Name: _____ Roll: _____

CSE 556: Natural Language Processing (NLP) — EndSem

Date: 28th Apr 2025 **Duration:** 2 hrs. **Max Marks: 60**

---

1. Given a grammar G, produce the **most probable** parse tree using the probabilistic CYK algorithm for the following sentence. **[10]**

We are expecting:

- Appropriate boundary marking **[0.5]**
- Row and column number **[0.5]**
- Complete parsing table with arrows and indices. **[7]**
- Explicit location and justification for probability-based disambiguation during the parsing process. **[1]**
- Most probable parse tree. **[1]**

Note: ***No partial marking*** for the incorrect/incomplete parsing table/parse tree.

Grammar G:

S	→ NP VP	[1.0]
NP	→ PRP	[0.3]
NP	→ DT NN	[0.5]
NP	→ NP PP	[0.2]
VP	→ VBD NP	[0.4]
VP	→ VP PP	[0.6]
PP	→ IN NP	[1.0]
PRP	→ I	[1.0]
NN	→ telescope	[0.5]
NN	→ girl	[0.5]
VBD	→ saw	[1.0]
DT	→ a	[1.0]
IN	→ with	[1.0]

1 I 2 saw 3 a 4 girl 5 with 6 a 7 telescope 8 (Put boundary markers here)

(↓) Row number Column number (→)

	2	3	4	5	6	7	8
1	PRP [1,2] (1.0) --> NP [1,2] (0.3)	-	-	S [1, 5] (1.0)	-	-	S [1, 8]
2		VBD [2,3] (1.0)	-	VP [2,5] (0.4)	-	-	VP [2, 8] (0.4) VP [2, 8] (0.6)
3			DT [3,4] (1.0)	NP [3,5] (0.5)			NP [3, 8] (0.2)
4				NN [4,5] (0.5)	-	-	-
5					IN [5,6] (1.0)	-	PP [5, 8] (1.0)
6						DT [6,7] (1.0)	NP [6, 8] (0.5)
7							NN [7,8] (0.5)

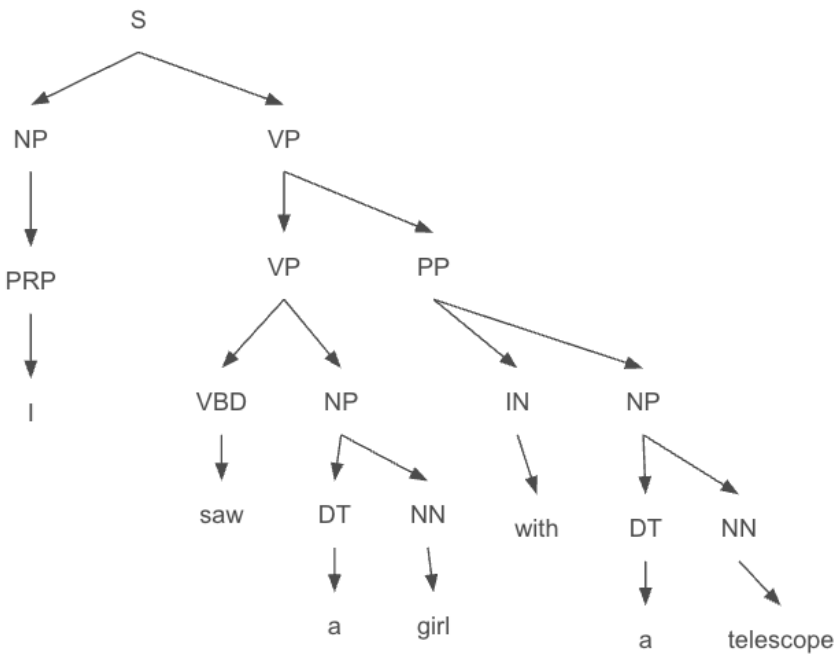
Ignore if someone has not mentioned probability values at each location, but it must be there for cell [2,8], either in the parsing table and/or in the justification part.

Name: _____ Roll: _____

Location and Justification for probability-based disambiguation:

Disambiguation is at location [2,8] in the parsing table. Two reductions of VP are possible with probability (VP → VBD NP) 0.4 and (VP → VP PP) 0.6 respectively. So the winner is with higher probability, ie., (VP → VP PP)

Parse Tree:



2. With reference to fact verification in the TabVer paper, answer **either** (a) or (b) **[10]**

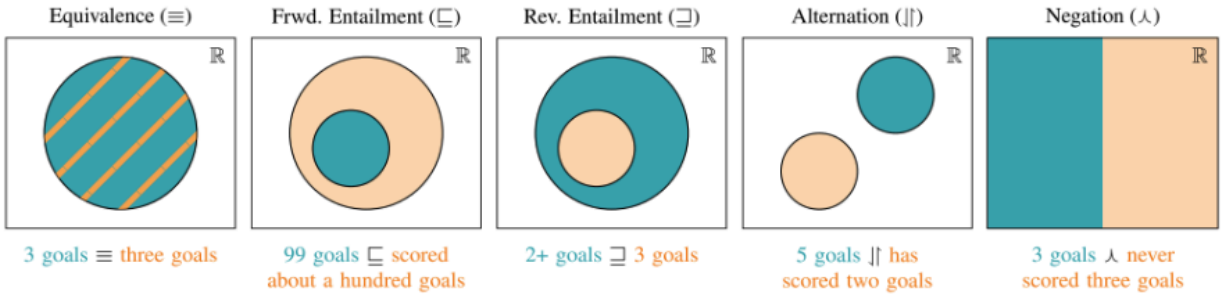
2a. Relations on transition and their names are required. No marks without names. Symbols can be different.

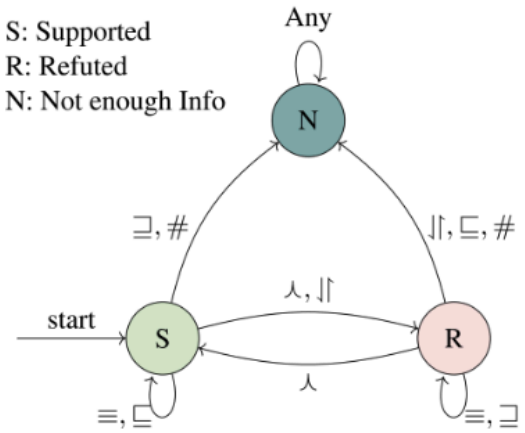
1 mark for each transition (Take any correct 10 relations for marking).

2b. Each relation must be defined and an example should be given. Diagrams are also acceptable.

2 marks for each relation (1 for definition + 1 for example)

Name: _____ Roll: _____





3. What is the objective of the Pointer-Generator Network? [2]

To copy factual details from the input instead of generating it. (Primary- give full marks)

It does so by adding the attention distribution with the vocabulary distribution at the decoding stage. (Optional)

4. In what ways, we evaluate ASR and TTS models? Name them. [2]

ASR: WER (Word-Error-Rate)

TTS: MOS (Mean Opinion Score)

5. Formulate DiloGPT using equations. Each variable must be defined appropriately. [2]

Dialog utterance: x_i N: Number of dialog utterances 0 < m <= N

Source $S = x_1, \cdots, x_m$

Target $T = x_{m+1}, \cdots, x_N$

$$p(T|S) = \prod_{n=m+1}^{N} p(x_n|x_1, \cdots, x_{n-1})$$

Name: _____ Roll: _____

1 mark for the equation and 1 mark for the variables.

6. Categorize the following terms based on societal biases and interpersonal behaviors [0.5*6]
hate speech, ***prejudice***, ***offensive speech***, ***toxic***, ***abuse***, ***aggression***

Societal biases: ***hate speech***, ***prejudice***, ***offensive speech***
Interpersonal behaviors: ***toxic***, ***abuse***, ***aggression***

7. In comparison to GPT-3, mention the main concept used in GPT-3.5. [1]

RLHF (Reinforcement Learning with Human Feedback)

8. Define unsupervised training and its advantages. How is continued pre-training different from it? [2]

Unsupervised Training: In unsupervised training, we leverage a large scale raw data and create training instances without human intervention. Next word prediction (LM, MLM), next sentence prediction, etc., are a few examples of unsupervised training. Advantages: Human effort is not required for training data. **[1 mark]**

Continued Pre-training: Unsupervised pre training usually happens for domain-agnostic data. In continued pretraining, we start with a large-scaled domain-agnostic trained model (load their weights) and perform unsupervised training on relatively small-scale domin data. **[1 mark].**

9. Describe the purpose of negative sampling in the Word2Vec model. Also, give its loss function. [4]

Negative sampling was introduced to avoid the costly denominator computation in the loss function of word2vec. It helps the model to discriminate between semantic unrelated words – they should be far apart from the word under consideration. **[2 marks]**

Either one of these equations: **[2 marks]**

$$\sum_{(c,w)\in D} \log P(D=1\,|\,c,w) + \sum_{(i,w)\in P_n} \log P(D=0\,|\,i,w)$$

$$\sum_{(c,w)\in D} \log P(D=1\,|\,c,w) + \sum_{(i,w)\in P_n} \log\left(1 - P(D=1\,|\,i,w)\right)$$

$$\sum_{(c,w)\in D} \log \sigma(v_c v_w) + \sum_{(i,w)\in P_n} \log\left(1 - \sigma(v_i v_w)\right)$$

$$\sum_{(c,w)\in D} \log \sigma(v_c v_w) + \sum_{(i,w)\in P_n} \log \sigma(-v_i v_w)$$

Name: _____ Roll: _____

10. Explain LoRA. Highlight its **advantages** using an example/diagram. [4]

Fine-tuning an LLMs is difficult due to its high number of parameters. With LoRA (low rank adapter), we attach a secondary/auxiliary network (aka. adapter) to the main LLM network and fine tune the adapter. Furthermore, LoRA introduces the concept of rank which further allows it to reduce the number of learnable parameters. For example, given the dimension of input and output layers of the adapter as M (say, 100) and N (say, 500), respectively, we need MxN (50000) number of learnable parameters (ignoring biases). With rank R (say, 10) << M and N, we need only (M x R + R x N) (1000 + 5000) learnable parameters.

Note: Diagram is also fine for example.

11. Which loss function is typically used for Multiclass classification problem and Multilabel classification problem. Name them and write their equations. [2]

Multiclass classification problem: Categorical Cross-Entropy Loss:

$$div(y_i, d_i) = \sum_i d_i \log d_i - \sum_i d_i \log(y_i) = -\log(y_i)$$

Multilabel classification problem: Binary Cross-entropy loss:

$$div(y_i, d_i) = -d_i \log y_i - (1 - d_i)\log(1 - y_i)$$

0.5 marks for Name + 0.5 for the equation.

12. Differentiate between non-contextual and contextual word embeddings. ***No marks for names only***. [3]

Name: _____ Roll: _____

- In non-contextual embeddings, we obtain the same embedding for a word, irrespective of its usage, context, and/or sense. E.g., Both senses of bank will have the same embeddings. **(Primary)**
- Input to non-contextual embeddings is the word itself.
- Output is the embedding of that word only.

- In contextual embeddings, we can obtain different embeddings for a word depending on its context. **(Primary)**
- Input to contextual embeddings is a sentence or discourse.
- Output is the embedding for each word.

Note: The last two points are not necessary. Give marks if the first point is covered.

13. Explain Masked Self Attention module. Justify its purpose. [2]

Masked Self Attention is used in the decoder. Since we don't know the future tokens (i.e., y(t+i) for the timestep t), we can't compute attention scores for them. Therefore, we apply masking over the future tokens and distribute the probability mass (softmax) over the known tokens only.

14. Between LM vs MLM, which task is easier to perform and why? [2]

LM is a generative task, but MLM is a classification task; hence MLM is much easier to perform.

15. Define n-gram language models. What is the effect of increasing/decreasing the value of n? [3]

N-gram LM: To predict the next token, we consider n previous tokens only. **[1 mark]**

Increasing n: **[1 mark]**

(+): More coherent sentences

(-): It is very difficult to find high frequency counts of longer sequences. Needs a huge amount of training data and time to compute the probability.

Decreasing n: **[1 mark]**

(-): Less coherent sentences

(+): Easy to find sequences and relatively easier to compute.

16. How do we assess the goodness of a smoothing function? [2]

By computing reconstructed counts as close as to the original counts.

Name: _____ Roll: _____

17. In the context of CFG, discriminate between the human and programming languages. [2]

Human CFG can be ambiguous. PL CFG are not ambiguous.

18. Describe the process of tokenization in Spacy. ***No step marking***. [4]

a. Split the sentences into tokens using whitespace chars (space, tab, etc.)

b. From left to right, Does the token require special attention?

i. Yes, Check whether some prefix, suffix, or infix can be split.

ii. No, continue;