CS 274A Natural Language Processing (Spring 2024)
Final Exam

Instructions

- Time: 10:15–11:45am (90 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 (7 pt)

1.1 Regular Expression (2 pt)

Given the text below:

```
The company announced a revenue increase of $5.2 million (surpassing
expectations by 15%), while the research and development team, including
Dr. Jane Smith, successfully launched the highly anticipated product,
earning a 4.5-star rating from customers!
```

Choose all of the following regular expressions that accept at least one substring in the text. Suppose we are using Python's `re.search` function.

□ `$\d.\d`

□ `(\w+)`

□ `\d+%`

□ `customers.`

Solution:
BCD

1.2 Byte-Pair Encoding (3 pt)

A recent study finds that even in popular LLMs, there might be hundreds of unreachable tokens in the vocabulary. That is, no input string can produce these tokens. To see how this could happen, consider the following BPE merge operations (which is produced by a post-processing method not covered in class): `(7, _), (2, 7), (7, 4), (2, 74), (2, 7_)`. The initial vocabulary is `{_, 2, 7, 4}`. Which one of the following tokens is unreachable after these merge operations?

- ○ `7_`
- ○ `27`
- ○ `74`
- ○ `274`
- ○ `27_`

Solution:
D

1.3 Word Normalization (2 pt)

Lily finds a strange sentence in a text corpus: "Tim and Tom were play football, shout at the top of their voice." She thinks the sentence may have been normalized by some strategy. Which one of the following strategies could have been used to normalize the sentence?

- ○ case folding
- ○ lemmatization
- ○ stemming

Solution:
C

2 Text Representation (13 pt)

2.1 Co-occurrence Matrices (6 pt)

Given the following term-context co-occurrence matrix for words in a corpus:

	mouse	eyes
dog	1	0
cat	4	5

2.1.1 PPMI (3 pt)

What is the PPMI value for the word pair (dog, mouse)? Suppose we are using the log base 2.

- ○ 0
- ○ 0.1
- ○ 0.167
- ○ 0.693
- ○ 1

Solution:
E

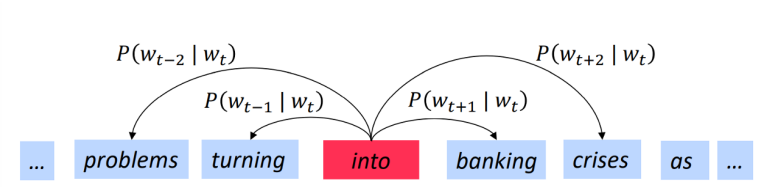


Figure 1: Word2Vec

2.1.2 Weighting PPMI (3 pt)

If we apply laplace (add-2) smoothing to the co-occurrence matrix, will the PPMI value for the word pair (dog, mouse) increase or decrease?

- ○ increase
- ○ decrease
- ○ remain the same

Solution:
B

2.2 Word2Vec (5 pt)

2.2.1 Basics (3 pt)

Given the Word2Vec model in Fig. 1, which of the following statements are true?

- □ This figure illustrates the Skip-gram model instead of the CBOW model.
- □ The context window size is 2.
- □ To calculate $P(w_{t+1}|w_t)$, we only need to know the vector representations of w_t and w_{t+1}.

Solution:
AB

2.2.2 Negative Sampling (2 pt)

Suppose c is a center word and o is a context word. What if we only maximize $P(co-occur|c,o)$ for all c/o pairs in the training corpus without negative sampling?

- ○ The model will be properly trained.
- ○ $P(co-occur|c,o)$ will possibly be negative.
- ○ $P(co-occur|c,o)$ cannot be properly optimized and it will be very small.
- ○ For any word pair w_1, w_2, $P(co-occur|w_1,w_2)$ will be nearly 1.

Solution:
D

2.3 Document Representation (2 pt)

Suppose we have 20 documents and the word "NLP" appears in all the 20 documents. It also appears in the first document for 10 times. What is the TF-IDF value of the word "NLP" in the first document?

- ○ 0
- ○ 0.5
- ○ 1
- ○ 1.5
- ○ 2

Solution:
A

3 Text Classification (10 pt)

3.1 Generative or Discriminative (3 pt)

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

- ○ a generative model.
- ○ a discriminative model.
- ○ neither a generative nor a discriminative model.

Solution:
B

3.2 Classification (3 pt)

Suppose we have a logisitic regression model for binary classification with `w = [0.1 -0.5 0.2]` and `b = 0.3`. Given the input feature vector `x = [1 2 3]`, what is the predicted probability of the positive class?

- ○ 0.5
- ○ 0.7
- ○ 0.8
- ○ 0.9

Solution:
A

3.3 Evaluation (4 pt)

Given the following confusion matrix for a binary classification task:

	Predicted Positive	Predicted Negative
Actual Positive	10	5
Actual Negative	3	12

What is the F1 score?

- ○ $\frac{2}{7}$
- ○ $\frac{3}{7}$
- ○ $\frac{4}{7}$
- ○ $\frac{5}{7}$

Solution:
D

4 Text Clustering (10 pt)

4.1 Expectation-Maximization (3 pt)

Consider a Mixture of Gaussian (MoG) model with k components. We run the EM algorithm on a set of data points. Suppose in one iteration, we revise the mean of all the components. Which step of the EM algorithm is this in?

- ○ E-step
- ○ M-step
- ○ neither E-step nor M-step

Solution:
B

4.2 LDA (3 pt)

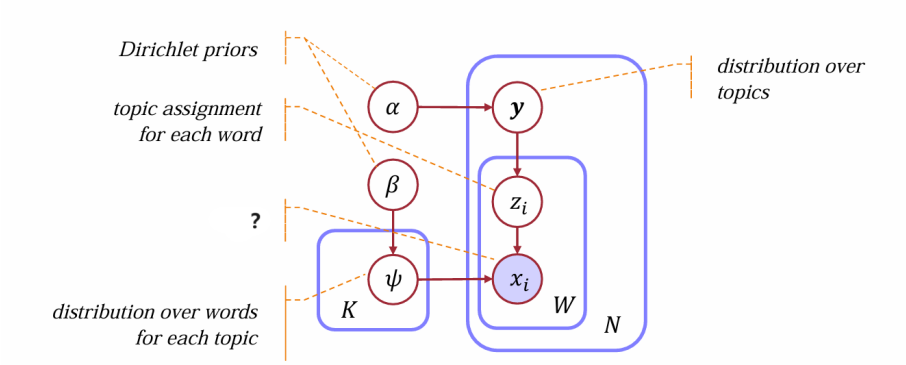


Figure 2: LDA

Figure 2 shows the graphical model of Latent Dirichlet Allocation (LDA). What does x_i stand for in the model?

- ○ an observed word

- ○ an observed topic
- ○ a prior distribution over words
- ○ a latent representation of a document

Solution:
A

4.3 Evaluation (4 pt)

Suppose we have the following clustering results:

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

Which one is larger, purity or inverse purity?

- ○ purity
- ○ inverse purity
- ○ purity and inverse purity are equal

Solution:
C

5 Language Modeling (23 pt)

5.1 N-gram Language Model (8pt)

1. n-gram is not a perfect language model as it
   - □ cannot be used to generate text.
   - □ requires space exponential to n.
   - □ is hard to model long-range dependencies.
   - □ is hard to generalize to words not in the train set.
2. You want to build a n-gram model (with maximum likelihood estimation) based on several sentences and find that there are 10 unique words in these sentences. The following tables show the number of occurrences of the 10 words in these sentences.

the	fox	brown	jumped	...	Total
8	6	4	4	...	32

Table 1: Number of times each word occurs.

x\y	the	fox	brown	jumped	...
the	0	0	0	0	...
fox	3	0	0	1	...
brown	1	0	2	0	...
jumped	0	3	0	0	...
...	...	...	...	...	...
Total	8	6	4	4	...

Table 2: Number of times word x occurs after word y.

Now you need to compute the probability of the following sentence (**please write your answer as a fraction**):

$$\textit{the brown fox jumped}$$

(a) With a uni-gram model only, the probability of the sentence is _____.

(b) With a bi-gram model (with the probability of the first word computed using a uni-gram model), the probability of the sentence is _____.

Based on perplexity, which model is better in this case?

- ○ The uni-gram model.
- ○ The bi-gram models.
- ○ Cannot tell, since their vocabularies are different.

Solution:
BCD, 3/4096, 0, A

5.2 RNN (6pt)

1. Which of the following statements is/are correct?
   - □ Consider a model that predicts each next token based on the current token regardless of previous tokens. This model can be seen as a neural unigram model.
   - □ LSTM is good at modeling long-range dependency as it introduces a mechanism to preserve information through time.
   - □ Exploding gradient is a problem that cannot be solved in RNN.
   - □ Vanishing gradient makes it harder for a model to learn long-distance dependencies.
2. Please select all tasks where Bi-LSTM can be used.
   - □ Sequence labeling.
   - □ Language modeling.
   - □ Sentence classification.
   - □ Masked language modeling.

Solution:
BD, ACD

5.3 Attention (6pt)

1. Which of the following statements is/are correct?
   - □ We need to mask the future tokens in the attention of RNN.
   - □ The attention mechanism solves the information bottleneck problem in RNN and LSTM.
   - □ The attention mechanism uses key vectors to find the queries to attend to.
   - □ The objective of the attention mechanism is to increase the number of parameters.
2. Which of the following may speed up inference on long sentences?
   - □ Change the original attention to linear attention.
   - □ Add relative position encoding to the attention module.
   - □ Use band sparse attention in the attention module, i.e., each token only attends to a small range of nearby tokens.
   - □ Use a stronger GPU.

Solution:
B, ACD

5.4 Transformer (3pt)

Which of the following statements is/are correct?

- ☐ Layernorm is important in the transformer as it normalizes the means and standard deviations of hidden states into 0 and 1.
- ☐ The feed-forward layer in the transformer introduces non-linearity to the model and can be seen as attention to the global memory.
- ☐ An advantage of autoregressive transformers (such as GPT) over RNN is that it can always generate tokens in parallel.
- ☐ Transformer is always better than RNN in all kinds of tasks.

Solution:
AB

6 Seq2Seq (7pt)

6.1 Basics

1. Please select all tasks where Seq2Seq models can be used.
   - ☐ Machine translation.
   - ☐ Document summarization.
   - ☐ Semantic role labeling.
   - ☐ Dialogue generation.
2. Which of the following statements is/are correct?
   - ☐ Beam search consumes more memory than greedy decoding.
   - ☐ The encoder and decoder are two modules, so we need to train them independently.
   - ☐ Naive non-autoregressive decoding achieves generation quality at least as good as autoregressive decoding since it can generate all words simultaneously.
   - ☐ In the seq2seq transformer, attentions to input and output sentences are computed in different attention modules.

6.2 Sampling

We may sample each token instead of using greedy or beam-search decoding to encourage diversity. Please match the following four sampling methods with their effects.

A. $s'_t = s_t + \alpha$ if t is in the input; otherwise $s'_t = s_t$.

B. $s'_t = s_t - \alpha$ if t is in the output; otherwise $s'_t = s_t$.

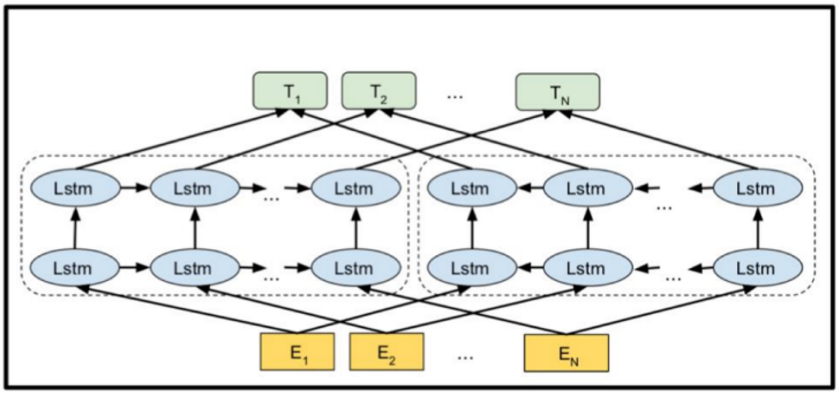
C. $s'_t = s_t$ if $s_t > \alpha$; otherwise $s'_t = -\infty$.

D. $s'_t = s_t - l\alpha$ if t is not the end of sequence token; otherwise $s'_t = s_t$.

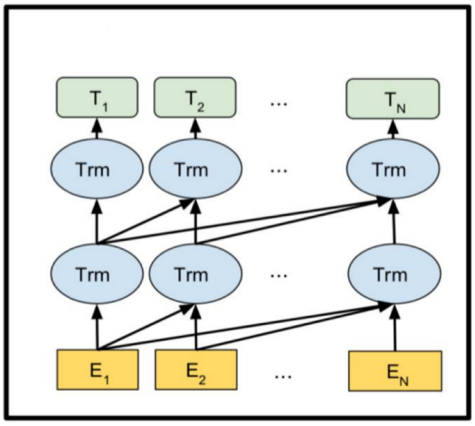
where s_t is the log probability of token t computed by the model, s'_t is a modified score of token t, α is a positive constant, l is the number of tokens that are already generated. The actual token distribution that we sample from at each position is obtained by applying softmax to s'_t. Now please match the following effects with the above sampling methods.

_____ may prevent the model from generating overly long responses.

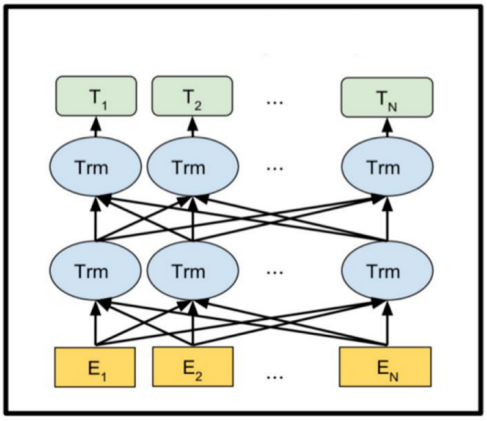
_____ may prevent accidentally generating a token of low probability while preserving some randomness.



(a) Architecture 1



(b) Architecture 2



(c) Architecture 3

Figure 3: Model Architectures

_____ may prevent repeating.

_____ may encourage the model to be faithful to the input.

Solution:
ABCD, AD, DCBA

7 Pre-trained Language Model (10 pt)

7.1 Architecture and Pre-training (4 pt)

1. Given the three pre-trained language model (PLM) architectures in Fig. 3, please select all correct statement(s): (2 pt)

   □ Architecture 1, 2, and 3 refer to ELMo, BERT, and GPT correspondingly.

   □ Language modeling is a common pre-training task for Architecture 2, though it is an encoder-only PLM architecture.

   □ Masked language modeling is a common pre-training task for Architecture 3.

   □ When pre-training encoder-decoder PLMs (not shown in the figure), training tasks are converted into the text-to-text format.

2. Select all correct statement(s) below: (2 pt)

   □ Masked language modeling shows an advantage over language modeling because it is relatively easy. That is why dominant PLMs, such as ChatGPT, use masked language modeling as the pre-training task.

□ Reinforcement Learning with Human Feedback (RLHF) is used in the training procedure of InsturctGPT.

□ Nowadays, PLMs are not limited to Transformer-based models.

Solution:
CD, BC

7.2 Fine-tuning PLM (3 pt)

Select all correct statement(s) below:

□ Prompt Tuning, a method for Parameter Efficient Fine-Tuning (PEFT), freezes models and keeps soft prompt tunable.

□ Most PEFT methods such as LoRA do not introduce extra parameters when fine-tuning the model.

□ When fine-tuning GPT for classification tasks, we do prediction from the output of the special token at the beginning of the sentence, which is the same as done in BERT fine-tuning.

□ The pre-train+fine-tune paradigm is one of the dominant practices in NLP nowadays, replacing previous practice of training models from scratch for specific tasks.

Solution:
AD

7.3 Utilizing PLM (3 pt)

Select all correct statement(s) below:

□ We may choose whether to freeze PLMs or keep their parameters trainable when utilizing them.

□ In vanilla BERT, span representations are usually obtained from embeddings of specific special tokens.

□ Designing prompt, including chain-of-thought prompt, is becoming important to utilizing PLMs.

□ Though Large Language Models (LLMs) are powerful, they still suffer from knowledge hallucination and poor zero/few-shot performance.

Solution:
AC

8 Sequence Labeling (20 pt)

8.1 Hidden Markov Model (HMM) (8 pt)

1. Denote x_i as the token and y_i as the label at position i. Select all correct statement(s) below (4 pt):

   □ Viterbi Algorithm is used to find the most likely label sequence under the model.

   □ Forward Algorithm can be thought of keeping track of the maximum probability of any path to each label at token t, instead of the total probability of all paths.

   □ Compared with Forward Algorithm, Forward-Backward Algorithm costs exactly twice the time when used to compute the marginal probability $P(x_1 \cdots x_n, y_i)$ for a specific position i.

□ $P(x_1 \cdots x_n, y_i) = P(x_1 \cdots x_i, y_i)P(x_{i+1} \cdots x_n | y_i)$ is the key formula to derive Forward-Backward Algorithm. It is based on the conditional independence within HMM.

2. Select all correct statement(s) below (4 pt):

□ In HMM supervised learning, maximum likelihood estimate can be applied and a closed-form solution exists.

□ In E-step of Baum-Welch Algorithm for HMM unsupervised learning, Forward Algorithm is applied to compute the marginal probability of the input sentence, which serves as the normalization factor when computing expected counts.

□ In M-step of Baum-Welch Algorithm for HMM unsupervised learning, simply normalizing expected counts is incorrect.

□ For HMM unsupervised learning, instead of using the EM algorithm, we may directly optimize $P(x_1 \cdots x_n)$ by gradient descent.

Solution:
AD, ABD

8.2 Conditional Random Filed (CRF) (8 pt)

1. Select all correct statement(s) below (4 pt):

□ Max-Entropy Markov Models (MEMM) is a generative model because label y_t of token x_t is dependent on the complete input text.

□ Unlike HMM, MEMM considers contextual information in the input.

□ Viterbi Algorithm can also be used to do decoding for CRF.

□ CRF utilizes local normalization instead of global normalization.

2. Select all correct statement(s) below (4 pt):

□ In CRF supervised learning, we may maximize the conditional (log) likelihood.

□ Another CRF supervised learning loss is the margin-based loss. It can be optimized in exactly the same way as optimizing the conditional (log) likelihood.

□ In CRF unsupervised learning, we may maximize the marginal likelihood $P(x_1 \cdots x_n)$.

□ Another option for CRF unsupervised learning is CRF Autoencoder, which applies a CRF to the input and then predicts each word from its label.

Solution:
BC, AD

8.3 Neural models (4 pt)

Please choose true (T) or false (F) for each statement.

1. In neural softmax, changing from BiLSTM to Transformer can solve the problem that relations between neighboring labels are not utilized.

○ T ○ F

2. The inference and learning of neural CRF is similar to that of traditional CRF.

○ T ○ F

Solution:
F, T

9 Constituency parsing (20 pt)

9.1 Evaluation (3 pt)

Suppose we have a corpus with 2 sentences with the gold and predicted parse trees shown in Figure 4. What is the micro-F1 score of the parser?

- ○ $\frac{17}{18}$
- ○ $\frac{17}{19}$
- ○ $\frac{17}{20}$
- ○ $\frac{17}{21}$

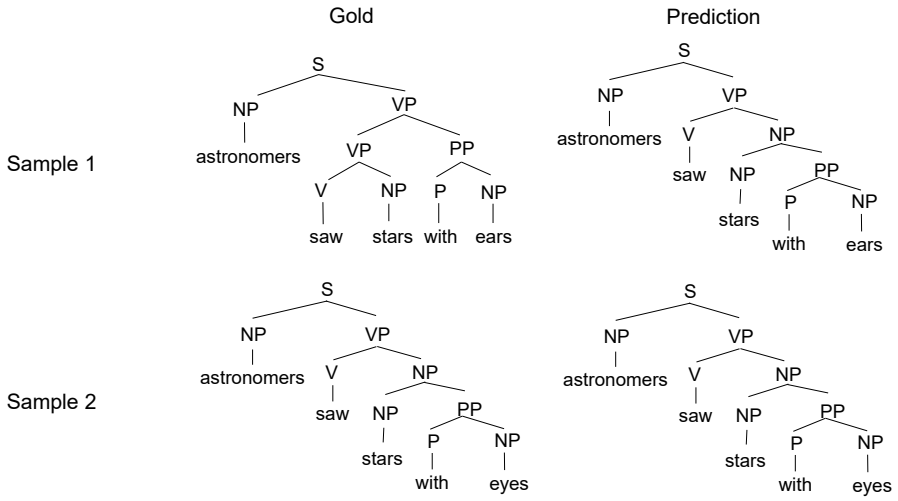


Figure 4: A corpus of constituency parse trees.

Solution:
A

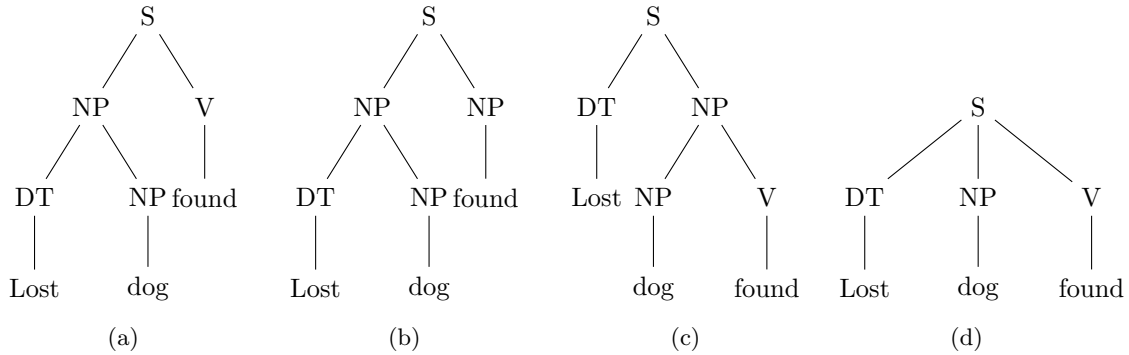
9.2 Span-based Constituency Parsing (3 pt)

Consider the sentence "Lost dog found". Given the span scores in Table 3, which one is the best parse tree?

	Lost	dog	found
Lost	2 (DT)	1 (NP)	0 (S)
dog	-	3 (NP)	2 (NP)
found	-	-	2 (V)

Table 3: Span scores for the sentence "Lost dog found".

- ○ (a)
- ○ (b)
- ○ (c)



○ (d)

Solution:
C

9.3 Context-Free Grammar (8 pt)

Given the grammar below:

```
[0.6] S -> NP Adv      [0.5] N -> time
[0.4] S -> N VP        [0.5] N -> flies
[0.2] NP -> N N        [0.3] V -> time
[0.8] NP -> NP PP      [0.7] V -> flies
[0.4] VP -> V Adv      [0.5] P -> with
[0.6] VP -> VP PP      [0.5] P -> in
[1.0] PP -> P NP       [1.0] Adv -> fast
```

9.3.1 The Rules (3 pt)

John would like to add a rule `S -> VP` to the grammar above. If we do not consider the probabilities, which one of the following statements is true?

○ If we add the rule, the grammar will no longer be context-free.

○ If we add the rule, the grammar is still context-free, but we cannot do parsing with CKY directly with this grammar.

○ If we add the rule, the grammar is still context-free and we can do parsing with CKY directly with this grammar.

Solution:
B

9.3.2 Ambiguous (2 pt)

Consider the sentence "time flies fast". Is this sentence ambiguous under the grammar above?

○ Yes

○ No

○ Cannot determine

Solution:
A

9.3.3 CYK Parsing (3 pt)

What is the probability of the best parse tree for the sentence "time flies fast" with the grammar above? If there is only one possible parse tree, please choose the probability of that parse tree.

- ○ 0.044
- ○ 0.056
- ○ 0.084
- ○ 0.088
- ○ 0.096

Solution:
B

9.4 Inside-Outside Algorithm (3 pt)

Figure 6 illustrates the bottom-up computation process of some probabilities in the inside-outside algorithm. Suppose w_i represents the i-th word in the sentence. Which algorithm does the figure illustrate?

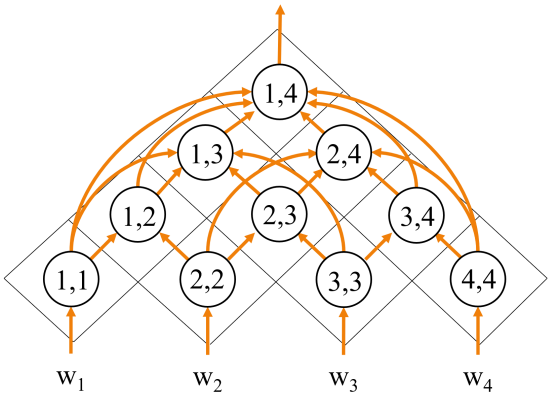


Figure 6: Part of the Inside-Outside Algorithm

- ○ Inside
- ○ Outside
- ○ Viterbi

Solution:
A

9.5 Transition-based Parsing (3 pt)

Suppose in transition-based parsing, we have the following stack and buffer before and after one transition:

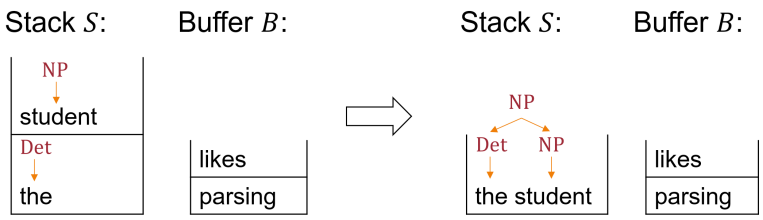


Figure 7: A Transition Step in Transition-based Parsing

Which action does the parser take in this transition?

- ○ SHIFT
- ○ LEFT-ARC
- ○ RIGHT-ARC
- ○ UNARY-NP
- ○ REDUCE-NP

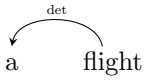
Solution:
E

10 Dependency Parsing (20 pt)

10.1 Basics (8 pt)

10.1.1 Head and Dependent

In the following dependency arc, which word is the head and which word is the dependent



- ○ 'a' is the head, 'flight' is the dependent.
- ○ 'a' is the dependent, 'flight' is the head.

10.1.2 Projectivity

Figure 8: Two dependency parses of the sentence "book me a flight".

Are these dependency parse trees projective?

- ○ (a) is projective; (b) is not projective.
- ○ (a) is not projective; (b) is projective.
- ○ Both of them are projective.
- ○ Neither of them are projective.

10.1.3 Evaluation

Suppose (a) is the predicted parse tree, while (b) is the gold parse tree. What is the Unlabelled Attachment Score (UAS)? What about the Labelled Attachment Score (LAS)?

- ○ UAS = 75%, LAS = 50%
- ○ UAS = 50%, LAS = 50%
- ○ UAS = 50%, LAS = 25%
- ○ UAS = 25%, LAS = 50%

10.1.4 Dependency and Constituency

If we use a context-free grammar to model the dependency parse tree of Figure 8(a), then which of the two constituency trees in Figure 9 (next page) would it produce?

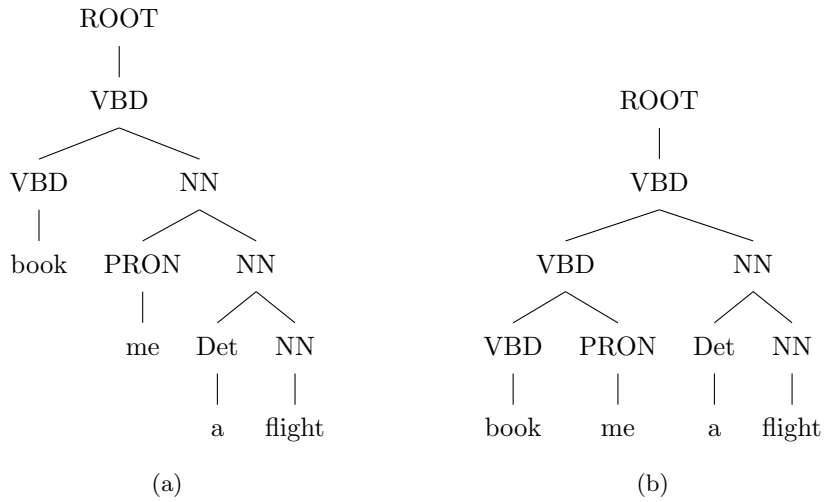


Figure 9: Two constituency parses of the sentence "book me a flight".

- ○ Only (a) is possible.
- ○ Only (b) is possible.
- ○ Both (a) and (b) are possible.

Solution:
B, A, C, B

10.2 Graph-Based Dependency Parsing (8 pt)

Table 4 shows the score matrix for sentence "book me a flight". Answer the following questions.

	book	me	a	flight
ROOT	7	3	1	2
book	-	7	1	8
me	6	-	0	1
a	1	0	-	3
flight	4	0	1	-

Table 4: A score matrix for sentence "book me a flight".

10.2.1 The Eisner Algorithm (1)

Recall in the Eisner algorithm, a triangle represents a partial tree whose root is the word marked by the vertical edge and no words except the root expect more children. What is the score of the triangle covering words "book me", as shown in Figure 10?

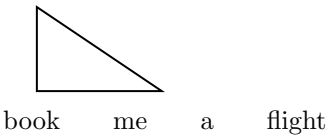


Figure 10: Triangle score in Eisner

○ 0 ○ 3 ○ 6 ○ 7 ○ 8 ○ 10

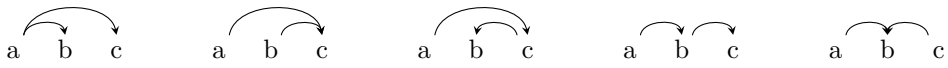
10.2.2 The Eisner Algorithm (2)

Given the arc scores in Table 4, what's the head of "book" if we run the Eisner Algorithm? (Hint: you don't have to really run the Eisner algorithm by hand.)

○ book ○ me ○ a ○ flight ○ ROOT

10.2.3 Second-order Graph-based Dependency Parsing (2pt)

In second-order graph-based dependency parsing, each connected pair of arcs has a score. The tree score is the sum of arc-pair scores. Siblings and grandparents are allowed in second-order graph-based dependency parsing. Please select all the allowed pairs.



(a) □ (b) □ (c) □ (d) □ (e) □

Solution:
D, E, ACD

10.3 Transition-Based Dependency Parsing (4 pt)

10.3.1 Transitions (2 pt)

What action can be used to convert the source state into the target state as shown in Figure 12.

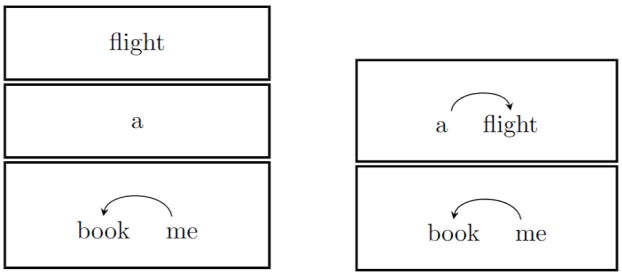


Figure 12: Caption

○ RIGHT-ARC ○ LEFT-ARC ○ SHIFT ○ NONE

10.3.2 Number of Actions (2 pt)

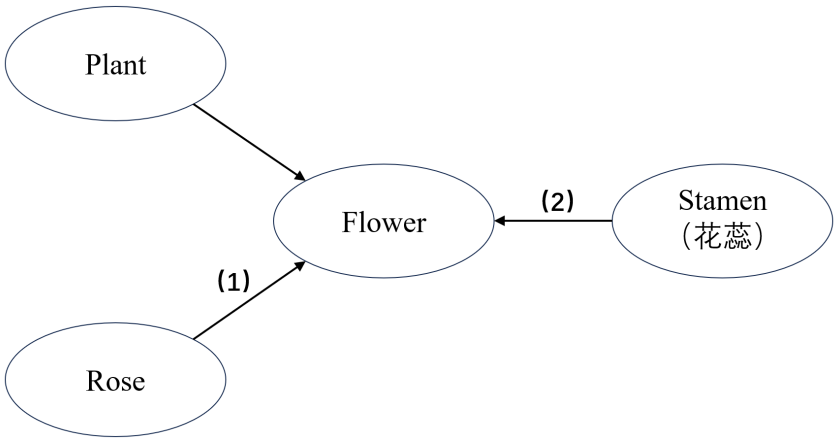
How many actions ('SHIFT', 'LEFT-ARC', 'RIGHT-ARC') should we take to perform transition-based dependency parsing on sentence "book me a flight"?

○ 0 ○ 1 ○ 4 ○ 7 ○ 8 ○ 16

Solution:
A, E

11 Semantics (15 pt)

11.1 Lexical Semantics (3 pt)



(a) WordNet

Noun
① S: (n) **plant**, works, industrial plant (buildings for carrying on industrial labor)
② S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)[(植物学)缺乏运动能力的活生物体]
③ S: (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
④ S: (n) **plant** (something planted secretly for discovery by another)

Verb
⑤ S: (v) **plant**, set (put or set (seeds, seedlings, or plants) into the ground)
⑥ S: (v) implant, engraft, embed, imbed, **plant** (fix or set securely or deeply)
⑦ S: (v) establish, found, **plant**, constitute, institute (set up or lay the groundwork for)
⑧ S: (v) **plant** (place into a river)
⑨ S: (v) **plant** (place something or someone in a certain position in order to secretly observe or deceive)
⑩ S: (v) **plant**, implant (put firmly in the mind)

(b) Word Senses

Figure 13: Lexical semantics

Given the semantic relations of `flower` and Word Senses of `plant` in WordNet as shown in Fig. 13 (next page), please select all correct statement(s):

□ Word embedding is a kind of explicit sentence representation.

□ Relations (1) and (2) in Fig. 13a are Hyponymy (subset; is-a relation) and Meronymy (part-of relation) correspondingly.

□ In the sentences "Jerry plants a thought in the my mind. I should finish planting flowers in the garden beside the power plant before sunset.", The three words, "plants", "planting" and "plant" have word sense ⑩, ⑤, ① correspondingly.

□ Word Sense Disambiguation can be formulated as a sequence labeling task.

Solution:
BCD

11.2 Semantic Parsing (3 pt)

Select all correct statement(s) below:

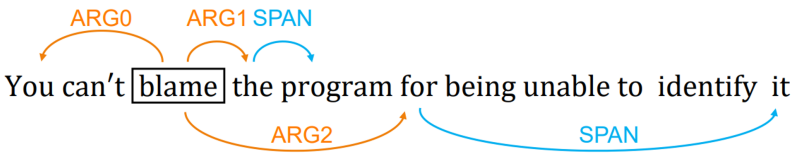
□ In various semantic graphs, nodes are always restricted to represent words.

□ $\forall x, \exists y, \neg Hates(x, y)$ means "There exists something that nobody hates".

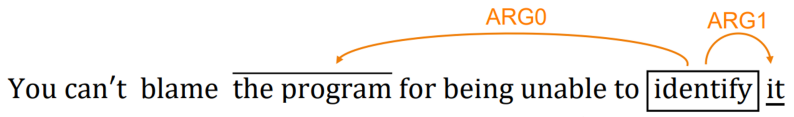
□ $[\lambda f.f(a, b)](\lambda x.\lambda y.Friends(x, y))$ reduces to $[\lambda x.\lambda y.Friends(x, y)](a, b)$ firstly, and finally reduces to $Friends(a, b)$.

□ Neural semantic parsing can be formulated as a parsing-to-graph task or a sequence-to-sequence task.

Solution:
CD



(a)



(b)

You	can't	blame	the	program	for	being	unable	to	identify	it
①	O	B-V	B-ARG1	I-ARG1	B-ARG2	I-ARG2	I-ARG2	②	I-ARG2	I-ARG2
O	O	O	B-ARG0	③	O	O	O	O	④	B-ARG1

Figure 14: Semantic Role Labeling with graph-based methods and sequence labeling.

11.3 Semantic Role Labeling (6 pt)

Please select the correct missing slots to help convert the following two graph-based semantic role labeling results to the sequence labeling format, as Fig. 14 (next page) shows.

①	○ O	○ B-ARG0	○ B-V
②	○ I-ARG2	○ B-V	○ O
③	○ B-ARG2	○ O	○ I-ARG0
④	○ B-V	○ O	○ B-ARG2

Hint: BIO here means beginning of a span, inside of a span, outside of a span correspondingly. The first line of the sequence labeling table corresponds to Fig. 14a, and the second line corresponds to Fig. 14b.

Solution:
BACA

11.4 Information Extraction (3 pt)

Select all correct statement(s) below:

- ☐ Sequence labeling can be applied to solve all kinds of NER tasks.
- ☐ Nested NER can be viewed as constituency parsing with partially-observed trees.
- ☐ Joint entity and relation extraction is a replacement for the "entity-relation" pipeline.
- ☐ Large language models such as ChatGPT cannot be used for information extraction tasks.

Solution:
BC

12 Discourse Analysis (5 pt)

12.1 Discourse Analysis Overview (2 pt)

Please choose true (T) or false (F) for each statement.

1. Discourse parsing can be divided into two stages: elementary discourse units (EDU) segmentation and rhetorical structure theory (RST) parsing.

   ○ T ○ F

2. A hierarchical discourse structure contains only asymmetric relations (represented with curved arrows).

   ○ T ○ F

Solution:
T, F

12.2 Coreference Resolution

12.2.1 Coreference Resolution Overview (2 pt)

Please choose true (T) or false (F) for each statement.

1. Coreference Resolution can be divided into two stages: Mention Detection and Mention Clustering.

   ○ T ○ F

2. Let m_i denote mention i, $q(m_j, m_i)$ denote the probability that mention i and j belong to the same cluster, and $y_{ij} \in \{0, 1\}$ denote the ground truth of mention clustering. Maximizing the probability: $\sum_{j=1}^{i-1} \mathbf{1}(y_{ij} = 1) \cdot q(m_j, m_i)$ is only suitable for learning Mention Ranking, not for other Mention Clustering methods.

   ○ T ○ F

Solution:
T, F

12.2.2 Mention Ranking Inference (1 pt)

Given the sentence "Mary asked Jenny if she could borrow her book because she lost hers yesterday.", the result of Mention Ranking is shown in Fig. 15. We do inference by taking the transitive closure and getting the clusters. Which one is a correct mention cluster below?

○ Mary she her ○ Marry she hers ○ Jenny she ○ NA Jenny her

Solution:
B

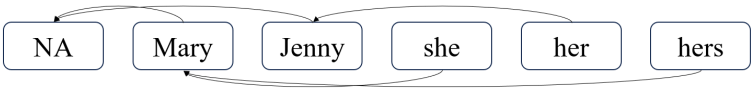


Figure 15: Mention Ranking