First Name: SOLUTIONS **Last Name:** SOLUTIONS

McGill ID: / **Section:** /

Faculty of Science
COMP-599 - Introduction to Natural Language Processing (Fall 2015)
Midterm Examination

November 10th, 2015
13:05 – 14:25

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. The study of the sound patterns in natural language and the rules that govern them is:

   (A) Phonetics
   (B) Morphology
   (C) Phonology
   (D) Syntax

2. Recursion in natural language is responsible for:

   (A) The hierarchical structure of language.
   (B) The compositional nature of natural language.
   (C) The lack of an upper bound on sentence length.
   (D) Tractable algorithms for parsing natural language sentences.

3. Zipf's Law states that:

   (A) The frequency of a word type is proportional to its rank by frequency.
   (B) The frequency of a word type is inversely proportional to its rank by frequency.
   (C) The frequency of a word token is proportional to its rank by frequency.
   (D) The frequency of a word token is inversely proportional to its rank by frequency.

   This question was poorly worded

4. Which of the following is *not* a problem when using maximum likelihood estimation to obtain the parameters to a language model?

   (A) Unreliable estimates when there is little training data
   (B) Out-of-vocabulary items
   (C) Overfitting
   (D) Smoothing

5. Which of the following is *not* a part of speech?

   (A) Noun phrase
   (B) Verb
   (C) Interjection
   (D) Determiner

6. Which of the following is *not* a good example of a cohesive device?

   (A) Discourse markers
   (B) Pronouns
   (C) Demonstratives
   (D) Prepositions

7. What is one advantage of a Porter stemmer over a full morphological analyzer?

   (A) The stemmer is better justified from a theoretical point of view.
   (B) The stemmer does not require a detailed lexicon to implement.
   (C) The output of the stemmer results in better performance in most downstream tasks, compared to the output of the morphological analyzer.
   (D) The output of the stemmer is always a valid word, unlike the output of the morphological analyzer.

B

8. A Hidden Markov Model is a (1), which means that the model is specified by (2). (Select the response that correctly completes both missing pieces.)

   (A) (1) discriminative model (2) $P(X, Y)$
   (B) (1) discriminative model (2) $P(Y|X)$
   (C) (1) generative model (2) $P(X, Y)$
   (D) (1) generative model (2) $P(Y|X)$

C

9. Grammar induction is best viewed as a(n):

   (A) Unsupervised learning problem.
   (B) Supervised learning problem.
   (C) Linear-chain Conditional Random Field problem.
   (D) Maximum A Posteriori (MAP) estimation problem.

A

10. What problem does IOB tagging solve in chunking?

    (A) Overfitting
    (B) Determining the boundaries of chunks
    (C) Defining arbitrary linguistic features in HMMs
    (D) Semi-supervised learning

B

11. What is a consequence of the convexity of a Linear-Chain Conditional Random Field?

    (A) We can determine the MLE of the model parameters analytically.
    (B) We can use gradient descent to find the global optimum of model parameters in terms of the training corpus likelihood.
    (C) We can condition on any part of the observed sequence when defining the features of the model.
    (D) We can use a version of the forward algorithm to find the normalization constant.

B

12. Which of the following is not a part of a semantic frame in Fillmore's Frame Semantics?

    (A) Linguistic expressions that realize the semantic frame
    (B) Distinctions between core and non-core frame elements
    (C) A schematic description of the real-world situation in terms of frame elements
    (D) A list of potential anaphors that can refer to the frame elements

D

13. The words *room* and *house* are in a lexical semantic relation, in which *room* is the (1) and *house* is the (2).

(A) (1) hypernym (2) hyponym
(B) (1) hyponym (2) hypernym
(C) (1) holonym (2) meronym
(D) (1) meronym (2) holonym

14. Which word in the following sentence is the best example of a metonym? *With Madrid failing to act on the recommendations, this year a Catalan parliamentary commission also recommended rejoining GMT.*

(A) Madrid
(B) recommendations
(C) Catalan
(D) GMT

15. An advantage of Yarowsky's algorithm over Lesk's algorithm is:

(A) It does not depend on definitions from a lexical resource.
(B) It does not require the use of unlabelled training data.
(C) It can disambiguate between more than two possible word senses.
(D) It can be used to discover cause-effect relationships, unlike Lesk's algorithm.

Short Answer

16. What is the entropy (in bits) of the following loaded four-sided die? You do not need to evaluate or simplify the expression. (2 points)

Outcome	Probability
1	0.25
2	0.5
3	0.1
4	0.15

$-0.25\log_2(0.25) - 0.5\log_2(0.5) - 0.1\log_2(0.1) - 0.15\log_2(0.15)$

17. Let $\alpha_i(2)$ be the cell in the forward trellis of a bigram hidden Markov model corresponding to hidden state i at time step 2, and w_2 be the word at that timestep. Write down the calculation of $\alpha_i(2)$ in terms of the values in the previous column of the trellis, assuming that the model has N hidden states. Use a_{ij} to indicate the transition probability from state i to state j and $b_i(w)$ to indicate emission probability of observation w from state i. (2 points)

$$\alpha_i(2) = \sum_{k=1}^{N} \alpha_k(1)\, a_{ki}\, b_i(w_2)$$

18. Identify three pairs of text spans in the passage below that are coreferent. Clearly indicate which text span corefers with which text span. (3 points)

Venezuela rejoined the 30-minuters in 2007, after its president, Hugo Chavez, issued a decree that moved the clocks back an hour. Chavez claimed school children wouldn't be as tired in the morning after the change.

3 of those:
- Venezuela — its
- its president — Hugo Chavez
- Hugo Chavez — Chavez
- (a decree that) moved the clocks back an hour — the change

19. Give the formal definition of a context-free grammar (CFG). (3 points)

N : set of non-terminals
T : set of terminals
$S \in N$: a unique start symbol
R : a list of rules or productions in the form
$A \rightarrow \beta_1 \dots \beta_k$
where $A \in N$
$\beta_1 \dots \beta_k \in N \cup T$

20. Extend your previous answer to give a formal definition for a *probabilistic* context-free grammar (PCFG). Remember to indicate what should form a probability distribution. (3 points)

N : set of non-terminals
T : set of terminals
$S \in N$: a unique start symbol
R : a list of rules or productions in the form
$A \rightarrow \beta_1 \dots \beta_k$, $\gamma \in [0, 1]$
where $A \in N$
$\beta_1 \dots \beta_k \in N \cup T$, γ is $P(A \rightarrow \beta_1 \dots \beta_k)$
s.t. $\forall_{A \in N} \sum_{r \in R} P(r) = 1$
where LHS is A

Problem Sets

freq table

A	4
B	1
C	4
D	2
E	1
F	1
G	3
tot	16

21. Take the following series of tokens as a training corpus.

B A C D A D C E A C G G F G A C

a) Identify the hapax legomena in the corpus. (2 points)

B, E, F

-1 per missing or incorrect

b) Give the maximum likelihood estimate of a unigram language model trained on the corpus. Leave your answers as fractions. (2 points)

P(A) 4/16
P(B) 1/16
P(C) 4/16
P(D) 2/16
P(E) 1/16
P(F) 1/16
P(G) 3/16

-1 for a minor mistake

c) Give the Good-Turing estimate of a unigram language model trained on the corpus. Use the simple version of Good-Turning smoothing (i.e., does not require training a regression model). You may simply write out the calculations; you do not need to simplify the fractions. Be sure to include $P(\text{UNK})$. (3 points)

Count of counts:

i	f_i
1	3
2	1
3	1
4	2
5	0

$$P(\text{UNK}) = \frac{3}{16}$$

$$c_1^* = 2\frac{f_2}{f_1} = 2\left(\frac{1}{3}\right) = \frac{2}{3}$$

$$c_2^* = 3\frac{f_3}{f_2} = 3\left(\frac{1}{1}\right) = 3$$

$$c_3^* = 4\frac{f_4}{f_3} = 4\left(\frac{2}{1}\right) = 8$$

$$c_4^* = 5\frac{f_5}{f_4} = 5\left(\frac{0}{2}\right) = 0.$$

X :	P(X) :
A	0
B	$\frac{2}{3} \times \frac{1}{16} = \frac{1}{24}$
C	0
D	$\frac{3}{16}$
E	$\frac{1}{24}$
F	$\frac{1}{24}$
G	$\frac{8}{16} = \frac{1}{2}$
UNK	$\frac{3}{16}$

22. Japanese has two types of adjectives[1], *i*-adjectives and *na*-adjectives.

Below are two examples of each, conjugated by tense (present or past[2]) and polarity (positive or negative).

i-adjectives		*na*-adjectives	
oishii	is delicious	henda	is strange
oishikatta	was delicious	hendatta	was strange
oishikunai	is not delicious	henjanai	is not strange
oishikunakatta	was not delicious	henjanakatta	was not strange
takai	is tall	shizukada	is quiet
takakatta	was tall	shizukadatta	was quiet
takakunai	is not tall	shizukajanai	is not quiet
takakunakatta	was not tall	shizukajanakatta	was not quiet

a) What are the stems of the four adjectives above? What are the four affixes for each combination of the polarity (POS/NEG) and tense (PRES/PAST) for each class of adjectives? (2 points)

Stems:

oishi 'delicious' hen 'strange'
taka 'tall' shizuka 'quiet'

Affixes:

i-adjectives:

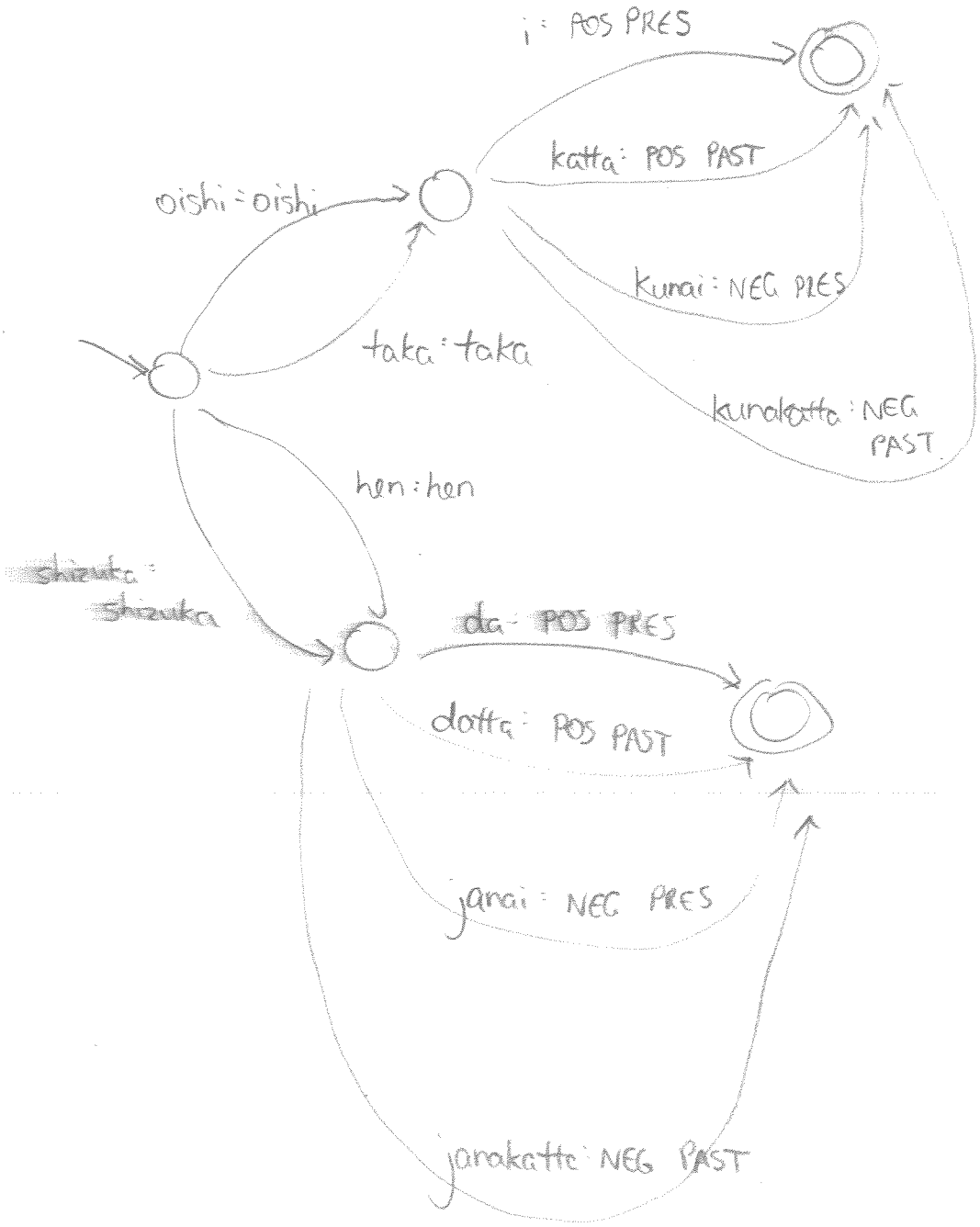
-i	POS	PRES
-katta	POS	PAST
-kunai	NEG	PRES
-kunakatta	NEG	PAST

na-adjectives:

-da	POS	PRES
-datta	POS	PAST
-janai	NEG	PRES
-janakatta	NEG	PAST

[1] Many linguists actually consider these adjectives to be a special subclass of verbs.
[2] This is linguistically inaccurate, but let's roll with it for the purposes of this exam question.

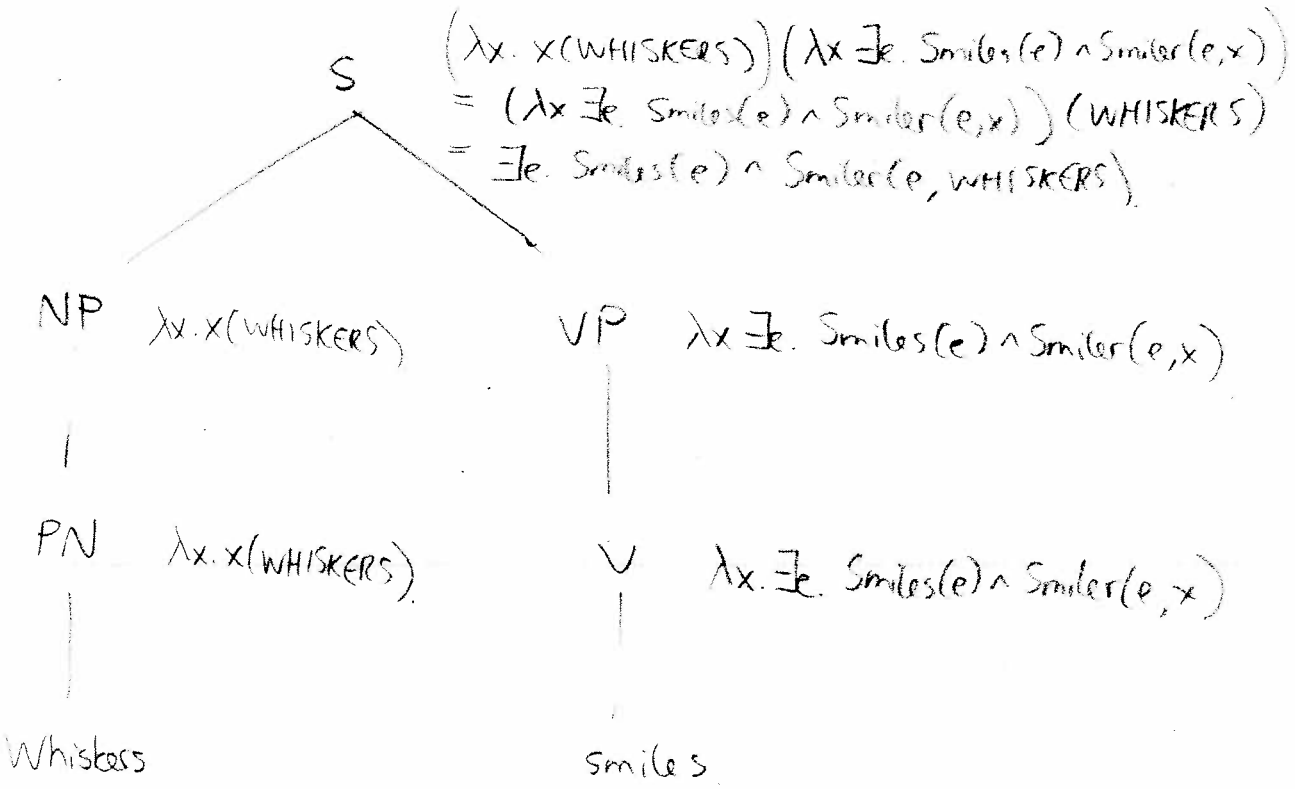
b) Draw a Finite-State Transducer, using the graphical representation presented in class, that parses an input adjective into its stem, polarity, and tense, in that order. You may include epsilon transitions and multi-character transitions to simplify your FST. (5 points)



23. Consider the following CFG, in which each rule (first column) is associated with a semantic attachment (second column), using the neo-Davidsonian event representation we saw in class. Note that one of the rules uses subscript indices so that the three VPs can be referred to, but the indices are not part of the syntactic category.

PN → *Whiskers*	{$\lambda x.x$(WHISKERS)}
NP → PN	{PN.sem}
V → *smiles*	{$\lambda x.\exists e.$*Smiles*$(e) \wedge$ *Smiler*(e,x)}
V → *purrs*	{$\lambda x.\exists e.$*Purrs*$(e) \wedge$ *Purrer*(e,x)}
VP → V	{V.sem}
Conj → *and*	?
$VP_1 \rightarrow VP_2$ Conj VP_3	{Conj.sem(VP_2.sem)(VP_3.sem)}
S → NP VP	{NP.sem(VP.sem)}

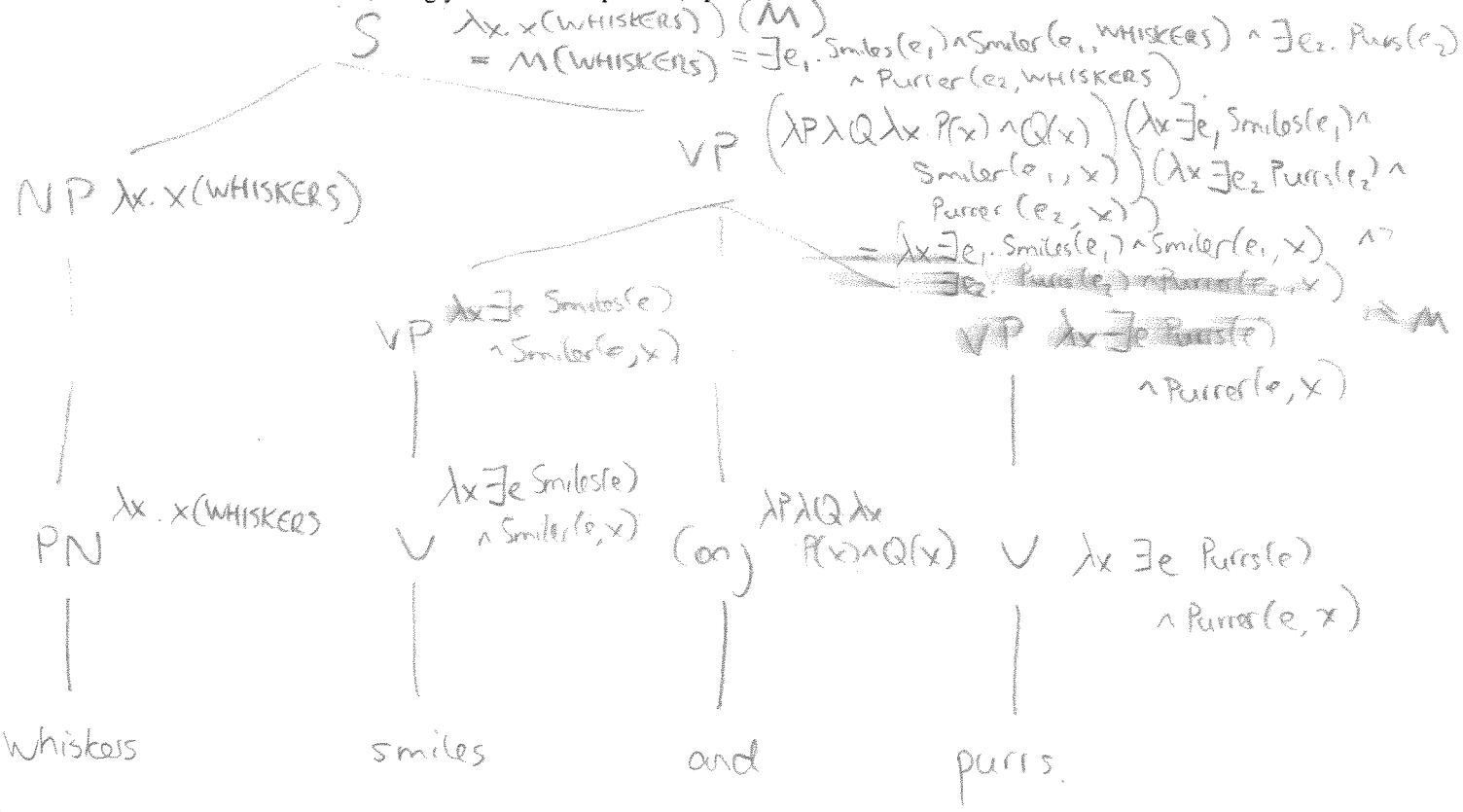
a) Draw the tree for *Whiskers smiles* with a lambda term meaning representation associated with each constituent in the tree, according to the above grammar. (3 points)



b) The desired representation of *Whiskers smiles and purrs* is $\exists e_1.Smiles(e_1) \wedge Smiler(e_1, \texttt{WHISKERS}) \wedge \exists e_2.Purrs(e_2) \wedge Purrer(e_2, \texttt{WHISKERS})$. What is the semantic attachment for the lexical entry of *and*? (3 points)

$\{ \lambda P \lambda Q \lambda x \; P(x) \wedge Q(x) \}$

c) Draw the tree for *Whiskers smiles and purrs* with a lambda term meaning representation associated with each constituent in the tree, using your rule from part b. (4 points)



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