CS 274A Natural Language Processing (Spring 2023), Midterm Exam

Instructions

- Time: 10:15–11:55am (100 minutes)
- This exam is closed-book, but you may bring one A4-size cheat sheet. Put all the study materials and electronic devices (with the exception of a calculator) into your bag and put your bag in the front, back, or sides of the classroom.
- You can write your answers in either English or Chinese.
- Two blank pieces of paper are attached, which you can use as scratch paper. Raise your hand if you need more paper.
- For multiple choice questions:
  - □ means you should mark ALL choices that apply;
  - ○ means you should mark exactly ONE choice;
  - When marking a choice, please fill in the bubble or square COMPLETELY (e.g., ● and ■). Ambiguous answer will receive no points.
  - For each question with □ choices, you get half of the points for selecting a non-empty proper subset of the correct answers.

1 Text Normalization (6 pt)

1.1 Subword Tokenization (2 pt)

Consider a Byte-Pair Encoding tokenizer C learned inside space-separated tokens tokenized by a white-space based tokenizer W, which **ONE** of the following statements is **NOT** correct?

○ When the vocabulary size of C is unlimited, C is guaranteed to produce the same segmentation as W on the training set.

○ When the vocabulary size of C is unlimited, C is guaranteed to produce the same segmentation as W on the test set.

○ Compared with W, C can reduce data sparseness and out-of-vocabulary issues by segmenting rare words into common subword units.

○ C can produce different segmentation depending on different hyperparameters (e.g., the vocabulary size) and different training sets. W and character tokenizers, which produce single-character segmentation, are unaffected by different training sets.

1.2 Regular Expression (2 pt)

Choose all of the following in which the regular expression accepts the string. "Accept" means all characters are matched. For example, for a RE "abc", string "abc123" is rejected because "123" is not matchable.

□ RE: "(\w+)(\d+(\.\d+)?)" String: "ShanghaiTech10.0"

□ RE: "abc+" String: "abcccc"

- ☐ RE: "(abc)+" String: "abcccc"
- ☐ RE: "f\d+:\d+(\.\d+)?" String: "f1:99.0"

1.3 Word Normalization (2 pt)

Which **ONE** of the following statements is correct?

- ○ Stemming: the stem of a word must be a substring of it. That is $X_1X_2 \dots X_n \xrightarrow{\text{stemming}} X_iX_{i+1} \dots X_j, 1 \leq i \leq j \leq n$ where X_i is the i-th character in the word.
- ○ Lemmatization chops off affixes crudely. For example, the lemma of "running" is "runn".
- ○ Affixes (词缀) are the core meaning-bearing units in morphemes (词素).
- ○ Word normalization can help reduce the appearance of unknown words on test set.

Solution:
1.1 B

Solution:
1.2 ABD

Solution:
1.3 D

2 Text Representation (14 pt)

2.1 Word Vector Zoo (3pt)

For the five different word vector models that we covered in the lectures, choose their key properties.

1. Please select the all the **sparse** word vectors:
   - ☐ One-hot vectors
   - ☐ PPMI vectors
   - ☐ BERT embedding
   - ☐ Word2vec skip-grams
   - ☐ LSA
2. Please select the all the **static** word vector:
   - ☐ One-hot vectors
   - ☐ PPMI vectors
   - ☐ BERT embedding
   - ☐ Word2vec skip-grams
   - ☐ LSA
3. Please select all the vectors obtained by training using neural networks:
   - ☐ One-hot vectors
   - ☐ PPMI vectors
   - ☐ BERT embedding
   - ☐ Word2vec skip-grams
   - ☐ LSA

Solution:
2.1.1 AB

Solution:
2.1.2 ABDE

Solution:
2.1.3 CD

2.2 Multiple choice (4 pt)

1. Select one or more correct statements.

   ☐ One-hot vectors can be used to compute the similarity between two words.

   ☐ Apply Add-k smoothing to PMI calculation will give frequent words slightly lower probabilities (k is small).

   ☐ One of the reasons we compute idf in tf-idf is that we want to reduce the effect of overly frequent words like "the", "it", or "they" that are not very informative about the context.

   ☐ In a word-word co-occurrence matrix, choosing a smaller window size (±1-3) can help represent more semantic information while choosing a larger window size (±4-10) can help represent more syntactic information in general.

2. Select one or more correct statements.

   ☐ Dense SVD compresses data in a sparse matrix into fewer dimensions.

   ☐ Skip-grams produces static word embedding.

   ☐ In CBOW, we use context words to predict center words.

   ☐ We can apply mean pooling over corresponding word embeddings to obtain phrase/sentence representations.

Solution:
2.2.1 BC

Solution:
2.2.2 ABCD

2.3 Skip-grams (7 pt)

In word2vec skip-grams, $\boldsymbol{u}_w$ denotes the word vector of context word w, and $\boldsymbol{v}_w$ denotes the word vector of center word w. Given a super tiny corpus "I love NLP" (yes, the corpus contains only 3 words), answer the following questions.

1. Choose the correct formula of $P(\text{NLP}|\text{love}) =$:

   ○ $\frac{e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{u}_{\text{NLP}}}}{e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{u}_{\text{NLP}}} + e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{u}_{\text{love}}} + e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{u}_{\text{I}}}}$

   ○ $\frac{e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{v}_{\text{NLP}}}}{e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{v}_{\text{NLP}}} + e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{v}_{\text{love}}} + e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{v}_{\text{I}}}}$

   ○ $\frac{e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{v}_{\text{NLP}}}}{e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{v}_{\text{NLP}}} + e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{v}_{\text{love}}} + e^{\boldsymbol{v}_{\text{love}}^T \boldsymbol{v}_{\text{I}}}}$

   ○ $\frac{e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{u}_{\text{NLP}}}}{e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{u}_{\text{NLP}}} + e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{u}_{\text{love}}} + e^{\boldsymbol{u}_{\text{love}}^T \boldsymbol{u}_{\text{I}}}}$

2. Suppose the window size m is 1, fill the blank to complete the formula of the **likelihood** of the corpus. Note 1: you should only use probabilities such as $P(\text{NLP}|\text{love})$ and math operations to form the formula. Note 2: when you do not have left context ($t = 1$), you only need to include the probability of predicting the right context word, and it's similar when you do not have right context ($t = 3$).

   $L(\theta) = \prod_{t=1}^{T} \prod_{\substack{m \le j \le m \\ j \neq 0}} P\left(w_{t+j} \mid w_t; \theta\right)$

   $= \boxed{} \times P(\text{I}|\text{love}) \times \boxed{} \times P(\text{love}|\text{NLP})$

3. Select all the correct statements about negative sampling in Skip-grams:

□ With negative sampling, we compute $P(\text{co-occur}|c, o)$ instead of $P(o|c)$

□ With negative sampling, we compute softmax over K negative samples plus one target sample to compute the probability

□ We use sigmoid activation instead of softmax with negative sampling to compute the probability of co-occurrence.

□ Our goal is to maximize the probability of context words and center word co-occurring.

4. Denote the size of vocabulary as $|V|$. Please select all the reason(s) why we should avoid computing softmax over the entire vocabulary in word2vec?

□ The time complexity of computing softmax is pretty high, which is $O(|V|^2)$.

□ When using skip-grams for training, the vocabulary size is often too large.

□ It is inaccurate to compute softmax over the entire vocabulary, because it ignores word order or context information.

Solution:
2.3.1 A

Solution:
2.3.2 $P(\text{love}|\text{I}), P(\text{NLP}|\text{love})$

Solution:
2.3.3 AC

Solution:
2.3.4 B

3 Naïve Bayes (20 pt)

In this section, we will use Naïve Bayes to analyse a language with only 4 words: A, B, C and D.

3.1 Text Classification (11 pt)

Suppose we have already known the classes (+/-) of the following samples:

No.	Class	Sentence
1	+	A B D
2	-	C D B A C
3	-	B C
4	+	B D C
5	-	B C A B

3.1.1 Generative or Discriminative (2pt)

Recall that discriminative models draw boundaries in the data space, while generative models try to model how data is placed throughout the space. Naïve Bayes classifier is

○ a generative classifier.

○ a discriminative classifier.

○ neither a generative classifier nor a discriminative classifier.

Solution:
A

3.1.2 Add-1 Smoothing (4pt)

Suppose we are doing add-1 smoothing. Calculate the likelihoods $P(w|c)$ from the training samples above. Recall the likelihoods come from the following formula:

$$p(w_i \mid c) = \frac{\text{count}(w_i, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c)\right) + |V|}$$

We have already done some calculations for you. Keep 2 decimal points in your answers.

$$P(A|+) = 0.20, P(A|-) = 0.20$$
$$P(D|+) = 0.30, P(D|-) = 0.13$$

$P(C|+) =$ [ ] $P(C|-) =$ [ ]

Solution:
0.20, 0.33

3.1.3 Prediction (3pt)

Suppose we are doing add-1 smoothing. Which class might the sentence 'C A D' belongs to? Remember to consider the pirors!

- ○ The positive class (+).
- ○ The negative class (-).
- ○ Equally likely for the two classes.

Solution:
B

3.1.4 Evaluation (2pt)

Suppose we have the following system predictions and their gold labels:

Prediction	Gold	Sentence
+	+	A C B D
+	-	A C B
+	-	A C
-	+	A B D C
+	-	A C A B

What is the F1-score of these predictions?

- ○ $\frac{1}{2}$
- ○ $\frac{3}{8}$
- ○ $\frac{1}{3}$
- ○ $\frac{1}{4}$
- ○ $\frac{1}{5}$

Solution:
C

3.2 Text Clustering (9 pt)

Now, suppose we do not know the classes of the samples. We want to run unsupervised Naïve Bayes to group these samples into 2 clusters.

No.	Sentence
1	A B D
2	C D B A C
3	B C
4	B D C
5	B C A B

3.2.1 The E step (3pt)

Recall in E step,

$$P\left(y_j = i \mid x_{j,1:w}, \theta^{(t)}\right) \propto \pi_i^{(t)} \prod_{k=1}^{w} P\left(x_{j,k} \mid \psi_i^{(t)}\right)$$

Given the model parameters below, perform one E step

$$p_{1,A} = 0.20, p_{2,A} = 0.30$$
$$p_{1,B} = 0.20, p_{2,B} = 0.10$$
$$p_{1,C} = 0.10, p_{2,C} = 0.40$$
$$p_{1,D} = 0.50, p_{2,D} = 0.20$$
$$\pi_1 = 0.2, \pi_2 = 0.8$$

where $p_{i,w}$ is the probability that word w appears in a sentence of cluster i, π_i is the prior probability that a sentence belongs to cluster i. Let θ be the model parameters, $P(y_i = j|\theta)$ be the probability that the ith sample belongs to the jth cluster, given the model parameters. We have provided the results of the first two samples. What is the label distribution of the third sample? Keep 2 decimal points in your answer.

$$P(y_1 = 1|\theta) = 0.45, P(y_1 = 2|\theta) = 0.55$$
$$P(y_2 = 1|\theta) = 0.05, P(y_2 = 2|\theta) = 0.95$$

$$P(y_3 = 1|\theta) = \boxed{}$$

Solution:
0.11

3.2.2 The M step (2pt)

For unsupervised Naïve Bayes, we

- ○ assign sentences proportionately to different clusters in M step.
- ○ revise each cluster based on its proportionately assigned words in M step.
- ○ minimize the log likelihood of each word given its sentence in M step.
- ○ take a cup of coffee and do nothing in M step.

Solution:
B

3.2.3 Evaluation (2pt)

Suppose we have the following clustering result:

- Predict: {1, 3, 5}, {2, 4}
- Gold: {1, 4}, {2, 3, 5}

Recall that in B-cubed metric, we compute a precision and a recall for each element, then average the individual precisions and recalls. What is the F1-score of the B-cubed metric for this result?

- ○ $\frac{2}{5}$
- ○ $\frac{7}{15}$
- ○ $\frac{8}{15}$
- ○ $\frac{3}{5}$

Solution:
C

3.2.4 LDA (2pt)

In topic modeling, Latent Dirichlet Allocation (LDA) may outperfom unsupervised Naïve Bayes. Below is an example of topics inferred by LDA:

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Linda argues that LDA cannot produce the table as shown above. Why?

- ○ Because LDA is not a model for text clustering. It requires the topic for each document in training.
- ○ Because LDA cannot group words. It could only group documents. Each document belongs to one specific topic.
- ○ Because LDA is not an efficient algorithm such that it is impossible to use it in practice.
- ○ Because LDA could group words, but it cannot summarize the topic names (the first line in the table).

Solution:
D

4 Language Modeling (12 pt)

4.1 Perplexity (4 pt)

John and Peter are designing language models on the same dataset. John proposes a model that predicts all words with the same probability. If John uses a vocabulary size of 1024 (including unknown token), on the test set he will reach a perplexity of

- ○ 0.0
- ○ $\frac{1}{1024}$
- ○ 1.0
- ○ 1024.0
- ○ $+\infty$

Peter trains a model using another vocabulary and achieves a perplexity of 9.0 on the test set. Decide which model is better.

- ○ John's model performs better because the perplexity is lower.
- ○ Peter's model performs better because the perplexity is lower.
- ○ John's model performs better because the perplexity is higher.
- ○ Peter's model performs better because the perplexity is higher.
- ○ We cannot tell which one performs better because the vocabulary is different.

Solution:
D, E

4.2 N-gram (5 pt)

The following is a toy corpus:

```
I love NLP
NLP is useful
I love CS274A
```

1. Without any smoothing, P(`NLP`|`love`) = _____.
2. Using Laplace smoothing with parameter $\lambda = 1$, P(`NLP`|`love`) = _____.
3. If the next word is only conditioned on the previous word, the model is a
   ○ unigram model. ○ bigram model. ○ trigram model.

Solution:
0.5, 0.25, B

4.3 Recurrent Neural Networks (RNN) (3 pt)

Which of the following statements is/are correct?

- □ Compared with fixed window neural language models, RNN language models can better remember history information.
- □ By scaling up the gradients when they are small, we can solve the vanishing gradient problem.
- □ Due to vanishing gradient, RNN language models are better at learning from syntactic recency than sequential recency.
- □ LSTMs can completely solve the vanishing gradient problem.

Solution:
A

5 Attention & Transformer (12 pt)

5.1 Attention (3 pt)

Which of the following statements is/are correct?

☐ The computational complexity of dot product attention scales linearly with sequence length.

☐ Attention in Transformer is efficient because it can be parallelized over positions.

☐ For the RNN with attention model taught in class, when we are at time step t and predicting the $(t+1)$-th word, the query vector is computed from the t-th hidden state and the key vectors are computed from the hidden states from time step 1 to $t-1$.

Solution:
B

5.2 Attention in language modeling (6 pt)

John implemented and trained a transformer for language modeling himself. He found the model can predict all the next words correctly when training, but has a poor performance when testing. He then printed the attention weight matrix in the training phase to inspect what is going on.

		Keys			
		<s>	I	Love	NLP
Queries	<s>	0.99	0.01	0.00	0.00
	I	0.00	1.00	0.00	0.00
	Love	0.02	0.00	0.97	0.01
	NLP	0.00	0.00	0.00	0.99

Table 1: Attention Weight Matrix

1. According to the attention weight matrix, decide what causes the poor testing performance.
   - ○ John forgot to add regularization to the parameters.
   - ○ John forgot to mask the unseen words.
   - ○ The model is trapped in a local minimum.
   - ○ Transformer is not good at language modeling.
2. After fixing the bug in 1, John found the variance of the attention weights is too large (all weights are close to either 0 or 1). If something is wrong in his attention module, which of the following is most likely?
   - ○ John forgot to scale down QK^T.
   - ○ John forgot to scale down V.
   - ○ John forgot to scale up QK^T.
   - ○ John forgot to scale up V.

Solution:
B, A

5.3 Transformer Components (3 pt)

Which of the following statements is/are correct?

□ Multi-head self-attention takes self-attention over the original input vectors multiple times and then concatenates the attention outputs together.

□ In linear attention, attention is always a valid distribution.

□ Layer normalization may reduce uninformative variation to make training more stable.

□ Attention is permutation invariant if no position embedding is used in the model. Permutation invariance means that when the input words are permuted, the outputs are also permuted in the same way.

Solution:
CD

6 Seq2Seq & Pretrained Language Model (16 pt)

6.1 Seq2Seq (3 pt)

Which of the following statements is/are correct?

□ In language modeling, beam search decoding performs at least as good as greedy decoding.

□ Cross-attention allows the decoder to look directly at source when decoding.

□ Length normalization is needed in decoding because longer sequences have a higher probability.

□ In cross-attention, the attention is computed using the queries from the decoder sequence and the keys from the encoder sequence.

Solution:
BD

6.2 Pretraining Tasks (3 pt)

Which of the following pretrained models is supposed to have the best performance in masked language modeling.

○ ELMo

○ BERT

○ GPT

Solution:
B

6.3 PLM Representations (3 pt)

Which of the follow representation is not suitable for a span represetation.

○ Embedding of the first token in a sentence.

○ Bilinear mapping of the embeddings of the start and end token of a span.

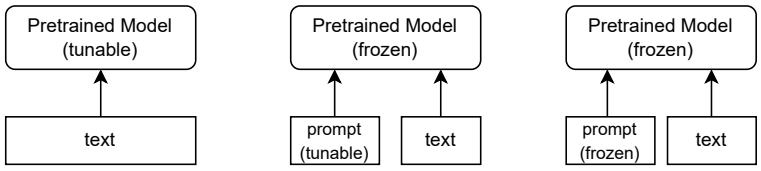
○ Max pooling of embeddings of the tokens in the span.

Solution:
A

6.4 Utilizing PLM (4 pt)

1. In each of the three blanks below, write A, B or C corresponding to the name of the method of using PLMs.

   A. Fine tuning
   B. Prompt engineering
   C. Prompt tuning



_____ _____ _____

2. In each of the three blanks below, write A, B or C to indicate the prompting method used in each case.

   A. Naive prompting
   B. In context learning
   C. Chain-of-thought prompting

```
Q: Jane Smith teaches mathematics at Harvard University.
  Jane Smith: professor

  Lines of code scroll past on the screen as he types furiously.
  he: programmer

  He winds up his right leg and strikes the ball with incredible force.
  He:
```

_____

```
Q: He winds up his right leg and strikes the ball with incredible force.
  What's "He"'s job?
A:
```

_____

```
Q: He winds up his right leg and strikes the ball with incredible force.
  What's "He"'s job?
A: Let's think step by step.
```

_____

Solution:
ACB, BAC

6.5 PLM Types (3pt)

Write $\surd$ in the cell if the model is of the corresponding type. Note that one model only has one type.

Solution:

	Encoder	Decoder	Encoder-Decoder
BERT			
GPT			
BART			
T5			
ChatGPT			

	Encoder	Decoder	Encoder-Decoder
BERT	√		
GPT		√	
BART			√
T5			√
ChatGPT		√	

7 Sequence Labeling (20 pt)

7.1 Choice Questions (10pt)

1. Which one of the following statements about Sequence Labeling is **NOT** correct?

   ○ Closed class words are usually function words. Open class words are usually content words.

   ○ Consider a label sequence $[..., y_i, y_{i+1}, ...]$ in sequence labeling with BIOES tagging scheme, y_i and y_{i+1} can be any label in $\{B, I, O, E, S\}$.

   ○ RNN models can be used in sequence labeling.

   ○ Transformer models can be used in sequence labeling.

2. Which of the following statements about Hidden Markov Models (HMMs) is/are correct?

   □ The algorithm that we use to find the best sequence of hidden states given a sequence of observations in HMMs is called the forward-backward algorithm.

   □ The time complexity of the forward-backward algorithm is $O(N^2L)$ where we denote N as the number of hidden states and L as the length of the sequence.

   □ HMMs can be used to handle the POS Tagging task.

   □ We do not have to normalize the transition matrix in HMMs before we run the Viterbi algorithm or the forward-backward algorithm.

3. Which of the following statements about Max-Entropy Markov Models (MEMM) and Conditional Random Fields (CRF) is/are correct?

   □ MEMM considers contextual information in the sentence.

   □ MEMMs solve some of the issues of HMMs. In particular, it does not suffer from the label bias problem.

   □ CRF uses global normalization instead of local normalization.

   □ One way to optimize CRF in supervised learning is to minimize the margin-based loss.

4. Which of the following statements is/are correct?

   □ MEMM is a globally normalized model.

   □ Emission scores in Neural CRFs can be computed by neural networks.

- □ CRFs can **NOT** be trained in an unsupervised manner.
- □ A bidirectional RNN can utilize the context of each word to predict its label.
- □ A transformer can utilize the context of each word to predict its label.

5. Consider an HMM with transitions and emissions shown in the tables below. Please select the optimal sequence of hidden states for the word sequence "Natural Language".

$\mathbf{Y}_{t-1}$	$\mathbf{P}(\mathbf{Y_t}\|\mathbf{Y}_{t-1})$		
	H_0	H_1	STOP
START	0.3	0.7	N/A
H_0	0.3	0.5	0.2
H_1	0.5	0.3	0.2

W	$\mathbf{P}(\mathbf{W_j}\|\mathbf{H}_i)$	
	H_0	H_1
Natural	0.4	0.7
Language	0.6	0.3

- ○ (H_0, H_0)
- ○ (H_0, H_1)
- ○ (H_1, H_0)
- ○ (H_1, H_1)

Solution:
7.1.1 B

Solution:
7.1.2 BC

Solution:
7.1.3 ACD

Solution:
7.1.4 BDE

Solution:
7.1.5 C

7.2 True or False (10 pt)

1. Time complexity of the Viterbi algorithm is higher than the forward-backward algorithm.
2. We can use back-propagation to compute expected counts in an HMM.
3. The Baum-Welch algorithm (EM for HMMs) will eventually converge to a global optimum.
4. An HMM is a directed graphical model, while a CRF is an undirected graphical model.
5. HMMs can be used for language modeling tasks.
6. CRFs suffer from the label bias issue.
7. An MEMM is an undirected graphical model.
8. Named Entity Recognition (i.e., recognizing entity spans in a sentence) can often be cast as a sequence modeling task.
9. By manipulating the transition matrix in a CRF, we can prevent certain labels from transiting to others.
10. We have to use some hand-designed features to make neural CRFs work.

Write down your answer (T/F):

1	2	3	4	5
6	7	8	9	10

Solution:
F T F T T F F T T F