The University of Melbourne

School of Computing and Information Systems

COMP90042
Natural Language Processing
Semester 1 2020

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

Length: This paper has 5 pages including this cover page.

Format: Open Book

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 [13 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) "Sparsity" is a key problem in text processing; explain what is meant by "sparsity" and outline an important impact this has on an "N-gram language model". [2 marks]

b) Contrast the evaluation metric "BLEU" and "ROUGE". Identify one similarity and one key difference. [2 marks]

c) Explain what is meant by the "contextual representation" of a word? Why might they be more useful than Word2Vec embeddings? [2 marks]

d) A copy mechanism is introduced to encoder-decoder models for abstractive summarisation. Would a copy mechanism help encoder-decoder models for translation? Explain. [1 mark]

Question 2: Formal Language Theory [3 marks]

a) "Regular languages" are closed under intersection. Explain what this means, and why this is important for language processing. [2 marks]

b) Describe how "finite state transducers" differ from "finite state automata". [1 mark]

Question 3: Topic Models [3 marks]

a) Name two hyper-parameters in "Latent Dirichlet Allocation", and describe their roles in how they influence the model in terms of topic quality and output distribution. [2 marks]

b) Explain why it is difficult to evaluate topic models, and name one method for evaluation of topic models. [1 mark]

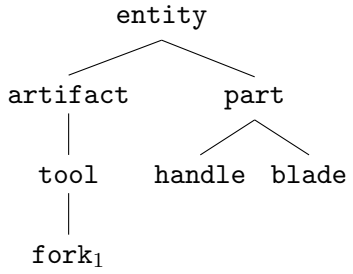
Continued overleaf ...

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: Lexical semantics [5 marks]

a) Based on the following WordNet-style graph of hypernymy:



where `artifact` means "a man-made object", and $\texttt{fork}_1$ means "an agricultural tool used for lifting or digging". Insert the following lemmas: $\texttt{fork}_2$, `knife`, `cutlery` (in their kitchenware senses), and `spade`. You are free to alter the structure of the graph to accomodate the new lemmas. Add distinguished edges (e.g., dashed) for at least two meronym relations. [2 marks]

b) Describe the "word2vec skip-gram" and "Latent Semantic Analysis" methods for learning vector representations of words. Compare and contrast these methods, considering their underlying modelling intuitions, data requirements, and algorithms for implementation. [3 marks]

Question 5: Information Extraction [4 marks]

Consider the following document, composed of only one sentence, its corresponding Named Entity annotation and a gold set of relations extracted from it:

- `Hugh Jackman is an actor born in 1968 in Sydney, NSW, Australia.`
- $\texttt{[Hugh Jackman]}_{\mathbf{PER}}$ `is an actor born in` $\texttt{[1968]}_{\mathbf{TIME}}$ `in` $\texttt{[Sydney]}_{\mathbf{LOC}}$, $\texttt{[NSW]}_{\mathbf{LOC}}$, $\texttt{[Australia]}_{\mathbf{LOC}}$.
- Gold relations:
  - `year-of-birth(Hugh Jackman, 1968)`
  - `place-of-birth(Hugh Jackman, Sydney)`
  - `city-state(Sydney, New South Wales)`
  - `state-country(New South Wales, Australia)`

a) Suppose you want to train a Named Entity Recogniser using an Hidden Markov Model. Rewrite the named entity annotated sentence into a sequence of (word, tag) elements using one of the schemes you learned in class. Write your answer in the following format: `word1/tag1 word2/tag2 ...` [2 marks]

b) The first step in Relation Extraction is to build a binary classifier that recognises if two entities have a relation or not. Assuming the example above is the only data you have available, how many positive and how many negative instances you would have in your training set for this classifier? [1 mark]

c) The second step in Relation Extraction is to build a multi-class classifier that, given a positive entity pair, predicts the relation between them. However, even if you have a perfect classifier the relations extracted from the sentence will not match the gold relations given above. Why is this the case? How would you solve this problem, so the relations match the gold standard? [1 mark]

Continued overleaf ...

Question 6: Dependency Grammar [6 marks]

a) Describe what it means for two words to be in a "dependency" relation, and provide an example. [1 mark]

b) Show the dependency parse for the sentence

> And I would eat them in a boat

You do not need to provide edge labels. [2 marks]

c) Show a sequence of parsing steps using a "transition-based parser" that will produce this dependency parse. Be sure to include the state of the stack and buffer at every step. [3 marks]

Continued overleaf . . .

Section C: Algorithmic Questions [12 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: N-gram language models [6 marks]

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

```
natural language processing
natural language understanding
natural language applications in the wild
```

a) Calculate a bigram language model over this data, using add 1 smoothing. Add additional symbols as needed. Hint: you should start by considering each context word, one at a time. [3 marks]

b) Compute the probability of the sentence "language understanding applications" under your bigram language model. [1 mark]

c) The "Kneser-Ney" method for language modelling differs in several ways to the language model used above. Explain two such differences, making reference to the above corpus, where appropriate, to support your answer. [2 marks]

Question 8: Context-Free Grammar & Parsing [6 marks]

This question is about using analyzing syntax. Consider the following ambiguous sentence:

```
Find the boy with an eye
```

a) Describe the syntactic ambiguity in this sentence [1 mark]

b) Write a set of linguistically-plausible CFG productions that can represent and structurally differentiate the two interpretations. [2 marks]

c) Perform CYK parsing of the sentence using your grammar. You should include the full chart described in the lecture, which will include the edges for both possible interpretations. You should label the order you fill in the cells. [3 marks]

— End of Exam —