N-gram Modeling

COMP 472/6721: Artificial Intelligence

Mohammad Reza Davari

Concordia University

Outline

Contents of the section

Problem Statement

- **Question:** 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:

Problem Statement

- **Question:** 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:
- **Training Corpus:** "Loola nikee. Aloka bibi vo. Vo bibi loola. Loola nikee bibi vo. Vo. Vo. Alokabibi loola. Loola aloka aloka. Loola loola. Nikee nikee nikee. Bibi vo. Bibi vo.Vo Vo. Nikee loola."

We will ignore case distinctions and sentence boundaries from here on.

Contents of the section

Basic Probability Calculations

Recap

- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Basic Probability Calculations

Recap

- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

Basic Probability Calculations

Recap

- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$
- $P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n)$

Basic Probability Calculations

Find the value of the followings:

- $P(\text{vo}|\text{bibi})$
- $P(\text{bibi vo})$

$P(\text{vo}|\text{bibi})$

$$P(\text{vo}|bibi) = \frac{P(\text{bibi} \cap \text{vo})}{P(\text{bibi})} \quad (1)$$

$P(\text{vo}|\text{bibi})$

$$P(\text{vo}|bibi) = \frac{P(\text{bibi} \cap \text{vo})}{P(\text{bibi})} \quad (1)$$

$$= \frac{P(\text{bibi} \cap \text{vo})}{\sum_n P(\text{bibi} \cap W_n)} \quad (2)$$

$P(\text{vo}|\text{bibi})$

$$P(\text{vo}|bibi) = \frac{P(\text{bibi} \cap \text{vo})}{P(\text{bibi})} \tag{1}$$

$$= \frac{P(\text{bibi} \cap \text{vo})}{\sum_n P(\text{bibi} \cap W_n)} \tag{2}$$

$$= \frac{\frac{\#(\text{bibi vo})}{\#(\text{bigrams})}}{\sum_n \frac{\#(\text{bibi } W_n)}{\#(\text{bigrams})}} \tag{3}$$

$P(\text{vo}|\text{bibi})$

$$P(\text{vo}|bibi) = \frac{P(\text{bibi} \cap \text{vo})}{P(\text{bibi})} \tag{1}$$

$$= \frac{P(\text{bibi} \cap \text{vo})}{\sum_n P(\text{bibi} \cap W_n)} \tag{2}$$

$$= \frac{\frac{\#(\text{bibi vo})}{\#(\text{bigrams})}}{\sum_n \frac{\#(\text{bibi } W_n)}{\#(\text{bigrams})}} \tag{3}$$

$$= \frac{\#(\text{bibi vo})}{\sum_n \#(\text{bibi } W_n)} \tag{4}$$

$P(\text{vo}|\text{bibi})$

$$
\begin{aligned}
P(\text{vo}|bibi) &= \frac{P(\text{bibi} \cap \text{vo})}{P(\text{bibi})} && (1)\\
&= \frac{P(\text{bibi} \cap \text{vo})}{\sum_n P(\text{bibi} \cap W_n)} && (2)\\
&= \frac{\frac{\#(\text{bibi vo})}{\#(\text{bigrams})}}{\sum_n \frac{\#(\text{bibi } W_n)}{\#(\text{bigrams})}} && (3)\\
&= \frac{\#(\text{bibi vo})}{\sum_n \#(\text{bibi } W_n)} && (4)\\
&= \frac{4}{6} && (5)
\end{aligned}
$$

$P(\text{bibi vo})$

$$P(\text{bibi vo}) \quad = \quad P(\text{bibi} \cap \text{vo}) \tag{1}$$

$P(\text{bibi vo})$

$$
\begin{aligned}
P(\text{bibi vo}) &= P(\text{bibi} \cap \text{vo}) && (1)\\
&= \frac{\#(\text{bibi vo})}{\#(\text{bigrams})} && (2)
\end{aligned}
$$

$P(\text{bibi vo})$

$$
\begin{aligned}
P(\text{bibi vo}) &= P(\text{bibi} \cap \text{vo}) && (1)\\
&= \frac{\#(\text{bibi vo})}{\#(\text{bigrams})} && (2)\\
&= \frac{4}{32} && (3)
\end{aligned}
$$

Contents of the section

Bigram Language Model: Without Smoothing

Frequencies

	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

Bigram Language Model: Without Smoothing

Probabilities

	loola	nikee	aloka	bibi	vo
loola	0.43	0.43	0.14	0	0
nikee	0.17	0.33	0.17	0.33	0
aloka	0.25	0	0.25	0.50	0
bibi	0.33	0	0	0	0.67
vo	0	0.11	0.11	0.22	0.56

Bigram Language Model: With Smoothing (Add 0.5)

Frequencies

	loola	nikee	aloka	bibi	vo
loola	3.5	3.5	1.5	0.5	0.5
nikee	1.5	2.5	1.5	2.5	0.5
aloka	1.5	0.5	1.5	2.5	0.5
bibi	2.5	0.5	0.5	0.5	4.5
vo	0.5	1.5	1.5	2.5	5.5

Bigram Language Model: With Smoothing (Add 0.5)

Probabilities

	loola	nikee	aloka	bibi	vo
loola	0.37	0.37	0.16	0.05	0.05
nikee	0.18	0.29	0.18	0.29	0.06
aloka	0.23	0.08	0.23	0.38	0.08
bibi	0.29	0.06	0.06	0.06	0.53
vo	0.04	0.13	0.13	0.22	0.48

Contents of the section

Sentence Probability Calculations

Using each bigram models (with and without smoothing) which of the following 2 sentences is more probable.

1. Aloka vo nikee aloka.
2. Vo nikee nikee aloka.

Sentence 1

Bigrams

1. Aloka vo
2. vo nikee
3. nikee aloka

Sentence 1

Bigrams

1. Aloka vo
2. vo nikee
3. nikee aloka

Using Bigram Model w/o Smoothing

$P(\text{Sentence 1}) = 0 \times 0.11 \times 0.17 = 0$

Sentence 1

Bigrams

1. Aloka vo
2. vo nikee
3. nikee aloka

Using Bigram Model w/o Smoothing

$P(\text{Sentence 1}) = 0 \times 0.11 \times 0.17 = 0$

Using Bigram Model w/ Smoothing

$P(\text{Sentence 1}) = 0.08 \times 0.13 \times 0.18 \approx 0.0019$

Sentence 2

Bigrams

1. Vo nikee
2. nikee nikee
3. nikee aloka

Sentence 2

Bigrams

1. Vo nikee
2. nikee nikee
3. nikee aloka

Using Bigram Model w/o Smoothing

$P(\text{Sentence 2}) = 0.11 \times 0.33 \times 0.17 \approx 0.0062$

Sentence 2

Bigrams

1. Vo nikee
2. nikee nikee
3. nikee aloka

Using Bigram Model w/o Smoothing

$P(\text{Sentence 2}) = 0.11 \times 0.33 \times 0.17 \approx 0.0062$

Using Bigram Model w/ Smoothing

$P(\text{Sentence 2}) = 0.13 \times 0.29 \times 0.18 \approx 0.0068$

Sentence Probability Calculations

More Probable Sentence

Sentence 2 is more probable using either model.