CSCI 3907/6907: Introduction to Statistical NLP
Fall 2019

Midterm Exam

Full Name: ..... GW ID: .....

- ✓ This exam consists of 8 questions (6 for undergrads) and 1 bonus question.
- ✓ Total points: 113 points (95 for undergrads) plus 10 bonus points.

Question 1. True or False [19 pts]

[T] After applying Laplace smoothing, the probability distribution is still valid (that is, the probabilities sum up to 1).

[F] Modern statistical machine translation systems rely on an intermediate meaning representation called Interlingua.

[T] A trigram language model makes a Markov assumption that a word is conditionally independent of past context given the two words that appear before it.

[T] PCFG rules impose independence assumptions that can result in poor probability estimates.

[F] Dependency structures are the same as constituency trees.

[T] Ambiguity in NLP refers to the problem of having multiple, alternative linguistic structures for the same segment.

[T] A confusion matrix is a way to analyze the performance of a system and detect labels that are confused with each other.

[F] Regular expressions are not equivalent to Finite State Automata (FSAs)

[F] Perplexity is a type of intrinsic evaluation of language models, where higher perplexity correlates with higher probability.

[F] The Naive Bayes classifier models dependencies (correlations) between input features.

[T] Cross Validation is used when we don't have much labeled data to separate into train and test sets.

[F] Using parts of the test set for model tuning helps reduce overfitting.

[F] BLEU is an evaluation metric for machine translation that is based on n-gram precision and recall.

[F] The problem of add-one smoothing is that unseen events don't get enough probability mass.

[T] Words are made up of meaningful subunits called morphemes

[T] In sequential classification, a sequence of input units (e.g. words or characters) is mapped to a sequence of corresponding outputs.

[F] Inflectional morphology refers to the combination of a stem and a grammatical morpheme that usually results in a word of a different class (e.g. fuzzy and fuzziness).

[T] Chunking is the process of identifying non-overlapping segments of texts that correspond to major constituent types

[T] Attachment ambiguity and coordination ambiguity are two types of structural ambiguity

Question 2: Short answers [16 pts]

i. How many aligned phrases can be extracted from the following alignment matrix?

2 phrases

ii. What is the difference between intrinsic and extrinsic evaluation? What are the pros and cons of each?

Intrinsic evaluation uses an evaluation metric to assess the performance. Extrinsic evaluation implements the task within other application and evaluate whether the system improves. Intrinsic might not reflect the impact of the task on real world applications. Extrinsic evaluation on the other hand takes long time.

iii. Why do we need a brevity penalty in BLEU score?

To penalize short sentences in the target language.

iv. Calculate the Naïve Bayes prior probabilities for a training set with 100 positive examples, 25 negative examples, and 50 neutral examples

P(positive)=100/175=0.57 P(negative)=25/175=0.14 P(neural)=50/175=0.29

v. What is the difference between lemmatization and stemming?

Lemmatization extracts the lemmas of words such that the generated lemmas have meanings in the languages and can correspond to the head dictionary. Stemming, on the other hand, chop off affixes according to certain rules and the generated stems do not necessarily have meanings.

vi. Describe how a back-off model is used to alleviate the problem of 0 n-gram counts.

In a back-off model, every time you had a zero count for an n-gram, you would back-off and use the counts for a lower n-gram. For example, if the count for a trigram is 0, you'd use the bigrams instead. If the bigram was 0 also, you could back off to unigram counts.

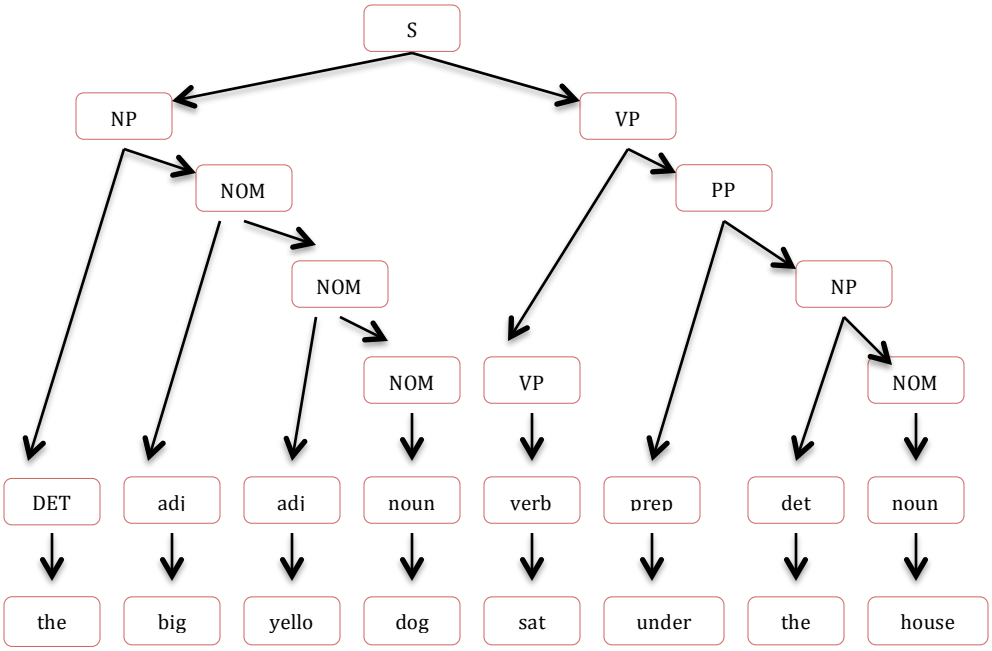
vii. What is the difference between a FSA and a FST?
FSA → Finite State Automata; FST → Finite State Transducer

FSA recognizes inputs while FST recognizes input and generates output

viii. Assuming the grammar below, show the parse tree for the sentence:
the big yellow dog sat under the house

S → NP VP
VP → VP PP
VP → verb NP
VP → verb
NP → DET NOM
NOM → ADJ NOM
NOM → NOUN
PP → PREP NP

DET → the
ADJ → big | yellow
NOUN → dog | house
VERB → sat
PREP → under



Question 3: N-gram Language Models [10 pts]

For the following questions, assume we are using a corpus completely summarized by the unigram counts below (thus |V| = 20):

brown **29**	missing **12**	short **28**
fox **34**	napkin **9**	options **13**
lazy **18**	cheap **22**	car **14**
dog **1**	fork **10**	concinnity **0**
plenty **41**	nickel **1**	neat **49**
tree **1**	chocolate **5**	syzygy **33**
skim **4**	syrup **9**	

SUM 333

i. What are the following probabilities? (Answer as a fraction or a whole number (e.g., "1/2" or "1")): **[2.5 points]**

P_{MLE}(short) = 28/333=0.08
P_{MLE}(concinnity) = 0/333 = 0

ii. Assume we are using Laplace smoothing. What are the following probabilities? **[2.5 points]**

$P_{Laplace}$(lazy) = 18+1/333+20=19/353
$P_{Laplace}$(concinnity) = 0+1/333+20=1/353

iii. Assume instead that we are using Good Turing smoothing. What probability do we assign to things with zero frequency in our training data? **[1 point]**

Probability of words that have been seen once

iv. Now assume that the above counts were just drawn from a larger corpus, and on that full corpus we collected the following counts of counts: **[4 points]**

N1 112849	**N6** 754
N2 41018	**N7** 283
N3 15608	**N8** 104
N4 5704	**N9** 37
N5 2111	**N10** 14

With this data, what are the smoothed counts c* for the following words (assuming the unigram counts given above are still valid)? C*=(c+1)*(Nc+1/Nc)

c*(skim) = (5)*(N5/N4)= 5*(2111/5704)=1.85 (c=4)

c*(syrup) = (10)*(N10/N9)= 10*(14/37)=10*0.378=3.78 (c=9)

Question 4: [10 pts] Write one or more regular expressions (with substitutions) to substitute some words in English with the corresponding forms. Assume input does not contain numeric, punctuation, or special characters.

Consider the following rules:

- For words that begin with a consonant letter, the initial consonant letter is moved to the end of the word, and "ay" is added:
  - pig latin → igpay atinlay
  - banana → ananabay
  - spain → painsay
- For words that begin with vowel, just add "way" to the end
  - egg → eggway
  - inbox → inboxway
  - a → away

s/^([^aeuio])(.*)/\2\1ay/g
s/^([aeuio].*)/\1way/g

Question 5: [20 pts] Describe the complete specifications of a Hidden Markov Model for a chunking system that identifies noun phrases using IOB Tagging. Show a diagram of the HMM and calculate all probabilities using the following training data.

Training Corpus:

the/BNP morning/INP flight/INP from/O Denver/BNP has/O arrived/O.

I/BNP need/O a/BNP flight/INP from/O Denver/BNP to/O Chicago/BNP.

book/O the/BNP morning/INP flight/INP to/O Chicago/BNP.

States: BNP (8 times), INP (5 times), O (8 times)

Observations: the, morning, flight, from, Denver, has, arrived, I, need, a to, Chicago, book

Transition Probabilities	BNP	INP	O
Start	2/3	0	1/3
BNP	0/8	3/8	3/8
INP	0/5	2/5	3/5
O	6/8	0/8	1/8

Likelihood Probabilities	BNP	INP	O
the	2/8	0	0
morning	0	2/5	0
flight	0	3/5	0
from	0	0	2/8
denver	2/8	0	0
has	0	0	1/8
arrived	0	0	1/8
I	1/8	0	0
need	0	0	1/8
a	1/8	0	0
to	0	0	2/8
Chicago	2/8	0	0
book	0	0	1/8

Question 6: [20 pts] Given the following probabilistic CFG:

S → NP VP	0.5	PRN →they	1.0
S → NP AUX VP	0.5	N → hunting	0.4
NP →N	0.4	N → dogs	0.6
NP → N N	0.3	V → hunting	0.6
NP → PRN	0.3	V→ are	0.4
VP → V NP	1.0	AUX → are	0.5
		AUX → is	0.5

Convert the grammar to Chomsky Normal Form *[Hint: the probability of a new dummy variable should be 1.0]*

S → NP VP	0.5	PRN →they	1.0
S → X1 VP	0.5	N → hunting	0.4
X1 → NP AUX	1.0	N → dogs	0.6
NP → hunting | dogs	0.4	V → hunting	0.6
NP → N N	0.3	V→ are	0.4
NP → they	0.3	AUX → are	0.5
VP → V NP	1.0	AUX → is	0.5

i. Fill in the CKY chart below for the sentence "**they are hunting dogs"** using the PCFG above to find the most likely parse tree. Draw the most likely parse tree and report its probability.

	they	*are*	*hunting*	*dogs*
	1	2	3	4
0	1. PRN (1.0) 2. **NP** (0.3)	**X1 (NP AUX)** → (0.3)*(0.5)*(1.0) =0.15	S (X1 VP) → (0.15)*(0.24)*(0.5)= 0.018	1. **S (X1 VP)** → (0.15)*(0.24)*(0.5)=**0.018** 2. S (NP VP) → (0.3)*(0.0288)*(0.5)= 0.00432
1		1. **AUX** (0.5) 2. V (0.4)	VP (V NP) → (0.4)*(0.4)*(1.0)=0.16	VP (V NP) → (0.4)*(0.072)*(1.0)= 0.0288
2			1. N (0.4) 2. NP (0.4) 3. **V** (0.6)	1. NP (N N) → (0.4)*(0.6)*(0.3)=0.072 2. **VP (V NP)** → (0.6)*(0.4)*(1.0)=0.24
3				1. N (0.6) 2. **NP** (0.4)

!! NOTE Question 7 and 8 are NOT required for undergraduate students (and won't be considered for bonus points).

Question 7 [8 pts]:

i. **True or False [6 points]**

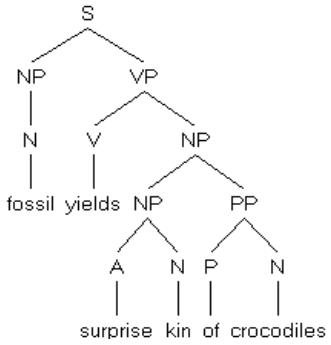
[T] Parent annotation is used to model dependencies between lexical heads.

[T] In arc-standard dependency parsing, if a word is asserted as a dependent using either LEFTARC or RIGHTARC operation, it's removed from the stack and is no longer available for additional relations.

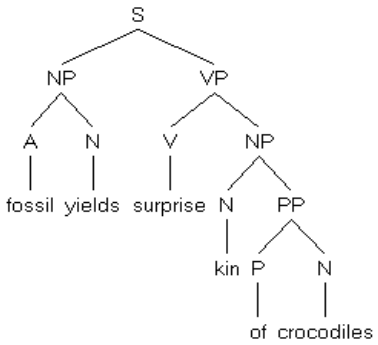
[T] Expectation maximization can be used for unsupervised parameter estimation in machine translation, POS tagging, and syntactic parsing.

[F] The root of a dependency tree has at most one incoming arc.

ii. [2 points] Given the following hypothesis and reference constituency parse trees, calculate labeled precision and recall.



(Reference)



(Hypothesis)

Reference	**Hypothesis**
S → NP VP	S → NP VP
NP → N	NP → A N
N→ fossil	A → fossil
VP → V NP	N → yields
V → yields	VP → V NP
NP → NP PP	V → surprise
NP → A N	NP → N PP
A → surprise	N → kin
N → kin	PP → P N
PP → P N	P → of
P → of	N → crocodiles
N → crocodiles	

Labeled precision: 5/12

Labeled recall: 5/11

Question 8: **Expectation Maximization for IBM Model 1** [10 pts]

We apply EM algorithm to estimate the parameters for a Foreign to English Machine Translation system. We use a simplified version of IBM Model 1 such that we do not consider NULL word or alignments in which English words do not align with any word in the source language.

Assume that you have the following two parallel sentences as your training corpus:

Foreign:	das Haus	das Buch	ein Buch
English:	the house	the book	a book

Given the below translation probabilities initialized with uniform distribution,

(a) Apply one iteration of EM algorithm.

(b) What is the most likely alignment for each parallel sentence in the training corpus?

	das	**Haus**	**Buch**	**ein**
the	1/4	1/4	1/4	1/4
house	1/4	1/4	1/4	1/4
book	1/4	1/4	1/4	1/4
a	1/4	1/4	1/4	1/4

das Haus the house (1/4)*(1/4)=(1/16)/(2/16)=1/2	das Haus the house (1/4)*(1/4)=(1/16)/(2/16)=1/2
das Buch the book (1/4)*(1/4)=(1/16)/(2/16)=1/2	das Buch the book (1/4)*(1/4)=(1/16)/(2/16)=1/2
ein Buch a book (1/4)*(1/4)=(1/16)/(2/16)=1/2	ein Buch a book (1/4)*(1/4)=(1/16)/(2/16)=1/2

	das	**Haus**	**Buch**	**ein**
the	½+½=1/2	½/2=1/4	½/2=1/4	0
house	½	½	0	0
book	½/2=1/4	0	½+½=1/2= 1/2	½/2=1/4
a	0	0	½	½