The University of Melbourne

School of Computing and Information Systems

COMP90042
Natural Language Processing
Final Exam
Semester 1 2022

Exam duration: 165 minutes (15 minutes reading time + 120 minutes writing time + 30 minutes upload time)

Length: This paper has 4 pages (including this cover page) and 7 questions. You should attempt all questions.

Instructions to students:

- This exam is worth a total of 120 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 paper. 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 (-10% penalty for each unlabelled question, e.g. if 4 questions are unlabelled then you will receive -40% penalty to your exam marks).
- Please answer all questions. Please write your student ID and question number on every 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, 2022

Total marks: 120 (40% of subject)

Students must attempt all questions

Section A: Short Answer Questions [30 marks]

Answer each of the questions in this section briefly. Each answer should be no longer than several sentences.

Question 1: General Concepts [18 marks]

a) What is "smoothing" in "N-gram language models", and why do we need smoothing in N-gram language models? [6 marks]

b) What are the two assumptions in "hidden Markov models" as presented in class, and how do they relate to the model parameters? [6 marks]

c) Compare and contrast "discourse parsing" and "context-free grammar parsing", focussing on their purposes/aims and algorithms. [6 marks]

Question 2: Machine Translation [12 marks]

a) What are the two main approaches to "machine translation" as presented in class? Your answer should describe briefly how the two approaches work and their data requirements. [6 marks]

b) What is "beam search decoding", and why is it used in machine translation? [6 marks]

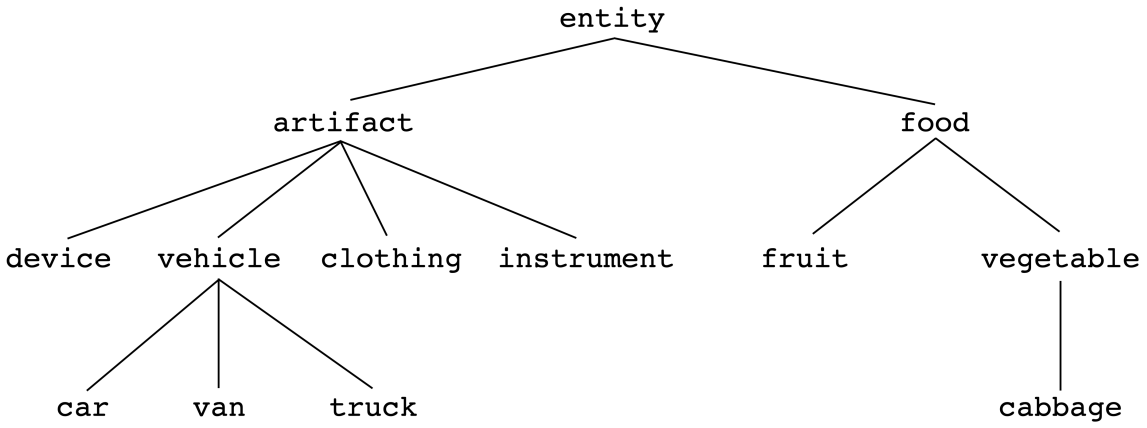
Continued overleaf ...

Section B: Method Questions [46 marks]

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

Question 3: Lexical Semantics [16 marks]

Based on the following "WordNet"-style graph of hypernymy:



a) Compute the "Path similarity" (simpath) for (`car, clothing`) and (`fruit, cabbage`). Compare the similarity values for these two pairs and explain whether they make sense. [3 marks]

b) Compute the "Wu-Palmer similarity" (simwup) for (`car, clothing`) and (`fruit, cabbage`). Compare the similarity values for these two pairs and explain whether they make sense. [6 marks]

c) "Lin similarity" (simlin) uses "information content" to estimate how abstract a sense is. Would this similarity measure produce similarity values for (`car, clothing`) vs. (`fruit, cabbage`) that make more sense? In your answer, you should explain how simlin uses information content and detail any assumptions that led to your conclusion. [7 marks]

Question 4: Finite-State Automata [18 marks]

Sketch out a "finite-state acceptor" diagram for the following languages, assuming the alphabet is $\{a, b\}$. Your answer should specify clearly the starting state and ending state(s).

a) Odd-length strings. [6 marks]

b) Strings that contain both *aa* and *bb* as substrings. [12 marks]

Question 5: Ethics [12 marks]

You're tasked to develop a chatbot (embedded in a toy) that can interact and talk to children. The interactions will be recorded and used as additional training data to improve the chatbot over time. Discuss at least **three** ethical implications of this application.

Continued overleaf . . .

Section C: Algorithmic Questions [44 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 6: Syntax [20 marks]

Consider the following ambiguous sentence:

```
cat guides lost hiker
```

a) Explain the two possible interpretations of the sentence in text. You should ground your interpretations using part of speech. [6 marks]

b) Draw two "dependency trees" for the sentence, illustrating its two interpretations. Be sure to label the root node, and include arrows to denote edge direction. You do not need to provide edge labels. [4 marks]

c) For one of the dependency trees, show a sequence of parsing steps using a "transition-based parser" that will produce it. Be sure to include state of the stack and buffer and the produced arc/edge (if applicable) at every step. [10 marks]

Question 7: N-gram language models [24 marks]

This question asks you to calculate the probabilities for N-gram language models that is applied to characters. You should leave your answers as fractions. Consider the following corpus, where each line is a "sentence":

abbacad
dababac

a) Compute the probability of the sentence *bd* under an "unsmoothed unigram language model". [3 marks]

b) Compute the probability of the sentence *bd* under an "bigram language model with add-k smoothing" where $k = 0.1$. [3 marks]

c) Compute the probability of the sentence *bd* under an "bigram language model with absolute discounting" where the discount factor $d = 0.1$. [12 marks]

d) The "Katz Backoff" method for language modelling differs in several ways to the bigram language model with absolute discounting above. Explain its core difference, and give 2 examples where the probabilities would differ by making reference to the above corpus to support your answer. [6 marks]

— End of Exam —