First Name: _____ **Last Name**: _____

McGill ID: _____ **Section**: _____

Faculty of Science
COMP-550 - Natural Language Processing (Fall 2018)
Midterm Examination

October 31st, 2018
18:05 – 19:25

Examiner: Jackie Chi Kit Cheung

Instructions:

- **DO NOT TURN THIS PAGE UNTIL INSTRUCTED**
- This is a **closed book** examination.
- Only writing implements (pens, pencils, erasers, pencil sharpeners, etc.) are allowed. The possession of any other tools or devices is prohibited.
- Answer **all** questions **on this examination paper** and return it.
- This examination has **15** pages including this cover page, and is printed on both sides of the paper.
- **MAKE SURE TO WRITE YOUR NAME AND STUDENT ID ON THE EXAM. MARKS WILL BE DEDUCTED IF INFORMATION IS MISSING.**

Sections

The exam consists of the following sections:

1. Multiple Choice: Questions 1 to 13
2. Short Answer: Questions 14 to 18
3. Problem Sets: Questions 19 to 21

Multiple Choice Questions (1 point each)

Circle the correct response.

1. Which of the following is **not** an example of a referring expression?

   (A) Demonstrative
   (B) Pronoun
   (C) Proper name
   (D) Quantifier

2. Specify the lexical semantic phenomenon that is exhibited by the word *boots* in the following statement:
   "At this critical point of the war, the army commander realized more boots were required on the ground."

   (A) Synonymy
   (B) Synecdoche
   (C) Holonymy/Meronymy
   (D) Polysemy

3. Which of the following parameter sets is **not** part of the standard model of an HMM?

   (A) Initial state probabilities
   (B) State transition probabilities
   (C) Initial observation probabilities
   (D) Emission probabilities

4. Which of the following represents the assumption of a trigram model, where C represents the context?

   (A) $P(w_n|C) = P(w_n|w_{n-1}, w_{n-2}, w_{n-3}, ..., w_1)$
   (B) $P(w_n|C) = P(w_n|w_{n-1})$
   (C) $P(w_n|C) = P(w_n)$
   (D) $P(w_n|C) = P(w_n|w_{n-1}, w_{n-2})$

5. Which of the following statements concerning the differences between Hidden Markov Models (HMM) and Linear-Chain Conditional Random Fields (LC-CRF) is true?

   (A) HMMs and LC-CRFs both allow features that depend on two words that are arbitrarily far apart.
   (B) An HMM is a generative model, while an LC-CRF is a discriminative model.
   (C) An HMM uses the Viterbi algorithm for inference, while an LC-CRF uses gradient ascent for inference.
   (D) All of the above

6. What are the following reasons one might use an long short-term memory neural network (LSTM) rather than a standard vanilla recurrent neural network (RNN)?

   (A) RNNs have more parameters that have to be learned than LSTMs
   (B) RNNs can suffer from exploding gradients during training, while LSTMs can avoid this
   (C) LSTMs can model any sequential data, whereas RNNs can only model text sequences
   (D) LSTMs can take advantage of distributional semantics by using word embeddings as feature inputs, whereas RNNs cannot

7. Which of the following models/algorithms **cannot** be used to produce word embeddings?

   (A) CYK algorithm

   (B) Deep neural networks

   (C) Latent semantic analysis

   (D) Singular value decomposition

8. Suppose we have a PCFG with the following rules defined, where the value in parentheses is the probability associated with that rule:

   N → *syntax* (0.7) | *subject* (0.3)
   V → *is* (1)
   D → *a* (1)
   A → *great* (1)
   NP → A N (0.1) | N (0.4) | D NP (0.4) | D N (0.1)
   VP → V NP (0.9) | V A (0.1)
   S → NP VP (1)

   Using these rules, there is only one legal parse of the sentence: *Syntax is a great subject*. What is the probability of that parse, according to the probabilities defined by the PCFG above?

   (A) 1

   (B) $0.0144 = 0.4 \times 0.9 \times 0.4 \times 0.1$

   (C) $0.00003024 = 0.4 \times 0.7 \times 0.9 \times 0.4 \times 0.1 \times 0.1 \times 0.1 \times 0.3$

   (D) $0.003024 = 0.4 \times 0.7 \times 0.9 \times 0.4 \times 0.1 \times 0.3$

9. Which of the following statements is **not** true?

   (A) In the word pair (*house*, *door*), *house* is a holonym and *door* is a meronym.

   (B) The different possible senses of the word *break* is an example of polysemy.

   (C) The phrase *the Pentagon often overthrows democratically elected governments* presents an example of metonymy.

   (D) In the word pair (*hot*, *hell*), *hot* is a hyponym and *hell* is a hypernym.

10. Which of the following is an example of a closed part-of-speech class in English?

    (A) Determiners

    (B) Nouns

    (C) Adjectives

    (D) Verbs

11. Let there be N documents in our training corpus. Stochastic Gradient Descent is used for training a neural network with a batch size of:

    (A) 1

    (B) N

    (C) N/2

    (D) Stochastic Gradient Descent is not used for training a neural network.

12. Consider a hidden Markov model with 5 possible states and 10 possible words. How many *free* parameters exist for its emission matrix?

    (A) 36

    (B) 40

    (C) 45

    (D) 50

13. Which of the following statements hold True for Viterbi algorithm: (i) It is a dynamic programming algorithm. (ii) It is used to compute the likelihood of a sequence of observations. (iii) It is used to compute the most likely sequence of hidden variables for a given sequence of observations.

    (A) Only (i)

    (B) (i) and (ii)

    (C) (ii) and (iii)

    (D) (i) and (iii)

Short Answer

14. Compare derivational morphology and inflectional morphology by giving **two** ways in which they are different. Also, give an example in English of each type of morphology (3 points).

15. a) Perform beta reduction on the following lambda calculus expression as much as possible. (2 points)
$(\lambda x.(\lambda y.yx)(\lambda z.xz))(\lambda w.ww)$

b) What is the role that lambda calculus plays when we are deriving the meaning representation of a sentence compositionally? (2 points)

16. Distributional semantic models rely on the hypothesis that words with similar distributions have related meaning. These models identify both synonyms and antonyms as related words, but cannot effectively distinguish between the two relations. Suggest **two** ways that this problem can be alleviated and briefly describe how they would work (in two or three sentences each). (4 points)

There is a question on this page. Did you answer it?

17. Translate the following sentences into first-order logic, using Neo-Davidsonian event semantics. List the names of each predicate and function that you create.

a) *At least one of the students in COMP-550 hates semantics.* (2 points)

Predicates: Functions:

b) *I know that John loves semantics.* (2 points)

Predicates: Functions:

18. Consider the following dictionary definitions of two different word senses from Wordnet for the verb "shoot":

    (1) (fire a shot) "the gunman blasted away"

    (2) (making a film or photograph of something) "take a scene"; "shoot a movie"

    Now, consider the following context sentence, upon which we would like to perform word sense disambiguation:

    - *In this scene, we will shoot the gunman running away from a fire he made.*

    a) Using Lesk's algorithm, what sense (1 or 2) would be predicted for shoot? Use all words in the entire definition, including the gloss and the example phrases, in your calculations. Use the number of unique word types as your overlap measure. Show your work. (2 points)

    b) Now, supposing all stopwords are removed and that each word is lemmatized, what word sense would Lesk's algorithm predict? Assume that your stopword list contains all determiners and demonstratives, but not prepositions or other words. Show your work. (2 points)

Problem Sets

19. Consider the following incomplete forward and backward trellises, which are produced during the process of running EM on a sentence of three words, in a labelling problem with two states, S_1 and S_2.

Forward trellis:

S_1	0.05	X	0.022855
S_2	0.3	0.0705	0.006085

Backward trellis:

S_1	0.082	0.4	1
S_2	Y	0.28	1

Potentially useful definitions:

$$P(\mathbf{O}|\theta) = \sum_i \alpha_i(t)\beta_i(t)$$

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{P(\mathbf{O}|\theta)}$$

$$\xi_{ij}(t) = \frac{\alpha_i(t)a_{ij}b_j(O_{t+1})\beta_j(t+1)}{P(\mathbf{O}|\theta)}$$

a) Give an expression for the missing entries in the trellises in terms of the other numerical entries. There is no need to simplify or evaluate the expression down to a single term. (2 points)

$X =$

$Y =$

b) What is the probability of being in state S_1 in the first timestep given the observation sequence? Let's call this quantity Z. (2 points)

$Z =$

20. (This question has **three** parts, and continues on the next page.) Nouns can have different word senses, depending on their context. For example, "board" can refer to a piece of wood, or to a group of people (as in a school board). Linguists have defined high-level semantic categories of hypernyms, called supersenses, into which word senses can be classified. For nouns, there are 25 such hypernymic **supersense labels** (including, for example, *plant, animal, artifact, group, person,* etc.). For example, the piece of wood reading above would be classified into *artifact*, whereas the school board reading would be classified into *group*.

    Suppose we have a dataset $\mathcal{D} = \langle X, Y \rangle$ such that for each sample (x, y), $x \in X$ is a sequence of tokens, and $y \in Y$ is a sequence of supersense labels for each token in the sequence x. Words that are **not** nouns receive an "O" label. We would like to train a model to **predict the supersense labels of text sequences**.

    a) Name **three different types of models** that could be used to solve this problem. For each model, indicate one advantage it would have over the other models for this task, given certain assumptions about the available data to learn from (e.g., one model may perform better based on a certain assumption about the size/distribution of labelled and/or unlabelled training samples). (3 points)

b) Suppose you were to use a deep learning-based neural model to solve this problem. Name **three deep learning design decisions** you would have to make **and explain why**. For example, some design considerations would be: what gradient descent optimization algorithm would you use, what model architecture would you use? (3 points)

c) In order to evaluate the quality of an NLP algorithm, it is critical to compare it to a simple baseline method. Describe a baseline algorithm/function that we could compare our model to. (2 points)

There is a question on this page. Did you answer it?

21. (This question has **three** parts, and continues on the next page.)

a) Give the formal definition of a Context Free Grammar. (2 points)

b) What does it mean for a grammar to overgenerate and to undergenerate? For the grammar below, give an example of each. (3 points)

S $\rightarrow$ NP VP	DT $\rightarrow$ *the*
NP $\rightarrow$ NN	NN $\rightarrow$ *man* $\mid$ *dog* $\mid$ *cat*
NP $\rightarrow$ DT NN	VB $\rightarrow$ *see* $\mid$ *sees*
NP $\rightarrow$ NP PP	
VP $\rightarrow$ VB NP PP	
PP $\rightarrow$ *in* NP	

Definitions of over- and undergeneration:

Examples:

c) Convert the grammar from part b) into Chomsky Normal Form. You may simply write out the rules to add and to remove. (3 points)

There is a question on this page. Did you answer it?

This page is left intentionally blank. You may do rough work on it.

MC:	13
Q14-16:	11
Q17-18:	8
Q19:	4
Q20:	8
Q21:	8
Total:	**52**