IFET COLLEGE OF ENGINEERING

DEPARTMENT OF CSE

23CSX502- Natural Language Processing

UNIT II WORD LEVEL ANALYSIS

QUESTION BANK

N-grams Models of Syntax - Counting Words - Unsmoothed N-grams, Evaluating N-grams, Smoothing, Interpolation and Backoff – Word Classes, Part of Speech Tagging, Rule-based, Stochastic and Transformation-based tagging, Issues in PoS tagging – The Viterbi algorithm - Hidden Markov and Maximum Entropy models.

PART A

Unit 2: N-gram Models of Syntax & Counting Words			
1.	What is an N-gram model in natural language processing?	2	K1
2.	Define unigram, bigram, and trigram models.	2	K1
3.	Why is word counting important in building N-gram models?	2	K2
4.	What is meant by Maximum Likelihood Estimation (MLE) in N-gram modeling?	2	K1
5.	How is the probability of a bigram computed using word counts?	2	K2
6.	Why do unseen N-grams cause a problem in unsmoothed models?	2	K2
Unit 2: Unsmoothed N-grams & Evaluation			
7.	What is an unsmoothed N-gram model**?**	2	K1
8.	What is perplexity in N-gram evaluation?	2	K1
9.	How does perplexity reflect the quality of a language model**?**	2	K2
10.	Name two evaluation metrics used for N-gram models.	2	K1
11.	Why does perplexity increase for higher-order N-grams with sparse data?	2	K2
12.	Why are unsmoothed N-gram models not suitable for real-world applications**?**	2	K2
Unit 2: Smoothing, Interpolation & Backoff			
13.	What is smoothing in N-gram models?	2	K1
14.	Name any two smoothing techniques.	2	K1
15.	How does Laplace (add-one) smooth handle zero probabilities?	2	K2
16.	What is the concept of interpolation in language modeling?	2	K1
17.	How does backoff work when a higher-order N-gram is unavailable?	2	K2
18.	Why is interpolation preferred over simple backoff in practice?	2	K2
Unit 2: Word Classes & Part-of-Speech Tagging			
19.	What are word classes in NLP?	2	K1
20.	Define Part-of-Speech (PoS) tagging**.**		K1
21.	Why is PoS tagging considered a syntactic analysis task**?**	2	K2
22.	How do word-class ambiguities affect PoS tagging accuracy?	2	K2
Unit 2: PoS Tagging Approaches			
23.	What is rule-based PoS tagging**?**	2	K1
24.	How does stochastic PoS tagging differ from rule-based tagging?	2	K2
25.	What is transformation-based tagging**?**	2	K1
26.	Compare rule-based and stochastic tagging in terms of adaptability.	2	K2
Unit 2: Issues in PoS Tagging			
27.	What is lexical ambiguity in PoS tagging?	2	K1
28.	Why is PoS tagging difficult for morphologically rich languages**?**	2	K2
Unit 2: Viterbi Algorithm, HMM & Maximum Entropy			
29.	What is the role of the Viterbi algorithm in PoS tagging?	2	K1
30.	How do Hidden Markov Models (HMMs) differ from Maximum Entropy models in tagging?	2	K2

PART B

1	A speech recognition system uses an unsmoothed trigram language model trained on a limited corpus. Consider the following corpus: <s> I like NLP </s> <s> I like ML </s> <s> I love NLP </s> a) Construct the vocabulary and compute the unigram, bigram, and trigram frequency counts (Counting Words). b) Calculate the unsmoothed unigram, bigram, and trigram probabilities using Maximum Likelihood Estimation (MLE). c) Using the bigram model, compute the probability of the sentence <s> I like NLP </s> (Evaluating N-grams). d) Apply Add-1 (Laplace) smoothing and compute the probability of an unseen bigram such as "love ML".	16	K3
2.	An NLP team develops bigram and trigram models for next-word prediction in a mobile keyboard application. They observe that the trigram model performs worse on new data. a) Explain the role of perplexity in evaluating N-gram models. b) Analyze why a higher-order N-gram may perform poorly with limited data. c) Discuss the impact of data sparsity on model evaluation. d) Recommend strategies to improve the trigram model.	16	K4
3.	A PoS tagging system designed for English shows reduced accuracy when applied to a morphologically rich language. a)Explain the challenges of PoS tagging in morphologically rich languages. b)Discuss the role of word classes in PoS tagging. c) Suggest suitable tagging approaches to handle these challenges.	16	K4
4.	An NLP application requires PoS tagging for real-time text processing. The team must choose between rule-based, stochastic, and transformation-based taggers. a)Explain the working principle of each tagging approach. b)Compare their strengths and limitations. c)Analyze which approach is most suitable for real-time systems. d) Justify your choice with appropriate reasoning.	16	K4
5.	Given the following word counts from a corpus: **Bigram** | **Count** (the, cat) | 40 (the, dog) | 20 the | 100 Vocabulary size = 5	16	K3

a)Compute the unsmoothed bigram probabilities.
b)Apply Laplace smoothing and recompute the probabilities.
c)Compute the probability of the bigram "the mouse" using Laplace smoothing.
e) Compute the perplexity of the bigram model for the sentence "the cat".

6 A trigram language model is trained on a corpus with the following statistics: 16 K3

- Count(the) = 500
- Count(the, students) = 120
- Count(the, students, study) = 60
- Count(students) = 300
- Count(students, study) = 90
- Vocabulary size $V = 8$

Test sentence: **"The students study".**
a)Unsmoothed Trigram Probability using MLE (4 marks)
b)Trigram Probability using Laplace Smoothing (4 Marks)
c)Effect of Data Sparsity in Higher-Order N-gram Models (4 Marks)
d)Role of Interpolation and Backoff in Improving Robustness (4 Marks)

7. Given an HMM with transition and emission probabilities, use the Viterbi algorithm to determine the most likely POS tag sequence for the sentence "fish sleep". Show all intermediate calculations. 16 K2

Given
Sentence: **fish sleep**
Tags: N (Noun), V (Verb)

Transition probabilities	**Emission probabilities**
• Start→N = 0.6 • Start→V = 0.4 • N→N = 0.3 • N→V = 0.7 • V→N = 0.8 • V→V = 0.2	• P(fish|N) = 0.4 • P(fish|V) = 0.6 • P(sleep|N) = 0.5 • P(sleep|V) = 0.5

a) Compute the probability of the tag sequence N → V for the sentence using the transition and emission probabilities.
b) Compute the probability of the tag sequence V → N for the same sentence.
c) Using the Viterbi algorithm, find the most likely sequence of POS tags for the sentence "fish sleep". Show all intermediate calculations.
d) Identify the final best tag sequence obtained using the Viterbi method.
e) Construct the Viterbi trellis (table) showing probabilities at each step.
f) State the maximum probability value corresponding to the best path.

8 (i) Describe the role of transition probabilities, emission probabilities, and the Viterbi algorithm. 8 K2

(ii) Explain the principle of Maximum Entropy models for PoS tagging. 8 K2