First Name: ________________________________ **Last Name**: ________________________________

McGill ID: ________________________________ **Section**: _______

Faculty of Science
COMP-550 - Natural Language Processing (Fall 2017)
Midterm Examination

November 9th, 2017
16:05 – 17:25

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 **13** 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 15
2. Short Answer: Questions 16 to 20
3. Problem Sets: Questions 21 to 23

Multiple Choice Questions (1 point each)

Circle the correct response.

1. Which of the following is **not** an example of a dynamic programming algorithm?

   (A) CYK algorithm
   (B) Backpropagation algorithm
   (C) Viterbi algorithm
   (D) Yarowsky's algorithm

2. The study of how sentences are structured is called:

   (A) Morphology
   (B) Pragmatics
   (C) Semantics
   (D) Syntax

3. An automatic tool has converted the input `itemized` into the output `itemize`. What task has this automatic tool performed?

   (A) Lemmatization
   (B) Morphological recognition
   (C) Parsing
   (D) Stemming

4. Assuming the same vocabulary and the same assumptions about the distributions responsible for generating samples, which of the following parameter estimation techniques will result in the greatest test corpus likelihood?

   (A) Maximum likelihood estimation (MLE)
   (B) Maximum a posteriori estimation (MAP)
   (C) Bayesian inference
   (D) It is impossible to say, with the given information

5. Let $\alpha_j(t)$ be the cell in the forward trellis of a Hidden Markov Model representing the state j at timestep t, and let $\delta_j(t)$ be the cell in the trellis when running the Viterbi algorithm on the same sequence, with the same state j and timestep t. What can we say about the relationship between $\alpha_j(t)$ and $\delta_j(t)$?

   (A) $\alpha_j(t) < \delta_j(t)$
   (B) $\alpha_j(t) = \delta_j(t)$
   (C) $\alpha_j(t) > \delta_j(t)$
   (D) None of the above

6. I would like to train a system that will predict the score that a game reviewer assigns to a newly released game, from 0.0 to 100.0, given the text of the review. This is best cast as what kind of problem?

   (A) Bootstrapping
   (B) Classification
   (C) Feature extraction
   (D) Regression

7. Let unary predicates $A(x)$ and $B(x)$ represent concepts, such that A is a hyponym of B. Which of the following statements is true?

   (A) $A(x) \rightarrow B(x)$
   (B) $A(x) \leftarrow B(x)$
   (C) $A(x) \leftrightarrow B(x)$
   (D) $A(x) \vee B(x)$

8. What is one reason to prefer using Context Free Grammars over Finite State Automata as a formal model of natural language?

   (A) CFGs overgenerate less than FSAs.
   (B) CFGs undergenerate less than FSAs.
   (C) CFGs can account for linguistic constructions that FSAs cannot.
   (D) There are treebanks labelled with CFG parse trees, and there are not for FSAs.

9. "An even number is an integer that is evenly divisible by two." This is an example of a(n):

   (A) Extensional definition
   (B) Intensional definition
   (C) Logical definition
   (D) Prototypical definition

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

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

11. 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

12. Compute the pointwise mutual information (PMI) association between "COMP550" and "awesome" in the following text. For the purposes of this question, two words are considered to co-occur if they are at most 4 words apart (i.e., separated by at most 3 words). Ignore punctuation.

    "At the end of the semester, the students found COMP550 awesome. It introduced all the fundamentals in NLP. Also, COMP550 gave an awesome coverage of classical techniques as well as recent advances in this awesome field. What an awesome course!"

    (A) log(8)
    (B) log(0.1)
    (C) log(10)
    (D) log(0.25)

13. 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

14. Which of the following represents the assumption of a bigram 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) $P(w_n|C) = P(w_n|w_{n-1}, w_{n-2})$

15. 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

Short Answer

16. Give **two** reasons why it is problematic to interpret the probability that we get from an N-gram language model as a grammaticality score. (2 points)

17. Lambda calculus

    a) Perform beta reduction on the following lambda calculus expression as much as possible. (2 points)

    $(\lambda$y.yy a$)(\lambda$x.x$)(\lambda$z.$(\lambda$u.u$)$z$)$

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

18. Explain why *-ness* is a more productive suffix in English than *-th*. (2 points)

19. Briefly explain the relationship between the following terms: time-delay neural network, recurrent neural network, feed-forward neural network, sequence modelling. (You should not need more than two sentences.) (4 points)

20. Distributional semantic models rely on the hypothesis that words with similar distributions have related meaning. Accordingly, these models will 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. (4 points)

Problem Sets

21. a) Translate the following sentence into first-order logic (FOL), using Russell's analysis of definite descriptions, using the predicates *Apple(x)*, *Yard(x)*, and *IsIn(x, y)*. Do not use neo-Davidsonian event semantics. (3 points)

    The apple is in the yard.

    b) Specify an interpretation of the FOL such that the above statement is true. You may assume that equality $=$ is defined as expected. (4 points)

22. **This question contains three parts, and continues on the next page.** Consider the following grammar (assume that S is the start symbol):

S → NP VP
NP → DT NN
NP → DT NNS
NP → NP PP
PP → IN NP
VP → VB NP
VP → VP PP

DT → the
NN → man | dog | cat
NNS → dogs | cats
VB → see | sees
IN → in | with

a) Show two examples of how the grammar overgenerates. (2 points)

b) Show two examples of how the grammar undergenerates. (2 points)

c) Apply the CYK parsing algorithm to the sentence:

The dog sees the cats.

Ignore case and punctuation. Don't forget the backpointers. Give the final output parse(s). (6 points)

23. **This question contains four parts, and continues on the next page.** We saw the Vernam cipher with a shifted key in Assignment 2, in which each letter was encoded by first shifting the sentence by an amount 'x', then added to the original plaintext. For this question, consider a Vernam cipher with the key shifted by **a single place**. Unlike the standard Vernam cipher, we will not wrap the sentence from the end back to the beginning. Thus, the first letter always remains the same. For example, the encoding for the message, 'smoke', is shown below:

s	**m**	**o**	**k**	**e**	{plaintext word, hidden layer}
	s	**m**	**o**	**k**	{key rotated by 1 place}
s	**e**	**a**	**y**	**o**	{cipher word, observed layer}

The vocabulary is restricted to the **26 English letters only**. It turns out that this cipher can be solved by a character-level HMM in which the output emission depends on the hidden states at the current and the previous timesteps.

a) Draw a directed graph representing the graphical model of the **modified HMM** in unrolled form. (2 points)

b) Using the plaintext word 'smoke' as your training text, give the MLE estimate of all the parameters of your **modified HMM**. If there are many entries with the same value, you may use a shorthand such as "all other letters". (3 points)

c) Introduce Laplace smoothing, and write down your new parameters. (3 points)

d) What are the total number of probability distributions and free parameters in **your modified HMM**? (2 points)

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