

Module-I (Part-2): Language Modeling

by:

Dr. Soumya Priyadarsini Panda

Sr. Assistant Professor

Dept. of CSE, SIT, Bhubaneswar

Contents

- Language Modeling
- Various types of languages and their modelling
- Grammar-based language modeling
- Statistical language modeling
- N-gram model

Language Modeling

- **Language:** Primary mean of communication used by human
- **Grammar:** Set of rules to generate sentences in a language
 - Provides the mean to specify natural language
- Automatic processing of language requires rules and exceptions of a language to be explained to the computer
- **Language Modeling:** Representation or model to process natural language through a computer based program

Types of Languages and their Modeling

- There are various types of languages used world wide.
- Each language has its own writing **script** covering a large collection of characters and symbols.

Examples:

Language:

Hindi:

Oriya (Odia)

Bengali

Script:

Devnagari

Odia

Bengali/ Bangla

Cont...

The Unicode standard:

- Unicode is a standard for consistent encoding representation and handling of text expressed in different writing systems.
- It reserves a range of values for different language alphabets and symbols to be used world wide.

Examples:

Language

Range

Devnagari(Hindi)

0900-097F

Bengali

0980-09FF

Oriya (Odia)

0B00-0B7F

Language and Grammar

- A language can be generated given its grammar

$G = (V, \Sigma, S, P)$, where

V = set of variables,

Σ = set of terminal symbols,

(which appear at the end of generation)

S = start symbol,

P = set of production rules

- The corresponding language of G is $L(G)$

Example-1

Tuples :

$V = \{S, NP, N, VP, V, Art\}$

$\Sigma = \{\text{boy, ice-cream, dog, bite, like, ate, the, a}\}$

Production Rules:

$P = \{S \rightarrow NP V P,$

$NP \rightarrow N,$

$NP \rightarrow ART N,$

$V P \rightarrow V NP,$

$N \rightarrow \text{boy} \mid \text{ice-cream} \mid \text{dog}$

$V \rightarrow \text{ate} \mid \text{like} \mid \text{bite},$

$Art \rightarrow \text{the} \mid \text{a}\}$

- Applying the rules from P sequentially sentences generated:

- The dog bites boy.
- **Boy bites the dog.**
- Boy ate ice-cream.
- The dog bite the boy.

Example-2

Lexicon:

DET \rightarrow a | the

ADJ \rightarrow beautiful | perching

N \rightarrow bird | birds | grain | grains

V \rightarrow peck | pecks | pecking

Production Rules:

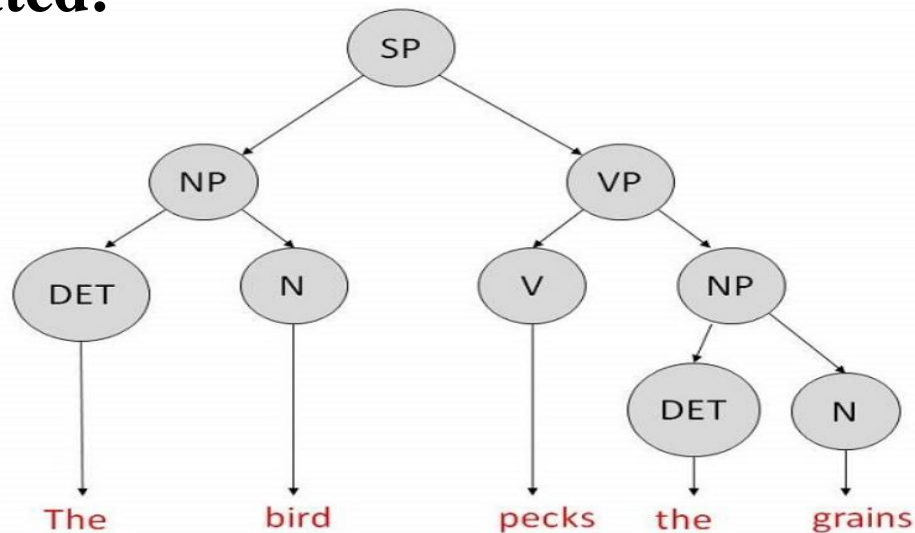
SP \rightarrow NP VP

NP \rightarrow DET N | DET ADJ

N

VP \rightarrow V NP

Parse tree generated:



Types of Language Modeling

- Uses the grammar of a language to create its model

Grammar-based

Example: **Transformational Grammar(TG)**

2 types

Statistical modeling

- Creates a language model by training it from a corpus

Example:

n-gram model

Grammar-based Language Modeling

- Uses the grammar of a language to create its model
- Represent the syntactic structure of language

Example:

S -> NP+VP //S: Sentence, NP: Noun Phrase, VP: Verb Phrase

- It utilizes, the structure of a sentence (NP, VP) and relationships between these structures.
- Examples of grammar based models:
 - **Transformational Grammar (TG)**, Government and Binding (GB), Lexical Functional Grammar (LFG), Paninian grammar

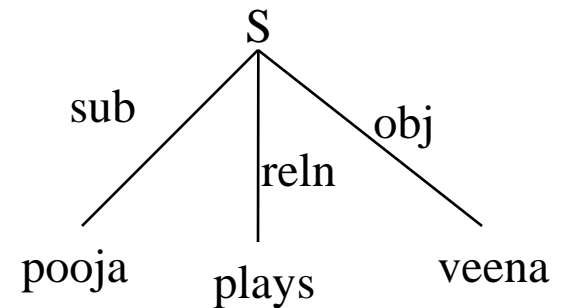
Transformational Grammar (Chomsky 1957)

- Assumes 2 levels of existence of sentences:
 - Surface level
 - Deep root level

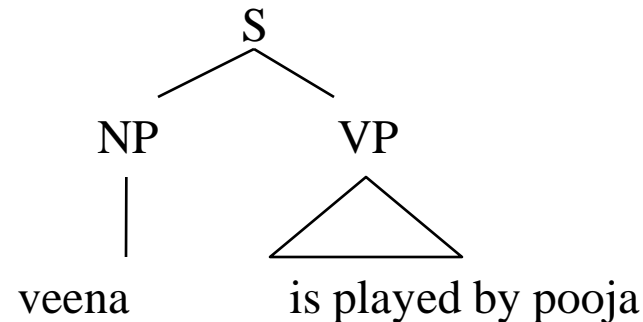
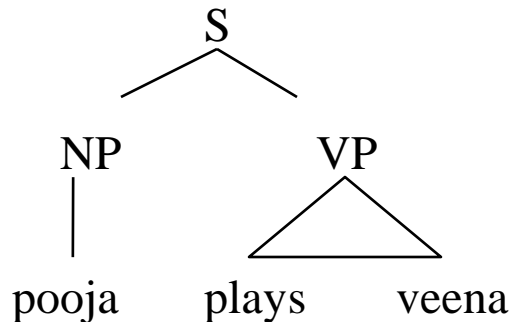
Example:

S: pooja plays veena

Deep structure:



Surface structure:



Cont...

- Transformational grammar has three components:

- 1. Phrase structure grammar**

- 2. Transformational rule**

- 3. Morphophonemic rule**

1. Phrase Structure Grammar

- Consists of rules that generate natural language sentences
- Assigns a structural description of natural languages

Example:

Set of rules:

$S \rightarrow NP + VP$

$VP \rightarrow V + NP$

$NP \rightarrow Det + N$

S: sentence

NP: Noun Phrase

VP: Verb phrase

Det: Determiner

N: Noun

V: Verb

- Sentences generated using these rules are called grammatical

2. Transformational Rule

- Used to transform one surface representation into another
- Applied on the terminal strings generated by the phrase structure rules

Example:

- converting active sentence into passive

Plays-> played

3. Morphophonemic rule

- Match each sentence representation to a string of phonemes

Example:

morpho-phonemic rule will convert:

catch +en → caught

Limitations of Grammar-based Language Models

- A large number of re-writable rules which are language specific
- Generation of a complete set of rules covering all languages is a challenging task.

Statistical Language Models

- A statistical language model is a probability distribution $P(s)$ over all possible word sequence (or words, sentences, paragraph, documents, etc).
- It creates a language model by training it from a corpus
- The goal of statistical language model is to estimate the probability (likelihood) of a sentence.

Example:

n-gram model

Cont...

- The n-gram model has its applications in-
 - Speech recognition
 - Machine translation
 - Spelling correction
 - Information retrieval, etc

Cont...

- The probability (likelihood) of a sentence can be estimated by decomposing sentence probability into a product of conditional probabilities using the chain rule:

$$\begin{aligned} P(s) &= P(w_1, w_2, w_3, \dots, w_n) \\ &= P(w_1) \cdot P(w_2/w_1) \cdot P(w_3/w_1 w_2) \dots P(w_n/w_1 w_2 \dots w_{n-1}) \\ &= \prod_{i=1}^n P(w_i / h_i) \end{aligned}$$

where h_i is history of word w_i defined as-

$$w_1 w_2 \dots, w_{i-1}$$

Cont...

- In order to calculate sentence probability,
 - the probability of each word is needed to be calculated,
 - given the sequence of words preceding it.
- The **n-gram** model approximates the probability of a word
 - given all the previous words by the conditional probability **given previous n-1 words** only.

n-gram Model

- An n-gram model calculates $P(w_i/h_i)$ by modeling language as Markov model of order $n-1$.
- i.e. by looking at *previous $n-1$* words only.

when $n=1$: **uni-gram**

$n=2$: **Bi-gram**

$n=3$: **Tri-gram**

$n=4$: **Four-gram**

$n=5$: **Five-gram**

.....

Example

- Given a sequence of letters what is the likelihood of the next letter.

Example:

“for ex....

covid....

covid-19

covid 19 India

covid vaccine

.....

Uni-gram Model

- In a uni-gram model, the probability of each word depends on its own probability in the document/ corpus

$$P(W) = \frac{\text{no. of times } W \text{ appears}}{\text{Total no. of words in corpus}}$$

Example-1:

If total no. of words in a corpus= 1,000,00 and the word 'the' appears 69971 times.

using unigram model,

$$P(\text{the}) = 69971/100000$$

$$= 0.69971$$

$$= 0.7$$

Example-2

For a training corpus with 10,000 words, find the uni-gram probability of the terms with respect to the below-given frequency counts:

<u>Term:</u>	<u>Frequency:</u>
football	58
cricket	112

Answer:

using unigram model,

$$\begin{aligned} P(\text{football}) &= 58/10000 \\ &= 0.0058 \end{aligned}$$

$$\begin{aligned} P(\text{cricket}) &= 112/10000 \\ &= 0.0112 \end{aligned}$$

Example-3

Find the unigram probability of the term "staff" in the training corpus sentence-

"the students and staff of this Institute"

Answer:

using unigram model,

$$\begin{aligned} P(\text{staff}) &= 1/7 \\ &= 0.142 \end{aligned}$$

Example-4

Find the unigram probability of the term "staff" in the training corpus sentence-

“The students and staff of the Institute are very professional”

Answer:

using unigram model,

$$P(\text{staff}) = 1/10$$

$$= 0.1$$

Bi-gram Model

- In a bi-gram model, the probability of each word depends on its previous word only.
- Probability of a word w_i can be calculated as-

$$P(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

- Probability of a sentence can be calculated by multiplying the bigram probabilities of each term in the sentence
- For a sentence S with words $w_1 w_2 \dots w_n$ the probability of the sentence:

$$P(s) = P(w_1 / \langle s \rangle) \cdot P(w_2 / w_1) \cdot \dots \cdot P(w_n / w_{n-1})$$

Example-1

Training set:

The Arabian Knights

These are the fairy tales of the east

The stories of the Arabian Knights are translated in many languages

Find the probability of the test sentence given below considering a bigram model

Test sentence:

The Arabian Knights are the fairy tales of the east

Answer

Training set:

<s>The Arabian Knights</s>

<s>These are the fairy tales of the east </s>

<s>The stories of the Arabian Knights are translated in many languages </s>

Bigram probabilities of words in training set:

$$P(\text{the}/<s>) = 2/3 = 0.67$$

$$P(\text{Arabian}/\text{the}) = 2/5 = 0.4$$

$$P(\text{Knights}/\text{Arabian}) = 2/2 = 1.0$$

Answer

Training set:

<s>The Arabian Knights</s>

<s>These are the fairy tales of the east </s>

<s>The stories of the Arabian Knights are translated in many languages </s>

Bigram probabilities of words in training set:

$P(\text{the}/<s>) = 2/3 = 0.67$ $P(\text{Arabian}/\text{the}) = 2/5 = 0.4$ $P(\text{Knights}/\text{Arabian}) = 2/2 = 1.0$

$P(\text{these}/<s>) = 1/3 = .33$ $P(\text{are}/\text{these}) = 1/1 = 1.0$ $P(\text{the}/\text{are}) = 1/2 = 0.5$

$P(\text{fairy}/\text{the}) = 1/5 = 0.2$ $P(\text{tales}/\text{fairy}) = 1/1 = 1.0$ $P(\text{of}/\text{tales}) = 1/1 = 1.0$

$P(\text{the}/\text{of}) = 2/2 = 1.0$ $P(\text{east}/\text{the}) = 1/5 = 0.2$

$P(\text{stories}/\text{the}) = 1/5 = 0.2$ $P(\text{of}/\text{stories}) = 1/1 = 1.0$ $P(\text{are}/\text{Knights}) = 1/2 = 0.5$

$P(\text{translated}/\text{are}) = 1/2 = 0.5$ $P(\text{in}/\text{translated}) = 1/1 = 1.0$ $P(\text{many}/\text{in}) = 1/1 = 1.0$

$P(\text{language}/\text{many}) = 1/1 = 1.0$

Estimating Sentence Probability

$P(\text{The Arabian Knights are the fairy tales of the east})$

$$\begin{aligned} &= P(\text{The}/<s>) \times P(\text{Arabian}/\text{the}) \times P(\text{Knights}/\text{Arabian}) \times P(\text{are}/\text{Knights}) \\ &\quad \times P(\text{the}/\text{are}) \times P(\text{fairy}/\text{the}) \times P(\text{tales}/\text{fairy}) \times P(\text{of}/\text{tales}) \\ &\quad \times P(\text{the}/\text{of}) \times P(\text{east}/\text{the}) \end{aligned}$$

$$\begin{aligned} &= 0.67 \times 0.4 \times 1.0 \times 0.5 \\ &\quad \times 0.5 \times 0.2 \times 1.0 \times 1.0 \\ &\quad \times 1.0 \times 0.2 \end{aligned}$$

$$= 0.00268$$

Example-2

For the below given training corpus find the bigram probabilities of all words:

Training Corpus:

I am Sam

Sam I am

I do not like eggs

Also, find the sentence probability of the test sentence:

I am Sam

Answer

Training Corpus:

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like eggs </s>

Bigram probabilities:

$$P(I/<s>) = 2/3 = 0.67$$

$$P(\text{Sam}/<s>) = 1/3 = 0.33$$

$$P(\text{do}/I) = 1/3 = 0.33$$

$$P(\text{am}/I) = 2/3 = 0.67$$

$$P(I/\text{Sam}) = 1/2 = 0.5$$

$$P(\text{not}/\text{do}) = 1/1 = 1.0$$

$$P(\text{Sam}/\text{am}) = 1/2 = 0.5$$

$$P(\text{like}/\text{not}) = 1/1 = 1.0$$

$$P(\text{eggs}/\text{like}) = 1/1 = 1.0$$

Training sentence probability:

$P(I \text{ am Sam})$

$$= P(I/<s>) \cdot P(\text{am}/I) \cdot P(\text{Sam}/\text{am})$$

$$= 0.67 * 0.67 * 0.5$$

$$= 0.2244$$

Example-3

Consider the following frequency matrix and Find the likelihood estimate of the below sentence using bigram model:

‘The Arabian Knights’

Term	frequency
<s>	3
The	5
Arabian	2
Knights	2

word sequence	Relative frequency
<s> The	3
The Arabian	2
Arabian Knights	2

Answer

$P(\text{The Arabian Knights})$

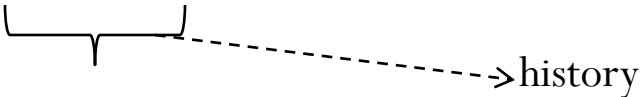
$= P(\text{the}/<s>) \cdot P(\text{Arabian}/ \text{the}) \cdot P(\text{Knights} / \text{Arabian})$

$= (3/3) \cdot (2/5) \cdot (2/2)$

$= 0.4$

Tri-gram Model

- In a tri-gram model, the probability of each word depends on its previous 2 word.
- For the word sequence-

$$w_1, w_2, w_3, \dots, w_{i-2}, w_{i-1}, w_i, \dots, w_n$$


- Probability of a word w_i can be calculated as:

$$P(w_i/w_{i-2}, w_{i-1}) = \frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$$

Example:

The Arabian **Knights**

P (**Knights**/ the Arabian)

Example-1

Find probability of the below given test sentence using a tri-gram model.

Training set:

The Arabian Knights

These are the fairy tales of the east

The stories of the Arabian Knights are translated in many languages

Test Sentence:

The Arabian Knights are the fairy tales of the east

Answer

Training set:

<s1><s2>The Arabian Knights </s2></s1>

<s1><s2>These are the fairy tales of the east </s2></s1>

<s1><s2>The stories of the Arabian Knights are translated in many languages </s2></s1>

Trigram probabilities:

$P(\text{The}/<s1><s2>) = 2/3 = 0.666 = 0.67$

$P(\text{Arabian}/<s2>\text{the}) = 1/2 = 0.5$

$P(\text{Knights}/ \text{the Arabian}) = 2/2 = 1.0$

.....

// find the tri-gram probabilities of other sequences

Estimating Sentence Probability

$P(\text{The Arabian Knights are the fairy tales of the east})$

$= P(\langle s1 \rangle \langle s2 \rangle \text{The Arabian Knights are the fairy tales of the east})$

$= P(\text{The} / \langle s1 \rangle \langle s2 \rangle) \times P(\text{Arabian} / \langle s2 \rangle \text{the}) \times P(\text{Knights} / \text{the Arabian})$
 $\times P(\text{are} / \text{Arabian Knights}) \times P(\text{the} / \text{Knights are}) \times P(\text{fairy} / \text{are the})$
 $\times P(\text{tales} / \text{the fairy}) \times P(\text{of} / \text{fairy tales}) \times P(\text{the} / \text{tales of})$
 $\times P(\text{east} / \text{of the})$

$= 0.67 * 0.5 * 1.0 * 0.5 * 0 * 1 * 1 * 1 * 1 * 1$

$= 0$

Data Sparseness Problem in n-gram Model

- An n-gram that does not occur in the training set is assigned a zero probability
- **Smoothing techniques:**
 - Add-one smoothing: assigns a count of 1 to unseen n-grams
 - Modified version of add-one smoothing technique adds a value of $(1/|V|)$ to unseen ngrams, where V is the vocabulary size.
 - Caching technique: assigns the probability of the most recent sequence

Applying smoothing technique on Example-1 of Trigram model

$P(\text{The Arabian Knights are the fairy tales of the east})$

$= P(\langle s1 \rangle \langle s2 \rangle \text{The Arabian Knights are the fairy tales of the east})$

$= P(\text{The} / \langle s1 \rangle \langle s2 \rangle) \times P(\text{Arabian} / \langle s2 \rangle \text{the}) \times P(\text{Knights} / \text{the Arabian})$
 $\times P(\text{are} / \text{Arabian Knights}) \times P(\text{the} / \text{Knights are}) \times P(\text{fairy} / \text{are the})$
 $\times P(\text{tales} / \text{the fairy}) \times P(\text{of} / \text{fairy tales}) \times P(\text{the} / \text{tales of})$
 $\times P(\text{east} / \text{of the})$

Applying add 1 smoothing

$$P = 0.67 * 0.5 * 1.0 * 0.5 * 1 * 1 * 1 * 1 * 1 * 1$$
$$= 0.1675$$

Example-2: Homework

For the below given training set find tri-gram probability of all terms in training set and find the probability of the given test sentence .

Training set:

I am Sam

Sam I am

I am not Sam

Test sentence:

I am not Sam

Answer: **0.2211**

Example-3: Homework

Using sentence S_1 , S_2 , S_3 as training data, find the probability of S_4 using a bigram model.

S_1 : The section of all the intelligent students

S_2 : The students

S_3 : The intelligent students of this college

S_4 : The section of all the intelligent students of this college

Answer: **0.0103**

Assignment Question-1:

Differentiate between a unigram and bigram model with an example. Find the **tri-gram probabilities** of all terms in the training set given below (S1, S2, S3, S4) and the probability of the test sentence S5. Use a modified version of add-one smoothing technique ($1/|V|$, where V is the vocabulary size) wherever needed.

Test Sentence:

S5: The applications of artificial intelligence in industry

Training set:

S1: The applications of data science in industry

S2: The applications of machine learning

S3: Applications of artificial intelligence

S4: human computer interaction technology with computational intelligence

Answer

$$P(\text{The}|\langle s1 \rangle \langle s2 \rangle) = 2/4 = 0.5$$

$$P(\text{applications}|\langle s2 \rangle \text{The}) = 2/2 = 1$$

$$P(\text{of}|\text{the applications}) = 2/2 = 1$$

$$P(\text{artificial}|\text{applications of}) = 1/3 = 0.33$$

$$P(\text{Intelligence}|\text{of artificial}) = 1/1 = 1$$

$$P(\text{in}|\text{artificial intelligence}) = 0/1 = 1/17 = 0.058$$

$$P(\text{industry}|\text{intelligence in}) = 1/0 = 1/17 = 0.058$$

....

Similarly estimate other training set probabilities.....

Estimating test sentence probability:

$$\begin{aligned} P(S5) &= P(\text{The}|\langle s1 \rangle \langle s2 \rangle) * P(\text{applications}|\langle s2 \rangle \text{The}) \\ &\quad * P(\text{of}|\text{the applications}) * P(\text{artificial}|\text{applications of}) \\ &\quad * P(\text{Intelligence}|\text{of artificial}) * P(\text{in}|\text{artificial intelligence}) \\ &\quad * P(\text{industry}|\text{intelligence in}) \\ &= 0.5 * 1 * 1 * 0.33 * 1 * 0.058 * 0.058 \\ &= \mathbf{0.000555} \end{aligned}$$

Assignment Question-2

Differentiate between a bigram and a trigram model with an example. Find the **bi-gram probabilities** of all terms in the training set given below and the probability of the test sentence.

Use a modified version of add-one smoothing technique ($1/|V|$, where V is the vocabulary size) wherever needed.

Training Set:

The section of all the intelligent students

Students of the college

The students of this college

The intelligent students of this college

Test Sentence:

The section of all the intelligent students of this college

Answer

Bigram probabilities of words in training set:

$$P(\text{the}|\langle s \rangle) = 3/4 = 0.75$$

$$P(\text{of}|\text{section}) = 1/1 = 1$$

$$P(\text{the}|\text{all}) = 1/1 = 1$$

$$P(\text{students}|\text{intelligent}) = 2/2 = 1$$

$$P(\text{of}|\text{students}) = 3/4 = 0.75$$

$$P(\text{college}|\text{the}) = 1/5 = 0.2$$

$$P(\text{this}|\text{of}) = 2/4 = 0.5$$

$$P(\text{section}|\text{the}) = 1/5 = 0.2$$

$$P(\text{all}|\text{of}) = 1/4 = 0.25$$

$$P(\text{intelligent}|\text{the}) = 2/5 = 0.4$$

$$P(\text{students}|\langle s \rangle) = 1/4 = 0.25$$

$$P(\text{the}|\text{of}) = 1/4 = 0.25$$

$$P(\text{students}|\text{the}) = 1/4 = 0.25$$

$$P(\text{college}|\text{this}) = 2/2 = 1$$

Estimating test sentence probability:

$P(\text{The section of all the intelligent students of this college})$

$$\begin{aligned} &= P(\text{The}|\langle s \rangle) * P(\text{section}|\text{the}) * P(\text{of}|\text{section}) * P(\text{all}|\text{of}) * P(\text{the}|\text{all}) \\ &\quad * P(\text{intelligent}|\text{the}) * P(\text{students}|\text{intelligent}) * P(\text{of}|\text{students}) \\ &\quad * P(\text{this}|\text{of}) * P(\text{college}|\text{this}) \end{aligned}$$

$$= 0.75 * 0.2 * 1 * 0.25 * 1 * 0.4 * 1 * 0.75 * 0.5 * 1$$

$$= \mathbf{0.005625}$$

