INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
Mid Semester Examination 2022-23

Date of Examination: /09/23 **Session(FN/AN)** __ **Duration** 3 hrs, **Marks** = 50
Sub No: CS60075 **Sub Name:** Natural Language Processing
Department/Centre/School : Computer Science and Engineering
Specific charts, graph paper,log book etc. required NO
Special Instructions (if any) ANSWER ALL questions. In case of reasonable doubt, make assumptions and state them upfront. Marks will be deducted for sketchy answers and claims without proper reasoning. All parts of a single question should be done at the same place. Keep probability values in fractional forms.

---

1. The following sentence, though bizarre and deliberately confusing, is grammatically correct!

   `The weasel that a boy that startles the cat thinks loves smiles eats.`

   Answer the following questions by identifying the correct parse (Hint: you might use square brackets as we used in class to represent the parse/breakup of the sentence). In some cases, the answers may be 'nobody/nothing in this sentence'.

   (a) What is the subject of this sentence? (Give a single-word answer.)
   (b) How many verbs are in the sentence?
   (c) Who startles whom or what?
   (d) Who thinks what?
   (e) Who loves whom or what?
   (f) Who smiles?
   (g) Who eats whom or what?

   [7]

   ```
   'Weasel' is the subject.
   Four: 'startles', 'thinks', 'loves' and 'eats'. 'Smiles' is a noun.
   A boy startles the cat.
   A boy thinks that a weasel loves smiles.
   A weasel loves smiles (at least in the mind of the boy).
   Nobody in this sentence (explicitly) smiles; 'smiles' is used as a noun.
   A weasel eats something unspecified
   ```

2. Consider the following toy example:
   Training data:
   `<s> I am Sam </s>`
   `<s> Sam I am </s>`
   `<s> Sam I like </s>`
   `<s> Sam I do like </s>`
   `<s> do I like Sam </s>`
   Assume that we use a bigram language model (no smoothing) based on the above training data. Compute the bigram probabilities.

   (a) What is the most probable next word predicted by the model for the following word sequences?

      i. `<s> Sam . . .`

ii. `<s> Sam I do . . .`

iii. `<s> Sam I am Sam . . .`

iv. `<s> do I like . . .`

(b) Which of the following sentences is better, i.e., gets a higher probability with this model? Compute the probabilities in each case and hence find the best one.

i. `<s> Sam I do I like </s>`

ii. `<s> Sam I am </s>`

iii. `<s> I do like Sam I am </s>`

(c) Consider again the same training data and the same bigram model. Compute the perplexity of `<s> I do like Sam`

(d) Take again the same training data. This time, we use a bigram language model with Laplace smoothing.

i. Give the following bigram probabilities estimated by this model: $P(\texttt{do|<s>})$, $P(\texttt{do|Sam})$, $P(\texttt{Sam|<s>})$, $P(\texttt{Sam|do})$, $P(\texttt{I|Sam})$, $P(\texttt{I|do})$, $P(\texttt{like|I})$

Note that for each word w_{n-1}, we count an additional bigram for each possible continuation w_n. Consequently, we have to take the words into consideration as well as the symbol `</s>`.

ii. Calculate the probabilities of the following sequences according to this model. Which of the two sequences is more probable according to our language model?

A. `<s> do Sam I like`

B. `<s> Sam do I like`

Solution:

Bigram probabilities:

$P(\texttt{Sam}|\texttt{<s>}) = \frac{3}{5}$ $P(\texttt{I}|\texttt{<s>}) = \frac{1}{5}$

$P(\texttt{I}|\texttt{Sam}) = \frac{3}{5}$ $P(\texttt{</s>}|\texttt{Sam}) = \frac{2}{5}$

$P(\texttt{Sam}|\texttt{am}) = \frac{1}{2}$ $P(\texttt{</s>}|\texttt{am}) = \frac{1}{2}$

$P(\texttt{am}|\texttt{I}) = \frac{2}{5}$ $P(\texttt{like}|\texttt{I}) = \frac{2}{5}$ $P(\texttt{do}|\texttt{I}) = \frac{1}{5}$

$P(\texttt{Sam}|\texttt{like}) = \frac{1}{3}$ $P(\texttt{</s>}|\texttt{like}) = \frac{2}{3}$

$P(\texttt{like}|\texttt{do}) = \frac{1}{2}$ $P(\texttt{I}|\texttt{do}) = \frac{1}{2}$

1. (1) and (3): "I".

   (2): "I" and "like" are equally probable.

   (4): `</s>`

2. Probabilities:

   (5): $\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{2}{3}$

   (6): $\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{2}$

   (7): $\frac{1}{5} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{2}$

   (6) is the most probable sentence according to our language model.

Exercise 2 *Consider again the same training data and the same bigram model. Compute the perplexity of*

<s> *I do like Sam*

Solution:
The probability of this sequence is $\frac{1}{5} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{150}$.
The perplexity is then $\sqrt[4]{150} = 3.5$

Exercise 3 *Take again the same training data. This time, we use a bigram LM with Laplace smoothing.*

1. *Give the following bigram probabilities estimated by this model:*
   $P(\texttt{do}|\texttt{<s>})$ $P(\texttt{do}|\texttt{Sam})$ $P(\texttt{Sam}|\texttt{<s>})$ $P(\texttt{Sam}|\texttt{do})$
   $P(\texttt{I}|\texttt{Sam})$ $P(\texttt{I}|\texttt{do})$ $P(\texttt{like}|\texttt{I})$

   Note that for each word w_{n-1}, we count an additional bigram for each possible continuation w_n. Consequently, we have to take the words into consideration and also the symbol </s>.

2. *Calculate the probabilities of the following sequences according to this model:*

   (8) <s> *do Sam I like*

   (9) <s> *Sam do I like*

   Which of the two sequences is more probable according to our LM?

Solution:

1. If we include </s> (this can also appear as second element of a bigram), we get $|V| = 6$ for our vocabulary.
   $P(\texttt{do}|\texttt{<s>}) = \frac{2}{11}$ $P(\texttt{do}|\texttt{Sam}) = \frac{1}{11}$ $P(\texttt{Sam}|\texttt{<s>}) = \frac{4}{11}$ $P(\texttt{Sam}|\texttt{do}) = \frac{1}{8}$
   $P(\texttt{I}|\texttt{Sam}) = \frac{4}{11}$ $P(\texttt{I}|\texttt{do}) = \frac{2}{8}$ $P(\texttt{like}|\texttt{I}) = \frac{3}{11}$
2. (8): $\frac{2}{11} \cdot \frac{1}{8} \cdot \frac{4}{11} \cdot \frac{3}{11}$
   (9): $\frac{4}{11} \cdot \frac{1}{11} \cdot \frac{2}{8} \cdot \frac{3}{11}$
   The two sequences are equally probable.

[3 + 2 + 4.5 + 1.5 + 3 + 4 = 18]

3. What is the tagging of the following sentence

computers process programs accurately

with the following HMM tagger (N is noun, V is verb, Adv is adverb):
(part of) lexicon (assume that the following words have no other POS apart from those given below):
computers N 0.123
process N 0.1
process V 0.2
programs N 0.11
programs V 0.15
accurately Adv 0.789
(part of) transitions:
$P(N|V) = 0.5, P(N|Adv) = 0.12, P(V|Adv) = 0.05, P(V|N) = 0.4, P(Adv|N) = 0.01, P(Adv|V) = 0.13, P(N|N) = 0.6, P(V|V) = 0.05$
Show the computations. **[10]**

4 choices (it's a lattice):

```
computers process programs accurately
    N        N        N        Adv
             V        V
```

Differences are (skept the common factors):

```
P(N|N) P(process|N) P(N|N) P(programs|N) P(Adv|N)
P(N|N) P(process|N) P(V|N) P(programs|V) P(Adv|V)
P(V|N) P(process|V) P(N|V) P(programs|N) P(Adv|N)
P(V|N) P(process|V) P(V|V) P(programs|V) P(Adv|V)
```

i.e.:

```
    0.6    0.1    0.6    0.11   0.01
--> 0.6    0.1    0.4    0.15   0.13 <--MAX
    0.4    0.2    0.5    0.11   0.01
    0.4    0.2    0.05   0.15   0.13
```

4. Assume that the texts to be POS tagged contain 1.5% of unknown words and that the performance of the POS tagger to be used is 98% on known words. What will be its typical overall performance if all unknown words are systematically wrongly tagged? **[3]**

1.5% is for sure wrongly tagged. For the rest (100%-1.5%), only 98% are correctly tagged. So the overall score is $0.985 \times 0.98 \simeq 0.96$.

(b) this is less obvious: still we have 0.985×0.98, but on the remaining 1.5% we cannot be sure:

5. In the dependency graph shown in Figure 1, find out the labels that are incorrect on the arcs. Correct them. `avmod` means adverbial modifier and `amod` means adjectival modifier. **[3]**

```
    'athlete' to 'agile' --> 'advmod' is wrong.
    It should be 'amod' because 'advmod' is adverbial
    modifier and 'amod' is adjectival modifier.

'athlete' is running and is the subject. Hence, the label
from 'runs' --> 'athlete' should be 'nsubj'. 'nobj' means
nominal object.

'runs' --> 'fast' should be 'advmod' and NOT 'amod'.
```

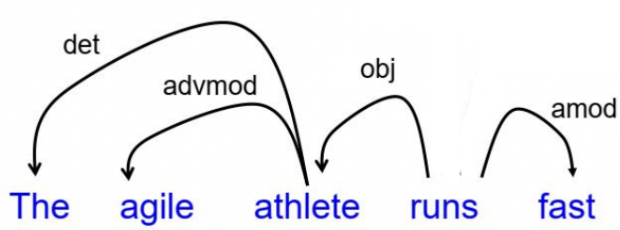

Figure 1: Dependency tree.

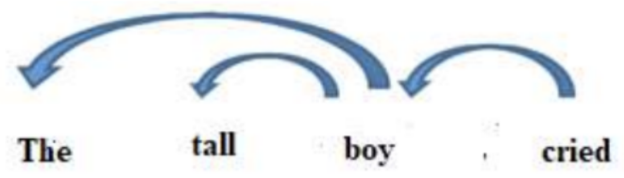

Figure 2: Reference sentence.

6. Give the correct sequence of arc eager parsing operations for the given sentence in Figure 2. Show the (a) action taken, (b) stack state, (c) buffer state. Provide a modified transition sequence where the parser mistakenly predicts the arc boy → cried, but gets the other dependencies right. **[2.5 + 2.5 = 5]**

[ ]	[Tha tall boy cried]	[]
[The]	[tall boy cried]	[Shift]
[The , tall]	[boy cried]	[Shift]
[The , tall]	[boy cried]	[LA]
[The]	[boy cried]	[LA]
[boy]	[cried]	[SH]
[ ]	[cried]	[LA]
[cried]	[]	[RA]

[ ]	[The tall boy cried]	[]
[The]	[tall boy cried]	[Shift]
[the , tall]	[boy cried]	[Shift]
[the , tall]	[boy cried]	[LA]
[the]	[boy cried]	[LA]
[, boy]	[cried]	[SH]
[boy]	[]	[RA]
[, boy cried]	[]	[RE]
[boy]	[]	[RE]

7. Suppose in a recommender system for online shopping, we have information about co-purchase records for items $x_1, x_2, ..., x_n$ (for example, item x_i is commonly bought together with item x_j). Explain how you would use ideas of distributional similarity to (i) define a context (ii) obtain embeddings (how and for what?) and (iii) recommend similar items to users who have shown interest in any one of the items. **[1 + 2 + 1 = 4]**

`We can treat items that are copurchased with x to be in` the 'con-text' of item x (1 point). We can use those copurchase records to build item embeddings akin to PMI/Word2Vec. (2 points, you need to mention that item embeddings are created). Then we can use a similarity metric such as finding the items with the largest cosine similarity to the average basket to determine item recommendations for users (1 point)