CSE556 NLP

Quiz 2 Rubrics

Date: 14 Oct, 2022 **Max Marks: 20**

1. Mention and define (1 line per type) the type of evaluations. [3]
   a. Human evaluation: Manually evaluate how likely the generated sentence is.
   b. Extrinsic evaluation: Utilize the generated sentence for building another system (e.g., MT)
   c. Intrinsic evaluation: Compute some evaluation metric.

   0.5 for mentioning and 0.5 for defining each type.

2. Utilizing BIO (Begin, Intermediate, and Outside) encoding, tag names in the following sentence. [3]
   Unarguably_O ,_O Federer_B ,_O Nadal_B ,_O and_O Djokovich_B are_O the_O best_O tennis_O players_O ever_O and_O they_O are_O leading_O the_O grand_B slam_I trophies_I in_O Wimbledon_B (_O 8_O)_O ,_O French_B Open_I (_O 14_O)_O ,_O and_O Australian_B Open_I (_O 9_O)_O ,_O respectively_O ._O.

   No partial marking

3. Differentiate between type and token. How many types and tokens are there in the Q2 sentence. [2]
   a. Tokens: Tokens are the total words present in your text; It is an individual occurrence of a linguistic unit. [0.5]
   b. Type: Type is just the unique words that are elements of the vocabulary. [0.5]

      Tokens : 43 [0.5]
      Types : 29 [0.5]

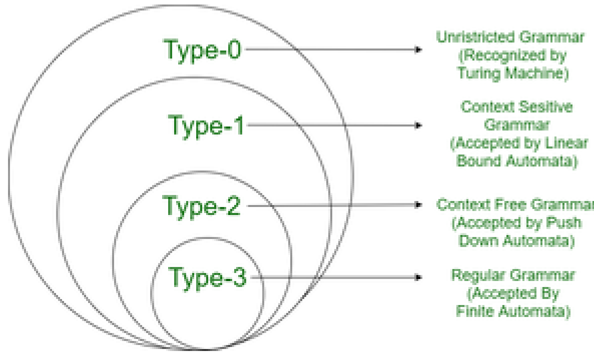
   Note: Definition may vary but must be correct.

4. According to the chomsky hierarchy, the most appropriate class of the following grammar is: [1]

   G: $A \rightarrow Aa \mid b$

   Regular Grammar or Type 3

   Note: All RG are CFG but not all CFG are RG.



5. Convert the grammar in Q4 into chomsky normal form. [2]

$A \rightarrow AB$ [1]
$A \rightarrow b$ [0.5]
$B \rightarrow a$ [0.5]

In CNF, need to follow two rules:

- A non-terminal generating two non-terminals. ($N \rightarrow NN$)
- A non-terminal generating terminal. ($N \rightarrow T$)

If these rules are not followed, then 0 will be awarded.

6. Is the following grammar suitable for top-down parsing? If not, suggest a solution. [3]

G : $S \rightarrow Aa \mid b$
$A \rightarrow Sc \mid d$

No, this is a case of non-immediate or indirect left recursive. [1]

Solution: Elimination of Left-Recursion

a. Substitute A by its production rules, to convert into direct left recursion. [1]

$S \rightarrow Sca \mid da \mid b$
$A \rightarrow Sc \mid d$

b. Convert into the CNF, using the given rule. [1]

$A \rightarrow A\alpha \mid \beta$ Can be replace by non-left-recursive productions $A \rightarrow \beta A'$ $A' \rightarrow \alpha A' \mid \varepsilon$

$S \rightarrow bS` \mid daS`$
$S` \rightarrow caS` \mid \varepsilon$
$A \rightarrow Sc \mid d$

Other valid production rules are also possible.

7. How do we evaluate a smoothing technique? Discuss with an example (assume your vocabulary) of bigram probability using Laplace smoothing.

We assume, $W = \{w_1, w_2 --- w_{i-1}, w_i\}$

U_c = Unigram count of w_i = C_{w_i}

B_c = Bigram count of (w_{i-1}, w_i) = C_{w_{i-1}, w_i}

B_p = Bigram Prob = $C_{w_{i-1}, w_i} / C_{w_{i-1}}$

B_{LP} = Laplacian Bigram Prob = $(C_{w_{i-1}, w_i} + 1) / (C_{w_{i-1}} + |V|)$

Smoothed Count = $B_{LP} \times U_c$

→ $|V|$ denotes vocabulary size

Let us assume our corpus = {A A A B B B C C A} ; $|V| = 3$

	A	B	C
A	2	1	0
B	0	2	1
C	1	0	1

Bigram Count
M1

⇒

	A	B	C
A	3/7	2/7	1/7
B	1/6	3/6	2/6
C	2/5	1/5	2/5

Laplace Bigram Prob
M2

⇒

	A*	B*	C*
A*	3/7×4	2/7×4	1/7×4
B*	1/6×3	3/6×3	2/6×3
C*	2/5×2	1/5×2	2/5×2

Smoothed Bigram Count
M3

Third matrix represents smoothed bigram counts (redistribution of counts)

Any smoothing function is good if M1 ≈ M3.

1. Initial Assumptions and upto bigram prob (displaying initial bigram prob). [1] mark
2. Providing the correct Laplacian smoothing prob formula. [1] mark
3. Providing the correct reconstructed bigram count formula (Eq 7). [1] mark
4. Applying the laplacian smoothing prob formula correctly to display the smoothed bigram prob matrix (2nd matrix). [1.5] marks
5. Correctly applying the smoothed bigram count matrix (3rd matrix). [1.5] marks

Extra Material: Complete details of the formulation, redistribution count effect and intuitive derivation of smoothed count is given below.

Q 7:

Input: Dataset D

Expected Output: Smoothed/reconstructed bigram counts B_i^*

Process:

Assume D: AAA BBBCCA — (1.)

Then,

Vocabulary V: {A, B, C}

$\boxed{|V| = 3}$ — (2.)

Unigram Count: C_w (w.r.t bigram counts this can also be written as $C_{w_{i-1}}$)

A	B	C
4	3	2

— (3.)

Bigram Count (B_i): $\boxed{C_{w_{i-1}, w_i}}$ — (4.)

$B_i =$

	A	B	C
A	2	1	0
B	0	2	1
C	1	0	1

From (3.) & (4.) we obtain the bigram probability (P_{bi}) as:

$$\boxed{P_{bi} = \frac{C_{w_{i-1}, w_i}}{C_{w_{i-1}}}} \quad \text{— (5.)}$$

Now, applying the laplacian smoothing on (5) we obtain the laplacian Bigram $\boxed{\text{Prob } (P_{bi\text{-}lap}) = \frac{C_{w_{i-1}, w_i} + 1}{C_{w_{i-1}} + V}}$ — (6.)

	A	B	C
A	(2+1)/(4+3)	(1+1)/(4+3)	(0+1)/(4+3)
B	(0+1)/(3+3)	(2+1)/(3+3)	(1+1)/(3+3)
C	(1+1)/(2+3)	(0+1)/(2+3)	(1+1) (2+3)

$\Rightarrow$

	A	B	C
A	3/7	2/7	1/7
B	1/6	3/6	2/6
C	2/5	1/5	2/5

Q7:

Finally we obtain the smoothed bigram counts as:

$$B_i^* = C^*_{w_{i-1}, w_i} = \underbrace{\frac{[C_{w_{i-1},w_i} + 1]}{[C_{w_{i-1}} + V]}}_{A} * \underbrace{C_{w_{i-1}}}_{B} \quad \text{— (7.)}$$

Here the part (A) is obtained from eq. (6.) & (B.) is obtained from eq. (3.)

Thus,

$B_i^* =$

	A^*	B^*	C^*
A^*	$(3/7)*4$	$(2/7)*4$	$(1/7)*4$
B^*	$(1/6)*3$	$(3/6)*3$	$(2/6)*3$
C^*	$(2/5)*2$	$(1/5)*2$	$(2/5)*2$

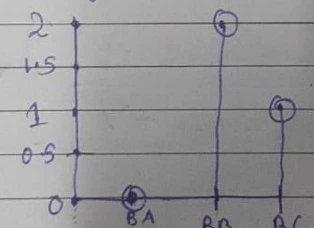
$\Rightarrow$

	A^*	B^*	C^*
A^*	1.71	1.14	0.57
B^*	0.5	1.5	1
C^*	0.8	0.4	0.8

B_i^* represents the redistribution of counts post smoothing.

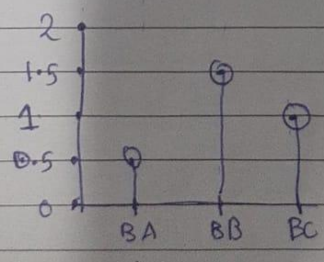
(Bonus): To bring things into prespective we display the shift in histogram counts w.r.t $C_{w_{i-1}}$ as B.
[NOT MARKED]

B_i for B



(ref. Eq (4.))

B_i^* for B



(ref. Eq (7))

For intution we also show the derivation of B_i^* [NOT MARKED]

Q7: (3.)

Intutive derivation of smoothed / reconstructed bigram counts c^*

$$\text{Bigram Prob} = \frac{\text{Bi\_count}[C_{wi-1, wi}]}{\text{Uni\_count}[C_{wi-1}]}$$

$$P_{bi} = \frac{C_{wi-1, wi}}{C_{wi-1}} \quad - (1.)$$

$$\text{or } C_{wi-1, wi} = P_{bi} * C_{wi-1} \quad - (2.)$$

Bigram Prob based on laplacian smoothing

$$P_{bi-lap} = \frac{C_{wi-1, wi} + 1}{C_{wi-1} + V} \quad - (3.)$$

Replacing P_{bi} in (2.) with the updated P_{bi-lap} we get

$$\boxed{C^*_{wi-1, wi} = \left[\frac{C_{wi-1, wi} + 1}{C_{wi-1} + V}\right] * C_{wi-1}} \quad - (4.)$$