The University of Melbourne

School of Computing and Information Systems

COMP90042
Natural Language Processing
Mock Exam
June 2020

Exam duration: Two hours (1 hour 45 minutes writing time; 15 minutes upload time)

Length: This paper has 4 pages including this cover page.

Instructions to students:

- This exam is worth a total of 40 marks and counts for 40% of your final grade.
- You can read the question paper on a monitor, or print it.
- You are recommended to write your answers on blank A4 papers. Note that some answers require drawing diagrams or tables.
- You will need to scan or take a photo of your answers and upload them via Gradescope. Be sure to label the scans/photos with the question numbers.
- Please answer all questions. Please write your student ID and question number on every answer page.

Format: Open Book

- While you are undertaking this assessment you are permitted to:
  - make use of the textbooks, lecture slides and workshop materials.
- While you are undertaking this assessment you must not:
  - make use of any messaging or communications technology;
  - make use of any world-wide web or internet-based resources such as wikipedia, stackoverflow, or google and other search services;
  - act in any manner that could be regarded as providing assistance to another student who is undertaking this assessment, or will in the future be undertaking this assessment.
- The work you submit must be based on your own knowledge and skills, without assistance from any other person.

Continued overleaf . . .

COMP90042 Natural Language Processing

Semester 1, 2020

Total marks: 40

Students must attempt all questions

Section A: Short Answer Questions [14 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in no more than a line or two.

Question 1: General Concepts [7 marks]

a) For higher order (n $\geq$ 2) N-gram language models, what is the key idea that differentiates more sophisticated "smoothing" techniques from stand-alone add-k smoothing? Mention one smoothing technique which instantiates this idea. [2 marks]

b) What is the vanishing gradient problem in recurrent neural networks? Explain one approach for tackling this. [2 marks]

c) What is discourse? Describe two common discourse applications. [3 marks]

Question 2: Machine Translation [5 marks]

a) Why is "machine translation" a difficult task? Explain with an example. [2 marks]

b) For "statistical machine translation", what is the rationale for decomposing the model into a language model and a translation model? [1 mark]

c) What is the "information bottleneck" issue in "neural machine translation"? Explain one approach for tackling this. [2 marks]

Question 3: Topic Models [2 marks]

a) Compare "Latent Semantic Analysis" and "Latent Dirichlet Allocation", identifying two important commonalities and two important differences. [2 marks]

Continued overleaf ...

Section B: Method Questions [15 marks]

In this section you are asked to demonstrate your conceptual understanding of the methods that we have studied in this subject.

Question 4: Text Classification [6 marks]

For this question, suppose you have a very large corpus of English texts written by people from 20+ different language backgrounds, and you want to build an automatic Native Language Identification system.

a) Name two types of "features" you think would be appropriate for this task and explain why. [2 marks]

b) Given the nature of the task and the features you have chosen, would you perform "lemmatisation" and/or "stop word removal" over your corpus? Explain why or why not for both preprocessing methods. [2 marks]

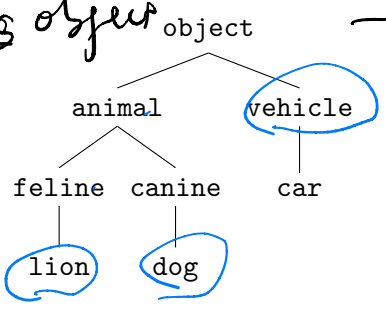
c) Given the task and the features you have chosen, do you think a Random Forest classifier would be appropriate? What about a Support Vector Machine? Justify your answers. [2 marks]

Question 5: Hidden Markov Models [4 marks]

a) Describe the assumptions that underlie Hidden Markov models, and provide a part-of-speech tagging example showing where these assumptions are inappropriate. [2 marks]

b) What classes of formal languages can be described by Markov models over word sequences? Relate this to context free grammars used in parsing. [2 marks]

Question 6: Lexical Semantics [5 marks]



The questions below are based on the partial lexical hierarchy above.

a) Fill in this sentence with the appropriate -nym: animal is a _____ of lion. [1 mark]

b) Based on simple "path-based" similarity, which is more similar to lion, dog or vehicle? What about with the "Wu-Palmer" similarity metric? [2 marks]

c) If we are using "Lin" similarity, is it possible that lion might be more similar to car than it is to dog? If so, show give the condition on the "information content" of dog that must hold (in terms of the IC of other nodes) for this to happen, or, if not, explain why not. [2 marks]

Continued overleaf . . .

Section C: Algorithmic Questions [11 marks]

In this section you are asked to demonstrate your understanding of the methods that we have studied in this subject, in being able to perform algorithmic calculations.

Question 7: Part-of-Speech and Parsing [6 marks]

This question is about using analyzing syntax. Consider the following newspaper headline:

`Eye drops off shelf`

a) First show the key ambiguity in the sentence by giving two possible part-of-speech tag sequences. You can use any existing POS tagset, or your own, provided it satisfies the basic properties of a tag set and is easily interpretable. The tag set you use need not distinguish inflectional differences. [1 mark]

b) Write a set of CFG productions that can represent and structurally differentiate these two interpretations. Your set of non-terminals should consist of S, NP, VP, and your POS tag set from above, and your rules should have no recursion. [2 marks]

c) Do a CYK parse of the sentence using your grammar. You must include the full table. Be sure to convert your grammar to Chomsky Normal Form, and show which productions must be changed. [3 marks]

Question 8: Viterbi Decoding [5 marks]

a) Why is decoding difficult for HMM at test time? Explain this in the context of part-of-speech tagging using a HMM. [2 marks]

b) Perform Viterbi decoding given the sentence `they can fish` and the following emission and transition tables. You should show the full table and the computation steps involved. [3 marks]

	they	can	fish
N	0.4	0.3	0.3
V	0.1	0.5	0.4

Table 1: Emission probabilities

	N	V
$< s >$	0.6	0.4
N	0.3	0.7
V	0.7	0.3

Table 2: Transition probabilities

— *End of Exam* —