

# COMP 6721 - Artificial Intelligence

## n-gram Modeling

**Question 1** Assume that we are working with the Shloutan language. If you don't know Shloutan, don't worry; it is a simple language made of only 5 words: loola nikee aloka bibi vo.

You want to build a word language model for Shloutan. The training corpus that you use is the following:

“Loola nikee. Aloka bibi vo. Vo bibi loola. Loola nikee bibi vo. Vo. Vo. Aloka bibi loola. Loola aloka aloka. Loola loola. Nikee nikee nikee. Bibi vo. Bibi vo. Vo Vo. Nikee loola.”

You can ignore case distinctions and sentence boundaries when answering the following questions.

- (a) What is the value of  $P(\text{vo} \mid \text{bibi})$ ?  $=c(\text{bibi vo})/c(\text{bibi})=4/6$
- (b) What is the value of  $P(\text{bibi vo})$ ?  $=c(\text{bibi vo})/Zc(\text{wiwj})=4/32$ . (wiwj为corpus中成对单词的数量)
- (c) Build a bigram language model based on this training corpus. Show the frequencies and the probabilities for each bigram.

### Frequencies

此排为给定条件

|       | loola | nikee | aloka | bibi | vo |
|-------|-------|-------|-------|------|----|
| loola | 3     | 3     | 1     | 0    | 0  |
| nikee | 1     | 2     | 1     | 2    | 0  |
| aloka | 1     | 0     | 1     | 2    | 0  |
| bibi  | 2     | 0     | 0     | 0    | 4  |
| vo    | 0     | 1     | 1     | 2    | 5  |

$$p(\text{wi}|\text{wj})=c(\text{wi-1wi})/c(\text{wi-1})$$

### Conditional Probabilities

|       | loola | nikee | aloka | bibi | vo |
|-------|-------|-------|-------|------|----|
| loola | 0.43  | 0.43  | 0.14  |      |    |
| nikee | 0.17  |       |       |      |    |
| aloka |       |       |       |      |    |
| bibi  |       |       |       |      |    |
| vo    |       |       |       |      |    |

- (d) Smooth your bigram language model using “add 0.5”. Show the frequencies and the probabilities for each bigram.

### Frequencies

|       | loola | nikee | aloka | bibi | vo  |
|-------|-------|-------|-------|------|-----|
| loola | 3.5   | 3.5   | 1.5   | 0.5  | 0.5 |
| nikee |       |       |       |      |     |
| aloka |       |       |       |      |     |
| bibi  |       |       |       |      |     |
| vo    |       |       |       |      |     |

### Conditional Probabilities

|       | loola | nikee | aloka | bibi | vo |
|-------|-------|-------|-------|------|----|
| loola |       |       |       |      |    |
| nikee |       |       |       |      |    |
| aloka |       |       |       |      |    |
| bibi  |       |       |       |      |    |
| vo    |       |       |       |      |    |

- (e) Using each language model from parts (c) and (d), which of the following 2 sentences is more probable. Show all your work.

$$p(s1)=p(vo \mid aloka)*p(nikee \mid vo)*p(aloka \mid nikee)=0$$

sentence 1: Aloka vo nikee aloka.

sentence 2: Vo nikee nikee aloka.