[Arabic text]

Question 1 (5 Marks)

B is a corpus which only contains one single bit string:

1 1 1 1 1 0 0 0 0 1 0 1 1 1 0 1 0 1 1 0 1 1 0

1.1) Calculate the following bigram probabilities from the corpus B using MLE (Maximum Likelihood Estimation). Answer with a ratio p/q, not a floating point number.

(a) $P(0 \mid 1)$ 1 mark

$P(0 \mid 1) = 6/15$

(b) $P(0 \mid 0)$ 1 mark

$P(0 \mid 0) = 3/9$

(Left page — handwritten)

b) Convert the grammar into Chomsky Normal Form. [2.5 marks]

$S \to E\ F$
$E \to \emptyset\ N$
$E \to P$
$F \to V\ E\ PP$
$PP \to D\ E$
$O \to From$ $N \to cheese$
$D \to The$ $P \to Tom$
$V \to steals$ $P \to Jerry$

c) Following the CKY algorithm, parse the following sentence with the above lexicon/grammar. "Tom steals the cheese from Jerry". [3 marks]

Tom	Steals	* The	Cheese	From	Jerry
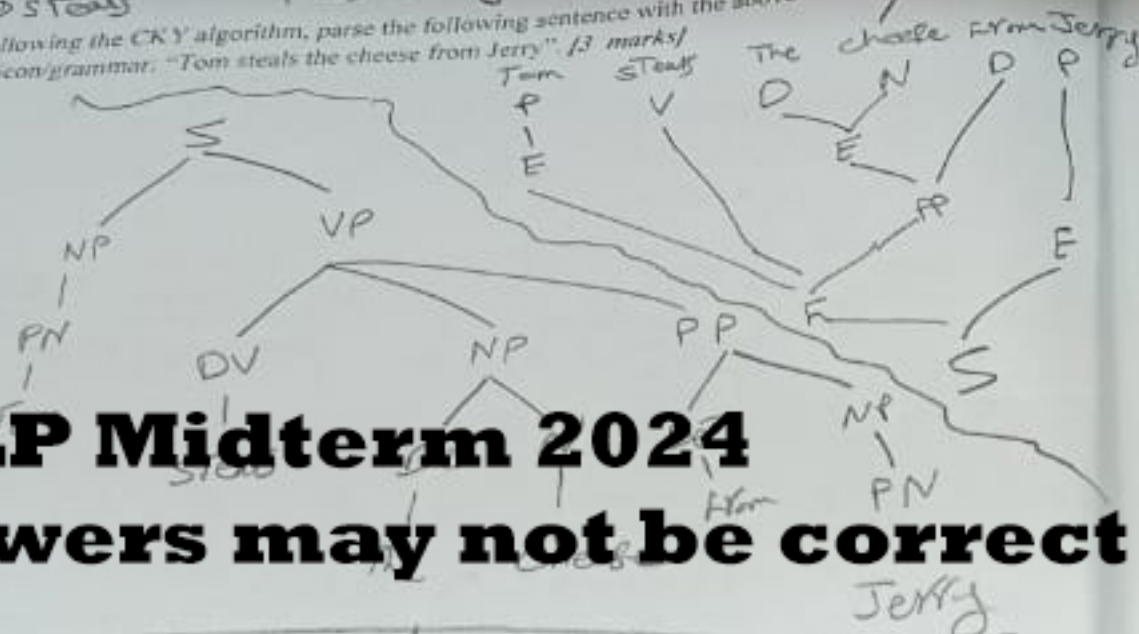
NLP Midterm 2024
These answers may not be correct

1.2) Assume a bigram language model created from corpus B. For each of the following bit strings, decide if it is more probable that x_1 resp x_2 is 0 or 1.

(c) $1010101x_1$ *1.5 marks*

$X_1 = 1$ is more probable

$P(0|1) = 6/15 < P(1|1) = 8/15$

(d) $0101010x_2$ *1.5 marks*

$x_2 = 1$ is more probable

$P(0|0) = 3/9 < P(1|0) = 5/9$

Question 2 (8 Marks)

1- (True or False. Explain your answer.) 4-grams are better than trigrams for part-of-speech tagging. (2 Marks)

false, there is no enough data for u-grams to out perform trigrams

2- In the pair (played, play+V+Past), what does "played" (resp. "play+V+ Past") correspond to? What is each of the two forms useful for? (3 Marks)

serface level is useful for interface
lexical level is useful for internal representation

3- Suppose you are running a shift-reduce dependency parser on a sentence of length n. How many shift operations are needed in terms of sentence length? [1 mark]

$n+1$

4- A sentence can easily have more than one parse tree that is consistent with a given CFG. How do PCFGs and non-probability-based CFGs differ in terms of handling parsing ambiguity? [2 marks]

PCFGs compute probability for every tree and tree with high probability is more efficient.

Question 3 (7 Marks)

Complete the context-free grammar below so that it generates (at least) the sentences listed.

GRAMMAR
S -> NP VP
NP -> Det N
NP -> PN
VP -> TV NP
VP -> DV NP PP
Det -> the
Det -> a
N -> mouse
N -> cat
PN -> Tom
PN -> Jerry
N -> cheese
IV -> sleeps
TV -> eats
DV -> steals

SENTENCES
The cat sleeps.
The mouse eats the cheese.
Tom steals the cheese from the mouse.

a) $PP \rightarrow$ Det NP
$D\bar{e}t \rightarrow$ from

[1.5 marks]