

CPSC 477/577
Natural Language Processing
Yale University
Spring 2021
Practice Questions

Question 1

Represent the following sentences in first-order predicate calculus (FOPC). There may be multiple ways to represent each of them. Give only one representation for each sentence.

- (a) Only one person understood the play.
- (b) Exactly two people understood the play.

Question 2

Give an example for each of the following verb subcategorization phrases.

- 1. NP
- 2. NP NP
- 3. \emptyset
- 4. V P_{to}
- 5. S

Question 3

Consider the five sentences shown in Figure 1. Think of them as your training data to build a probabilistic context-free grammar (PCFG).

Delta flight 411 leaves Toronto for Atlanta at 6 PM
This flight serves a light meal
When does this flight leave
Northwest **flight 29 leaves Atlanta for Detroit at 10 AM**
When does the Delta flight arrive
That flight may serve a meal

Figure 1: Training sentences.

The parse trees for these sentences are shown in Figure 2:

```

(S (NP (NNP Delta)
      (NN flight)
      (CD 411))
  (VP (VBZ leaves)
      (NP (NNP Toronto))
      (PP (IN for)
          (NP (NNP Atlanta)))
      (PP (IN at)
          (NP (QP (CD 6)
                  (RB PM))))))

(S (NP (DT This)
      (NN flight))
  (VP (VBZ serves)
      (NP (DT a)
          (JJ light)
          (NN meal))))

(S (WH (WRB When))
  (S (AUX does)
    (NP (DT this)
        (NN flight))
    (VP (VB leave))))

(S (NP (NNP Northwest)
      (NN flight)
      (CD 29))
  (VP (VBZ leaves)
      (NP (NP (NNP Atlanta))
          (PP (IN for)
              (NP (NNP Detroit))))
      (PP (IN at)
          (NP (CD 10)
              (RB AM))))))

(S (WH (WRB When))
  (S (AUX does)
    (NP (DT the)
        (NNP Delta)
        (NN flight))
    (VP (VB arrive))))

(S (NP (DT That)
      (NN flight))
  (VP (MD may)
      (VP (VB serve)
          (NP (DT a)
              (NN meal))))))

```

Figure 2: Parse trees for the sentences in Figure 1.

- (a) Build a PCFG using the training data. For each non-lexical rule (e.g., $S \rightarrow NP VP$), indicate its probability.
- (b) Build a probabilistic lexicon for the PCFG. Give the probability of each lexical rule (e.g., $NN \rightarrow \text{flight}$).
- (c) Smoothing (reserving probability mass for unobserved rules is very important when building PCFGs. Redo parts (a) and (b) above using 10% of the probability mass for each non-terminal or lexical category to cover unknown words. Example: if $A \rightarrow B C$ has a probability of .6 and $A \rightarrow B D$ has a probability of .4, you need to create a new rule $A \rightarrow \alpha$ with a probability of .1 and readjust downward the probabilities of the other two rules that have A on the left-hand side.
- (d) For each of the following two sentences "When does Northwest flight 77 leave for Milwaukee" and "Does this flight leave for Milwaukee", draw **one** parse tree according to the (smoothed) grammar in part (c). If you are getting any zero probabilities, return to part (c) and fix your grammar appropriately. What are the final probabilities for each of these two sentences? For this question, you don't need to find all possible parses of a given sentence. One parse per sentence will be enough.

Question 4

What (maximum likelihood) formula is used to estimate trigram probabilities $P(w_n | w_{n-1}, w_{n-2})$ using a corpus. Don't worry about smoothing or backoff.

Question 5

Why are "mildly context-sensitive grammars" like Tree Adjoining Grammars (TAG) and Combinatory Categorical Grammars (CCG) used in NLP? Give two reasons.

Question 6

English has the wonderful feature that it lets you stick two nouns together into a compound noun, whose meaning derives in some idiosyncratic way from the meanings of its parts:

water fountain: a fountain that supplies water

water ballet: a ballet that takes place in water

water meter: a device (called meter) that measures water

water barometer: a barometer that uses water instead of mercury (to measure air pressure)

water biscuit: a biscuit that is made with water

water glass: a glass that is meant to hold water

Even more fun is that one of the two nouns in the compound noun could itself be a compound noun, as in the case of ice cream soda. But what's the recipe for that beverage? It depends. You

make [[ice cream] soda] by dropping ice cream into soda, but you make [ice [cream soda]] by dropping ice into cream soda.

A. The paragraph above used [square brackets] to distinguish two possible meanings of ice cream soda, one of them being the conventional meaning. Add brackets to each compound below to indicate whether the most likely meaning corresponds to [[X Y] Z] or [X [Y Z]].

- a. ice cream soda
- b. science fiction writer
- c. customer service representative
- d. state chess tournament
- e. Mars Rover landing
- f. plastic water cooler
- g. typeface design report

B. Choose the most likely bracketing for the 4-word compound noun country song platinum album.

- a. [country [song [platinum album]]]
- b. [country [[song platinum] album]]
- c. [[country song] [platinum album]]
- d. [[country [song platinum]] album]
- e. [[[country song] platinum] album]

Give a plausible definition of [[space mission] [[control freak] show]]. (If you must use compound nouns in your definition, define them too.)

C. Show the most likely bracketing for the 8-noun sequence below. As in the examples above, your bracketing must have the form [X Y], where each of X and Y is either a single-word noun or a compound noun (which must also be written as a bracketing [X Y] and so on.)

family board game togetherness effect government study

D. A computer program knows less about the world than you do, so it may have more trouble interpreting these sequences of nouns. How many bracketings must it choose among? Complete the following table by inserting the correct number for f(5). Bonus if you give f(6) as well.

f(1) = 1

f(2) = 1

f(3) = 2

f(4) = 5 (see part B. above)

f(5) = ???

f(6) = ???

Question 7

Give first-order logic translations for the following sentences:

Vegetarians do not eat meat.

Not all vegetarians eat eggs

Question 8

Consider the HMM below, as defined in the following transition and emission tables:

Transition Probabilities

	to 0	to 1	to 2	to 3	Comment
from 0		0.4	0.6		Start state
from 1		0.4	0.5	0.1	
from 2		0.6	0.2	0.2	
from 3					Final state

For example, the probability to get from state 0 to state 2 is 0.6

Emission Probabilities

state	symbol	prob
1	a	0.7
1	b	0.3
2	a	0.2
2	b	0.8

For example, the probability of emitting "a" from state 1 is 0.7

8.1. What is the probability of generating "baa"?

8.2. What is the most likely state sequence associated with the output string "aba"?

Question 9

A. list three part of speech categories that are open class.

B. list three part of speech categories that are closed class.

Question 10

Consider the following Dr. Seuss rhyme:

One fish two fish red fish blue fish black fish blue fish

Show a table with the bigram counts for this corpus.

Given this table, give $P(\text{fish}|\text{two})$ and $P(\text{black}|\text{fish})$.

Question 11

You are in a noisy bar diligently studying for your midterm, and your friend is trying to get your attention, using only a two words vocabulary. She has said a sentence but you couldn't hear one of the words:

($w_1 = \text{hi}$; $w_2 = \text{yo}$; $w_3 = ???$; $w_4 = \text{yo}$)

Assume that your friend was generating words from this first-order Markov model:

$$\begin{aligned} p(\text{hi}|\text{hi}) &= 0.7 & p(\text{yo}|\text{hi}) &= 0.3 \\ p(\text{hi}|\text{yo}) &= 0.5 & p(\text{yo}|\text{yo}) &= 0.5 \end{aligned}$$

Given these parameters, what is the posterior probability of whether the missing word is “hi” or “yo”?

Question 12

Consider the sentence

Marina often gives Bill cold water

(a) Write a linguistically motivated **CCG** lexicon entries for the six **words** in this sentence. Make sure to capture the correct subcategorization frame for “gives”. Use standard symbols, such as S, S\NP, etc.

(c) Show the full CCG parse for the sentence.

Question 13

What is the formula for the sigmoid function described in class $f(z)$? Why is it commonly used in Neural Networks?

Question 14

We first need to introduce a function that compares similarity between two sentences. Recall the definition for the cosine distance between two vectors x and y :

Part A: How would you define a mapping from sentences s to vectors $f(s)$ such that the cosine distance

$$\text{cosine}(f(s_1), f(s_2))$$

is a measure of the similarity between two sentences s_1 and s_2 ?

Part B: Now, we will define a dynamic programming algorithm that computes sentence alignment of two translations. The alignment score is the sum of the similarity scores of the aligned sentences. Our goal is to find an alignment with the highest score.

We will consider alignments of the following form:

- a sentence can be aligned to an empty sentence. This happens when one of the translators omits a sentence.
- a sentence can be aligned to exactly one sentence.
- a sentence can be aligned to two sentences. This happens when one of the translators either breaks or joins sentences.

Our sentence alignment algorithm recursively computes the alignment matrix F indexed by i and j , one index for each sentence. The value stored in $F(i, j)$ is the score of the best alignment between the first i sentences of x and the first j sentences of y . $s(x_i, y_j)$ is the similarity between sentence x_i and sentence y_j .

- Define $F(0, 0)$, $F(i, 0)$ and $F(0, j)$.
- Define $F(i, j)$ for $i > 0$ and $j > 0$.

Part C: Next, we'll modify the alignment score to be the same as before, but to include a fixed penalty p each time a sentence is aligned to an empty sentence. p is a parameter, which is ≥ 0 , chosen by the user of the algorithm. Describe a modified dynamic programming method which takes this new penalty into account.

Question 15

Assume in the contexts of HMMs that S refer to a sequence of states, O refers to a sequence of observations, and M is a particular HMM model. What is the algorithm that computes the value: $\operatorname{argmax} P(S | O, M)$?

Question 16

Consider a logistic regression model used to classify documents into “sports” vs. “not sports”. The weights $\theta = (-\ln(4), \ln(2), -\ln(3))$. A given document D is represented as the feature vector $x = (1, 1, 1)$.

What is the probability that document D is about sports?

Please provide your answer in the form of a fraction.

Question 17

What problem of simple recurrent neural network does LSTM solve? How does it solve the problem? What is another variant of SRN that solves the same problem?

Describe how an LSTM works. Draw a diagram and explain all gates and all forward propagation formulas.

Question 18

Give two reasons why a Naive Bayes classifier may be preferred over a neural network for some tasks. What important statistical assumption does a Naive Bayes classifier make that isn't likely to be true of text?

Let's use Naive Bayes to predict whether or not a sandwich will be tasty given its ingredients. Assume we have two possible classes of tastiness, "good" and "bad". Given:

$P(\text{"pickles"} \mid \text{"good"}) = 0.8$
 $P(\text{"mayo"} \mid \text{"bad"}) = 0.7$
 $P(\text{"ham"} \mid \text{"bad"}) = 0.2$
 $P(\text{"anchovies"} \mid \text{"good"}) = 0.1$
 $P(\text{"cheddar"} \mid \text{"good"}) = 0.8$
 $P(\text{"bad"}) = 0.4$

What, according to a Naive Bayes classifier, is the probability a sandwich containing ham, cheddar, and mayo is good? What is the probability the sandwich is bad? What assumption might the Naive Bayes classifier be making that isn't true of combining ingredients in a sandwich to determine if the sandwich will taste good?

Question 19

What is BERT? What is the key innovation? Please give two examples of using BERT for NLP tasks?

Question 20

Consider a logistic regression model with weights $\beta = (-\ln(4), \ln(2), -\ln(3))$. A given document has feature vector $x = (1, 1, 1)$. Now, please provide your answer in the form of a fraction a/b.

Question 21

What is the purpose of an activation function in a neural network? What would happen if we didn't have one?

Which of the following is true of sigmoid?

Sigmoid overcomes the vanishing gradient problem.

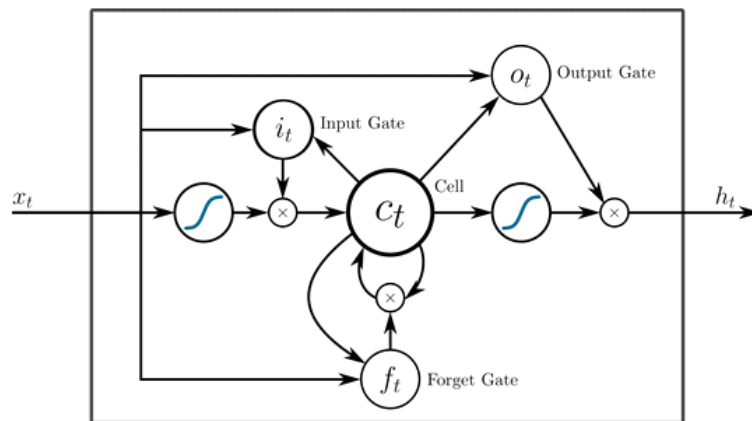
Optimization is easier in sigmoid than in tanh.

The resultant outputs are in the range $[0,1]$.

Sigmoid is preferred to ReLU.

Question 22

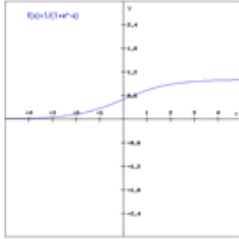
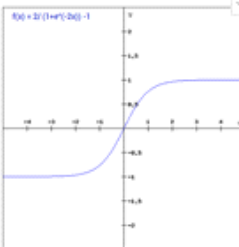
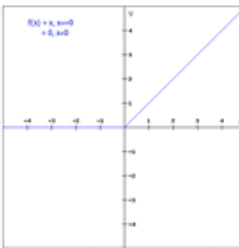
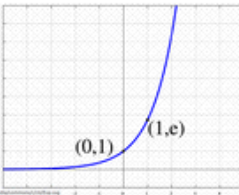
What does this figure represent?



- a. neural attention
- b. gated recurrent unit (GRU)
- c. long short-term memory network (LSTM)
- d. tree neural network
- e. support vector machine

Question 23

What are the names of the four functions below?

Mathematical Expression	Range	Plot
$\frac{1}{1 + e^{-x}}$	(0, 1)	
$2 * \text{sigmoid}(2x) - 1$	(-1, 1)	
$\max(0, x)$	$[0, \infty)$	
$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$	$[0, 1]$	

Question 24

Draw a neural network with a hidden layer that can compute $X \text{ XOR } Y$ for the Boolean variables X and Y . Show all biases and other weights and explain why your network works. Note that there are many possible answers.

Question 25

What is the idea behind “retrofitting embeddings”, a method introduced by Manaal Faruqui (answer in one sentence)?

Question 26

Give an example of a sentence that would be a valid Winograd schema. Do not use the examples shown in class.

Question 27

Which (zero, one, or more) of the following phrases are non-compositional?

- a. white cloud
- b. red herring
- c. green tree
- d. black market
- e. blue box

Question 28

BERT is trained to predict 15% of the tokens. What does it do with these 15%? Pick all correct answers (zero, one, or more):

- a. replaces the token with a mask
- b. replaces the token with a random token
- c. keeps the token unchanged
- d. duplicates the token
- e. writes the token backwards

Question 29

The formula below is used in what type of neural network:

$$\text{softmax} \left(\frac{\begin{matrix} \text{Q} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix} \times \begin{matrix} \text{K}^T \\ \begin{matrix} \square & \square \\ \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} \text{V} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix} \\ = \begin{matrix} \text{Z} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix}$$

- a. Bidirectional LSTM
- b. Transformer
- c. Stacked LSTM
- d. Pointer network

e. Convolutional neural network

Question 30

Please explain the difference between syntactic and semantic parsing. Consider questions such as: How is semantic parsing done? In what situation might semantic parsing be more useful?

Please explain how can the semantic meaning of a sentence be represented formally. Consider an example sentence such as "All children like toys." You may use words instead of symbols.

Please describe the approach to semantic parsing using Abstract Meaning Representation (AMR).

Please describe the approach to semantic parsing employing the SQL database language. Which do *you* think is more effective (AMR vs. SQL), and why? (No correct answer)

Question 31

LSTMs, which stand for Long Short-Term Memory networks, are a special kind of Recurrent Neural Networks. Answer the following questions:

RNNs are known for capable of dealing with variable-length input and producing variable-length output, making them suitable for a variety of NLP tasks using sequential text data. How do RNNs remember past information?

Vanilla RNNs, also known as Simple Recurrent Networks (SRNs), are known to have "short term memory", "vanishing gradient problem", and "exploding gradient problem". Answer these sub-questions.

Explain what these three terms mean and describe what about the vanilla RNN network subjects the network to these issues.

Briefly describe the impact of these issues on the network's performance, including referring to at least one specific NLP task

Explain how LSTMs solve these three issues. Specifically, describe what is special about the LSTM architecture and how it improves SRNs.

GRU (Gated Recurrent Unit) is a variant of LSTM. Describe how it is different.

Bidirectionality is often added to RNNs such as LSTM, GRU, and SRN. Explain: 1) how it is implemented in practice, 2) what is its advantage, including reference to at least one practical NLP task

RNNs are often used as the encoder and decoder of a sequence to sequence (seq2seq). Could bidirectionality also be used for the encoder and the decoder? If yes, given an example of a specific NLP task that takes advantage of this. If not, explain why.

Question 32

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others. Please answer the following questions related to BERT.

- 1) What is the key innovation for BERT? (3 points)
- 2) Please give two examples of using BERT for NLP tasks? (4 points)
- 3) What's GLUE result? (2 points)
- 4) BERT and other pre-trained language models are extremely large and expensive, how are companies applying them to low-latency production services? (2 points)

Question 33

A. Describe the difference between *recurrent* neural networks and *recursive* neural networks. In particular, which can be thought of as a generalization of the other?

B. Suppose you start running into the “vanishing gradient problem” with a vanilla recurrent neural network. What RNN variants should you try to fix the problem?

C. Suppose you want to write a neural network that predicts the sentiment of sentences on the basis of the dependency parse of the sentence. Would you want to use a *recurrent* neural network or a *recursive* neural network? Justify your choice.

D. Suppose you have some time-series data of stock prices, and want to use it to predict the future movement of stock prices. Would you want to use a *recurrent* neural network or a *recursive* neural network? Justify your choice.

E. Describe briefly how backpropagation works for a recursive neural network in terms of standard backpropagation (i.e. for a multilayer perceptron). (2)

Question 34

b. Perform the following feature structure unifications. If the feature structures do not unify, write “FAIL” (2 points).

b1.

CAT NP

CAT NP

AGR NUM 3rd U AGR GEND F
PERS SG

b2.

CAT NP CAT VP
AGR NUM 1st U AGR NUM 1st
PERS PL ARI 2

c. Consider the following feature structures:

CAT VP CAT VP
NUM 3rd PERS SG
PERS SG TNS PLUP

Draw each feature structure as a directed acyclic graph. If the feature structures unify, draw the unification of the feature structures. If not, explain why they do not unify (4 points).

Question 35

Applying Naive Bayes to WSD

Recall that $P(c)$ is the prior probability of a given sense of a word (counted in a labeled training set). $P(w|c)$ is the conditional probability of a word given a particular sense, such that $P(w|c) = \text{count}(w, c) / \text{count}(c)$. We can also generalize this paradigm to look at features besides words (i.e. $P(f|c)$, the conditional probability of a feature given a sense).

Consider the following data and vocabulary:

	DOC	WORDS	CLASS
TRAINING	1	elephant smoked elephant	e
	2	elephant line	e
	3	elephant smoked	e
	4	flute jazz line	f
TEST	5	line flute jazz jazz	?

$V = \{\text{elephant, smoked, line, flute, jazz}\}$

Recall the following formulas:

$$\hat{P}(c) = \frac{N_c}{N} \quad \hat{P}(w|c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

And assume the following prior probabilities:

$$P(e) = 2/3$$

$$P(f) = 1/3$$

Part A: Calculate the following conditional probabilities:

$$P(\text{line}|e) =$$

$$P(\text{flute}|e) =$$

$$P(\text{jazz}|e) =$$

$$P(\text{line}|f) =$$

$$P(\text{flute}|f) =$$

$$P(\text{jazz}|f) =$$

Part B: Comparing $p(e|d5)$ and $p(f|d5)$, choose the appropriate class.

Question 36 (922)

Use the following documents for this problem, where the frequency of word appearing (and not just the word's presence) in a document matters.

D1 = "cat, dog, fox"

D2 = "fish, tiger, cat"

D3 = "cat, fox, dog"

D4 = "fish, cat, fish"

You may find it helpful to transform the documents into frequency vectors using the table below:

	fish	cat	fox	tiger	dog
D1					
D2					
D3					
D4					

What is the Jaccard Similarity between D1 and D2?

What is the Euclidean Distance between D3 and D4?

Suppose we decide to use Cosine Similarity. Which Document is most similar to D2?

What are the ranges (in general, not for this particular problem) of the above similarity and distance functions: Jaccard Similarity, Euclidean Distance and Cosine Similarity? You may assume that all document vectors only consist of non-negative components.

Question 37

A Naive Bayes classifier has to decide whether document number 5 ‘London Paris’ is news about the United Kingdom (class U) or news about Spain (class S). You can think of documents 1 through 4 as independent Bernoulli trials.

	document	class
1	London Paris	U
2	Madrid London	S
3	London Madrid	U
4	Madrid Paris	S

(a) Estimate the probabilities that are relevant for this decision from the following four documents. Answer with fractions.

(b) Based on the estimated probabilities, which class does the classifier predict? Assume a uniform prior distribution over the classes U and S. Explain your answer, showing that you have understood the Naïve Bayes classification rule, and show your work.

(c) Practical implementations of a Naïve Bayes classifier often use log probabilities. Explain why.

Question 38

Probabilistic Context Free Grammars (PCFG), by design, make certain independence assumptions that can hurt the performance of a natural language parser. Techniques such as parent annotation, constituent splitting, and vertical markovization are often used to mitigate such shortcomings. In this example, we will split the NP constituent into NP-SUBJ and NP-OBJ.

Let’s consider the following probabilities:

$$p = P(\text{NP} \rightarrow \text{PRP})$$

$$n = 1 - p$$

$$p_s = P(\text{NP-SUBJ} \rightarrow \text{PRP})$$

$$n_s = 1 - p_s$$

$$p_o = P(\text{NP-OBJ} \rightarrow \text{PRP})$$

$$n_o = 1 - p_o$$

Notes: PRP is a personal pronoun (such as “I”, “she”, and “they”). NP-SUBJ is a noun phrase that serves as the subject of the sentence. NP-OBJ is a noun phrases that serves as the object of the sentence.

Assume a PCFG trained on the Wall Street Journal portion of the Penn Treebank. Which of the following inequalities is/are likely to be accurate? Pick 0, 1, or more answers.

a. <

- b. >
- c. >
- d. >

Question 39

Which one of the following can be considered as a "universal function approximator"?

- a. a two-layer neural network
- b. a hidden markov model
- c. a push-down automaton
- d. a finite-state automaton

Question 40

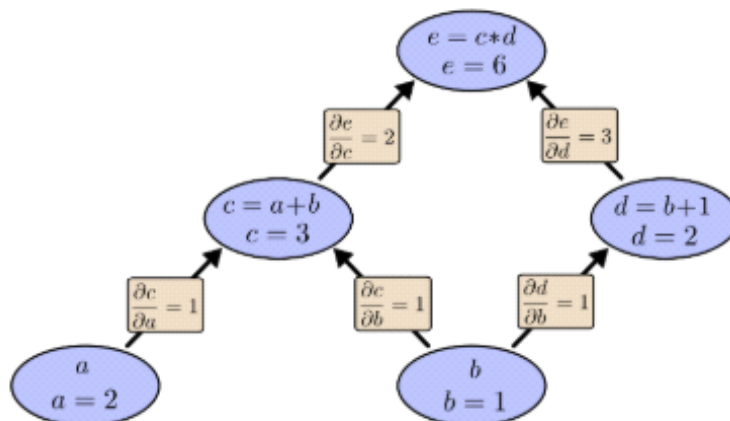
What is the difference between the Viterbi algorithm and the Forward algorithm for POS tagging? Give pseudo-code for both and also explain the difference in plain English.

Question 41

Compute the derivative of the logistic function:

Question 42

For the given computational graph below, calculate the following partial derivatives:



Question 43

In the following table we consider four candidates for **when**, which is the most likely correction? Why?

		c	c)	
deea	deer	a r	0.0001	0.003
deea	idea	de id	0.0007	0.0009
deea	dear	ea ar	0.0003	0.007
deea	deea		0.85	0.00000006

Question 44

Explain each of these terms and give an example (or formula, if appropriate) of each:

- LCS (lowest common subsumer)
- confusion matrix for binary classification
- the semantic compositionality principle
- negative sampling (for word embeddings)
- backpropagation over time for RNN (recurrent neural networks)
- HMM trellis
- softmax function
- horizontal markovization
- labeled dependency accuracy

Question 45

How does shift-reduce constituent parsing work?

Question 46

What is the update formula for a perceptron used to compute the value of w at time $i+1$ from the value of w at time i ?

$$w^{(i+1)} = \dots \dots w^{(i)} \dots \dots$$

Question 47

Why does the CKY parsing algorithm for CFG require the grammar to be in CNF?

Question 48

Write a CCG lexical entry for a transitive verb such as “watch” in the following sentence: “Akira watched the movie”.

Question 49

Describe how an LSTM works. Draw a diagram and explain all gates and all forward propagation formulas.

Question 50

1. The following sentence contains a syntactic ambiguity. Draw two possible parse trees for the sentence.

astronomers saw stars with ears

2. Assume the following probabilistic rules.

$S \rightarrow NP VP$	1.0	$NP \rightarrow NP PP$	0.4
$PP \rightarrow P NP$	1.0	$NP \rightarrow \textit{astronomers}$	0.1
$VP \rightarrow V NP$	0.7	$NP \rightarrow \textit{ears}$	0.18
$VP \rightarrow VP PP$	0.3	$NP \rightarrow \textit{saw}$	0.04
$P \rightarrow \textit{with}$	1.0	$NP \rightarrow \textit{stars}$	0.18
$V \rightarrow \textit{saw}$	1.0	$NP \rightarrow \textit{telescopes}$	0.1

Calculate the probability of each tree that you drew. According to your results, which parse is more probable?

Question 51

Fill in the Probabilistic CKY chart below for sentence: *Time flies like an arrow*

Assume the following rules and weights:

- 1 $S \rightarrow NP VP$
- 6 $S \rightarrow V_{st} NP$
- 2 $S \rightarrow S PP$
- 1 $VP \rightarrow V NP$
- 2 $VP \rightarrow VP PP$
- 1 $NP \rightarrow Det N$
- 2 $NP \rightarrow NP PP$
- 3 $NP \rightarrow NP NP$
- 0 $PP \rightarrow P NP$

	Time	flies	like	an	arrow
	1	2	3	4	
0	NP 3 Vst 3				
1		NP 4 VP 4			

2			P 2 V 5		
3				Det 1	
4					N 8

Extra credit: recover the best parse by tracing back-pointers.

Question 52

Calculate the TF*IDF for the terms listed below for documents 1 to 4. There are 10,000 documents in a collection. The number of times each of these terms occur in documents 1 to 4 as well as the number of documents in the collections are listed below. Use this information to fill in the TF*IDF scores in the table below.

Number of Documents Containing Terms:

- *reverse cascade*: 3
- *full shower*: 50
- *half bath*: 10
- *multiplex*: 3

Term Frequencies				
	Documents			
	Doc 1	Doc 2	Doc 3	Doc 4
reverse cascade	8	10	0	0
full shower	3	1	2	2
half bath	0	0	8	7
multiplex	2	2	2	9

First calculate the IDF of the mentioned terms; then fill the following table:

TFIDF for terms in documents				
	Documents			
	Doc 1	Doc 2	Doc 3	Doc 4
reverse cascade				
full shower				
half bath				
multiplex				

Question 53

Complete the following CCG derivation.

- (a) replace the four instances of ??? with words that form an appropriate sentence.
 (b) Fill in the rest of the derivation in the empty space.

--	--	--	--

????
NP

????
((S \ NP) / NP)

????
(NP / N)

????
N

S

Part 2: Multiple Choice

Question 54

Consider the noun-noun phrases such as "cat food", "baby goat", "attorney general". What is/are some reasonable CCG parses for these:

- (a) $N/N \ N \rightarrow NP$
- (b) $N \ N \backslash N \rightarrow NP$
- (c) $N/N \ N/N \rightarrow NP$
- (d) both a and b

Question 55

Consider the sentence "Bill drives a Honda from Albany to New York." Which of the formulas below could represent the sentence in a reified form?

- (a) $\exists w, x, y: \text{Driving}(\text{Bill}, w, x, y)$
- (b) $\exists z: \text{Driving}(\text{Bill}, \text{Honda}, \text{Albany}, \text{New York})$
- (c) $\exists w, x, y, z: \text{Driving}(\text{Bill}, \text{Honda}, \text{Albany}, \text{New York})$
- (d) $\exists w, x, y, z: \text{Driving}(w, x, y, z)$
- (e) None of the above.

Question 56

Represent the following sentence "No person is immortal." in First-Order Predicate Calculus (FOPC):

- (a) $\neg \exists x: \text{Person}(x) \wedge \text{Mortal}(x)$
- (b) $\neg \exists x: \neg \text{Mortal}(x)$
- (c) $\exists x: \text{Person}(x) \wedge \text{Mortal}(x)$
- (d) $\neg \exists x: \text{Person}(x) \wedge \neg \text{Mortal}(x)$
- (e) $\neg \exists x: \text{Mortal}(x)$

Question 57

Adverbs are used to specify all of the following EXCEPT:

- (a) place
- (b) time
- (c) manner
- (d) degree
- (e) agent

Question 58

The hypernym and hyponym relations in WordNet hold between which of the following notions (pick one):

- (a) words
- (b) lemmas
- (c) stems
- (d) synsets

Question 59

The dependency representation of the sentence "Jane has a cat" is the following (with the arrows pointing from parent to child node):

- (a) has -> cat, has -> Jane, cat -> a
- (b) cat -> has, Jane -> cat, a -> cat
- (c) has -> Jane, Jane -> cat, cat -> has
- (d) Jane -> has, has -> a, a -> cat
- (e) has -> Jane, Jane -> cat, a -> cat

Question 60

What is the Levenshtein edit distance between "apples" and "pears"? Assume the following costs: insertion=1, deletion=1, substitution=1.

- (a) 2
- (b) 5
- (c) 4
- (d) 1
- (e) 3

Question 61

Which of the following statements about syntactic constituents is false?

- (a) Constituents are non-crossing.
- (b) Each word is a constituent.
- (c) If two constituents share one word, then one of them must completely contain the other.
- (d) Constituents are continuous.
- (e) Constituents cannot be nested.

Question 62

Assuming that exactly two constituents get combined at each iteration, a sequence of three nouns can be parenthesized in two different ways:

(a(bc)) and ((ab)c).

A sequence of four nouns can be parenthesized in five different ways:

((ab)c)d (a(bc))d (ab)(cd) a((bc)d) a(b(cd)).

In how many ways can a sequence of five nouns (or, alternatively, an adjective followed by four nouns) be parsed?

- (a) 28
- (b) 8
- (c) 14
- (d) 16
- (e) 10

Question 63

Which of the following statements is true?

The operator ">" here means "strictly more expressive (powerful) than".

CCG = Combinatory Categorical Grammar, CSG = Context Sensitive Grammar, CFG = Context Free Grammar, TAG = Tree Adjoining Grammar, TSG = Tree Substitution Grammar.

- (a) CSG > CFG > TAG
- (b) CSG > TSG > CFG
- (c) CSG > CCG > CFG
- (d) CCG > CSG > CFG
- (e) CSG > CFG > CCG

Question 64

Consider the corpus:

cat cat cat dog dog rat rat bat bat bat bat bat fox

Using "Add One" Laplacian smoothing, what is the estimate for $P(\text{rat})$?

- (a) $2/14$
- (b) $3/14$
- (c) $3/18$
- (d) $3/16$
- (e) $3/19$

Question 65

A regular die has 6 sides, numbered from 1 to 6. If you throw two regular dice together, what is the probability that their sum is an even number?

- (a) $1/4$
- (b) $1/18$
- (c) $1/6$
- (d) $1/3$
- (e) $1/2$

Question 66

Which of the following statements about evaluation of relation extraction is **not** correct?

- (a) Precision = #correctly extracted relations / #all extracted relations
- (b) F1 = harmonic mean of Precision and Recall
- (c) Recall = #correctly extracted relations / #all existing relations
- (d) F1 = arithmetic mean (average) of Precision and Recall

(e) $F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$

Question 67

Which of the following part of speech categories is open class?

- (a) adjectives
- (b) prepositions
- (c) interjections
- (d) articles
- (e) conjunctions

Question 68

What does this logical expression mean in English?

$\exists e: \text{Arriving}(e) \wedge \text{Arriver}(e, \text{Speaker}) \wedge \text{Destination}(e, \text{NewYork}) \wedge \text{IntervalOf}(e, i) \wedge \text{EndPoint}(i, p) \wedge \text{Precedes}(p, \text{Now})$

- (a) I am arriving in New York
- (b) He is arriving in New York
- (c) I will arrive in New York
- (d) I arrived in New York
- (e) She will arrive in New York

Question 69

The CKY (Cocke-Kasami-Younger) parsing algorithm only works when...

- (a) the grammar only includes a single production for any non-terminal.
- (b) the grammar has been converted to Chomsky Normal form.
- (c) the input sentence has exactly one verb.
- (d) the input sentence is in English.
- (e) the input sentence is not ambiguous.

Question 70

In the absence of any other relevant information, how should an out of vocabulary (OOV) word be tagged?

- (a) determiner
- (b) adverb
- (c) noun
- (d) verb

(e) adjective

Question 71

The sentence "Stolen painting found by tree" exhibits what phenomenon:

- (a) prepositional phrase attachment ambiguity
- (b) coordinating conjunction attachment ambiguity
- (c) syntactic ambiguity
- (d) all of the above
- (e) none of the above

Question 72

What part of speech is *least likely* after an article?

- (a) noun
- (b) adjective
- (c) verb
- (d) numeral

Question 73

In the Bayes formula: $P(H|E) = P(E|H)P(H)/P(E)$, what do "H" and "E" stand for?

- (a) H=hypothesis, E=evidence
- (b) H=hyperparameter, E=evidence
- (c) H=hyperparameter, E=estimate
- (d) H=hypothesis, E=estimate
- (e) none of the above

Question 74

In a large corpus, the frequencies of the three most frequent words are approximately 10%, X%, and 3%. What is the value of X, assuming a Zipfian distribution?

- (a) 9
- (b) 7
- (c) 6
- (d) 5
- (e) 4

Question 75

In language modeling, the "add-1" method is an example of:

- (a) linear interpolation
- (b) backoff
- (c) smoothing
- (d) caching
- (e) hypothesis testing

Question 76

An experiment was done to measure the perplexity of unigram, bigram, and trigram models on a news corpus.

Which of the following sets of numbers makes the most sense?

- (a) unigram 1000, bigram 200, trigram 100
- (b) unigram 100, bigram 100, trigram 100
- (c) unigram 100, bigram 200, trigram 1000
- (d) unigram 100, bigram 0, trigram 0

Question 77

For a sentence with n words, the maximum number of boxes in the CKY table that can be non-empty is:

- (a) $n*n*n$
- (b) n
- (c) $n \log n$
- (d) $n*(n-1)/2$
- (e) $n*(n+1)/2$

Question 78

Which of the following sentences exemplifies "type coercion" in sentence parsing:

- (a) I saw a cat in the park.
- (b) I slept.
- (c) I had a tea in the morning.
- (d) I gave Mary a pretzel.
- (e) Get out!

Question 79

Which of the following sentences is non-projective:

- (a) The non-callable issue, which can be put back to the company in 1999, was priced at 99 basis points above the Treasury's 10-year note.
- (b) John saw a dog yesterday which was a Yorkshire Terrier.
- (c) Price details weren't immediately available.
- (d) The collateral is being sold by a thrift institution.
- (e) Ms. Haag plays Elianti.

Question 80

Which of the following relation(s) is/are symmetric:

- (a) brother(X,Y)
- (b) sister(X,Y)
- (c) mother(X,Y)
- (d) both a. and b. above
- (e) none of the above

Question 81

Which of the following propositional logic statements is always true? The symbol "==" here is used to express equivalence.

- (a) $\text{NOT } (A \text{ AND } B) == (\text{NOT } A) \text{ OR } (\text{NOT } B)$
- (b) $\text{NOT } (A \text{ OR } B) == (\text{NOT } A) \text{ AND } (\text{NOT } B)$
- (c) $\text{NOT } A == \text{NOT } B$
- (d) $A == \text{NOT } B$
- (e) more than one of the above

Question 82

Which of the following is true:

- (a) English is an SOV language, Japanese is an SVO language
- (b) English is an SOV language, Japanese is an SOV language
- (c) English is an SVO language, Japanese is an SVO language
- (d) English is an SVO language, Japanese is an SOV language
- (e) English is an SOV language, Japanese is an VSO language

Question 83

The BLEU evaluation metric is essentially:

- (a) n-gram recall with a penalty for brevity
- (b) n-gram recall with a penalty for excess length
- (c) n-gram precision with a penalty for brevity
- (d) n-gram precision with a penalty for excess length
- (e) none of the above