



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]

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>

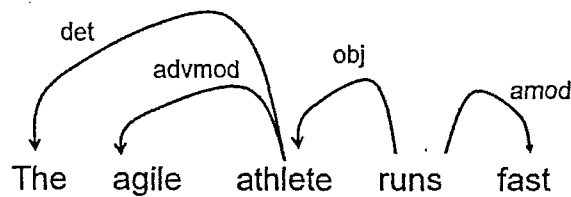


Figure 1: Dependency tree.

- 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.
- Give the following bigram probabilities estimated by this model: $P(\text{do}|\text{<s>})$, $P(\text{do}|\text{Sam})$, $P(\text{Sam}|\text{<s>})$, $P(\text{Sam}|\text{do})$, $P(\text{I}|\text{Sam})$, $P(\text{I}|\text{do})$, $P(\text{like}|\text{I})$
 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>`.
 - Calculate the probabilities of the following sequences according to this model. Which of the two sequences is more probable according to our language model?
 - `<s> do Sam I like`
 - `<s> Sam do I like`

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

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

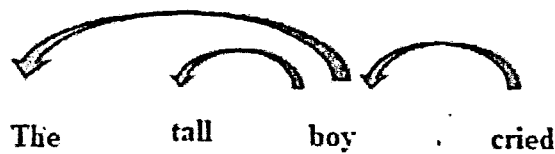


Figure 2: Reference sentence.

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 $\text{boy} \rightarrow \text{cried}$, but gets the other dependencies right.

[2.5 + 2.5 = 5]

7. Suppose in a recommender system for online shopping, we have information about co-purchase records for items x_1, x_2, \dots, 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]