe TWO methods for intrinsic and THREE for extrinsic evaluation. [30%]

2. A Hidden Markov Model (HMM) is a popular approach to automatic part-of-speech tagging.

a) A HMM tagger's estimate of the best tag sequence $\hat{y}$ for a given word sequence $x = \{x_1, ..., x_n\}$ is expressed by the formula:

$$\hat{y} = \underset{y \in \mathcal{Y}^{\mathcal{N}}}{\operatorname{argmax}} P(y|x)$$

where $y = \{y_1, ..., y_n\}$ and $\mathcal{Y}$ is a set of possible tags. This is approximated by assuming:

$$\hat{y} = \underset{y \in \mathcal{Y}^{\mathcal{N}}}{\operatorname{argmax}} \prod_{n=1}^{N} P(x_n|y_n)P(y_n|y_{n-1})$$

Explain using equations how this approximation is derived, including the simplifying assumptions that are used. [30%]

b) Consider the sentence *I play games*. The following counts are observed in a corpus: (i) unigram word/tag counts; (ii) bigram tag counts; and (iii) counts of occurrences of a word with a particular part-of-speech tag (see the corresponding tables below). Here <s> denotes a special start of sentence marker.

	Count
I	7342
play	100
games	253
<s>	50000
VB	148787
NN	253048
PRP	64520

(i) Word/tag counts

	VB	NN	PRP
<s>	789	5783	2304
VB	43	7432	1038
NN	1134	2276	1358
PRP	1492	68	9

(ii) Bigram tag counts

	I	play	games
VB	0	26	82
NN	0	35	171
PRP	7342	0	0

(iii) Tag-word counts

Write down the equations for computing the Maximum Likelihood estimates of $P(x_n|y_n)$ and $P(y_n|y_{n-1})$. Use them to tag the given word sequence by computing $\hat{y}$. You do not need to calculate all of the transition probabilities or the emission probabilities, but you should show how you would calculate one of each.

Note: If you do not have a calculator, you may leave your answer in the form of an arithmetic expression(s) involving integers.

[40%]

c) The Viterbi algorithm is the most commonly used algorithm to calculate the most probable path through a HMM efficiently. Describe the Viterbi algorithm using pseudocode and any auxiliary data structures it employs. [30%]

3. a) What are distributed word representations? How do they compare to one-hot encoding? [20%]

   b) Describe how one can obtain sparse distributed word representations from a corpus by counting word contexts, giving details on how the choices of context window size, context representation and count post-processing affect the representations learned. [30%]

   c) Describe in detail, using appropriate equations and diagrams, the skip-gram model and how it can be used to learn dense distributed word representations from a corpus. [30%]

   d) Give ONE example of intrinsic word representation evaluation and ONE example of extrinsic word representation evaluation. [20%]

END OF QUESTION PAPER