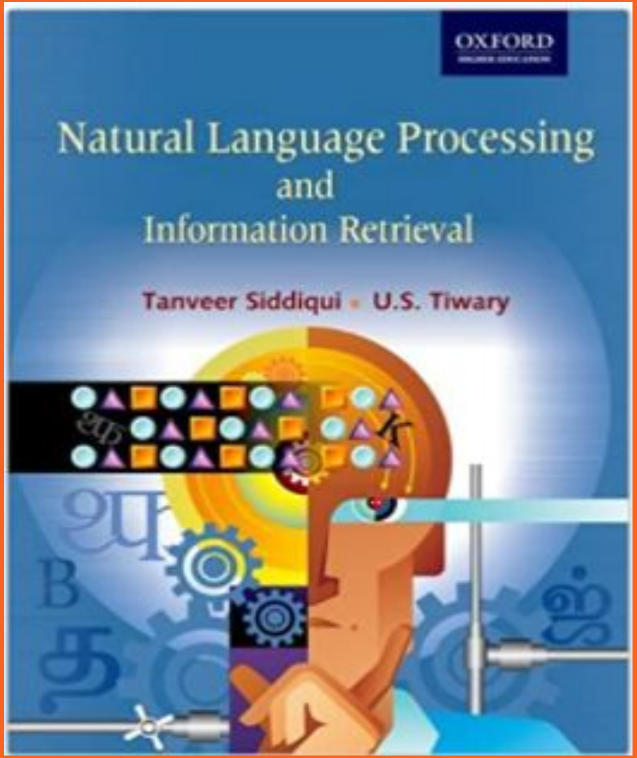
Module -1

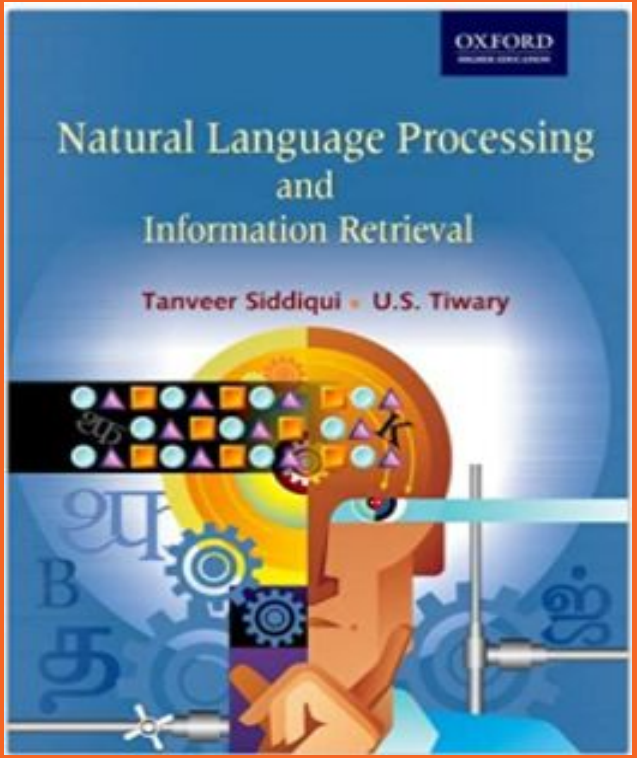
Overview and Language Modelling


OXFORD
Natural Language Processing and Information Retrieval
Tanveer Siddiqui · U.S. Tiwary

Chapter 1

Introduction

- ○Learn idea of natural Language Processing
- ○Origins of NLP
- ○Language and Knowledge
- ○ Role of grammar in language processing and transformational grammar
- ○ Challenges of NLP
- ○ Applications of NLP
- ○ Processing Indian languages
- ○ Information Retrieval



What is Natural Language Processing(NLP)

- Language is the primary means of communication by humans.
- Tool used to express ideas and emotions.
- It is hard to realize how we process natural language.
- But there must be some kind of representation in mind i.e. “*Content of Language*”.

Contd...

- NLP is concerned with *development of computational models* of aspects of human language processing.
- 2 main reasons for such development are:
  - ★ To develop automated tools for language processing.
  - ★ To gain a better understanding of human communication.
- Building computational models should have knowledge on how human acquires ,store and process language.
- Knowledge of world and language.

Contd…

- 2 major approaches of NLP are:
- Rationalist Approach– assumes *existence of some language faculty.*
    - Like it *is not possible to learn children complex things* (cognitive capacity shared by all normal human beings) in brain.

- Empiricist Approach – *Do not believe in the existence of a language faculty.*
    - Assume existence of general organization principles such as *pattern recognition, generalization and association*.
    - Learning of detailed structures take place through application of these principles on *sensory inputs* available to child.

Origins of NLP

- Mistakenly termed as Natural Language Understanding
- Understanding only involves interpretation of language
- NLP includes both
  - *Understanding (interpretation)*
  - *Generation (production)*

- NLP also includes Speech processing
- In the prescribed Text books we study:

  Text processing—covering area of Computational Linguistics, Theoretical linguistics,Psycholinguistics

Contd...

- **Theoretical Linguistics:**
  - Study the *structural description of natural language* and its semantics(meaning of word/Phrase/Text).
  - Identify the rules that describe and restrict the structure of languages(grammar)

- Example: Most of the languages have constructs like noun and verb phrases.

Contd...

- **Psycholinguistics:**
  - Explains *how human produce and comprehend natural language*
  - It is interested in understanding about:
    - how people identify the appropriate structure of a sentence
    - when they decide on the appropriate meaning for words
    - how are word meanings identified
    - when does understanding take place
  - Rely on *empirical investigations* to back up their theories.

Contd...

- **Computational Linguistics:**
  - Study of language using *computational models* of linguistic phenomena
  - Employ notions of algorithms and data structures from computer science.
  - Deals with the application of *linguistic theories and computational techniques* for NLP
  - Take advantage of what is known from all other disciplines.

Contd...

- **Computational Models deals about :**
  - How is the structure of sentences identified
  - How can knowledge and reasoning be modeled
  - How can language be used to accomplish specific tasks

- Problem is:
- How to represent a language?
  - Most Knowledge representations tackle only small part of knowledge. Because representing whole body of knowledge is almost impossible.

Contd...

- **Computational Models broadly classified under:**
  - Knowledge-driven
  - Data-driven

- Knowledge Driven
  - Rely on explicitly coded linguistic knowledge
  - Expressed as a set of *handcrafted grammar rules*.

  Problems
  - Acquiring and encoding such knowledge is difficult
  - Lack of sufficient coverage of domain knowledge

Contd...

- Data Driven
  - Presume the *existence of a large amount of data*
  - Usually employ some *Machine Learning techniques* to learn syntactic patterns

  - Advantages are:
  - Less human effort
  - Performance of such systems depend on quantity of data
  - Such systems are adaptive to noisy data

Language and Knowledge

- Language-medium of expression in which knowledge is deciphered.
- Processing language means of processing content of it.

- Language (Text) processing has 5 different levels of analysis are:
  - *Lexical Analysis*
  - *Syntactic Analysis*
  - *Semantic Analysis*
  - *Discourse Analysis*
  - *Pragmatic Analysis*

Language and Knowledge

- **Lexical Analysis:**
- Involves analyses of *words*
- Requires *morphological knowledge* (structure and word formation from basic units)
- Word formation rules are language specific.

Language and Knowledge

- **Syntactic Analysis:**
- Considers *sequence of words* as a unit (generally, sentence) and finds its structure.
- Decomposes a sentence into its constituents.
- Identifies how the constituents relate to each other.
- Captures grammaticality/ungrammaticality of sentences considering constraints such as word order, number, case agreement etc.
- **Example**: *'I went to the market '* is valid sentence but not valid sentence is : *'went the I market to'*

Language and Knowledge

- **Semantic Analysis:**

- Associated with the *meaning of the language*
- Concerned with creating meaningful representation of linguistic inputs.
- Map natural language sentences or utterances onto some representation of meaning.
- Grammatically valid sentences can be meaningless.

- **Example:** Colourless green ideas sleep furiously

Language and Knowledge

- **Discourse Analysis:**

- Attempts to interpret the structure and meaning of units *larger than sentence* such as *paragraph, document* etc. in terms of words, phrases, clusters and sentences.
- It requires the *discourse knowledge*, i.e., knowledge of how meaning of sentence is determined by preceding sentence.
- Requires the *resolution of anaphoric references* and *identification of discourse structure.*
- **Example**: *Radha* is a girl. *She* went to to market.*It* was too rush.

Language and Knowledge

- **Pragmatic Analysis:**

- Deals with *purposeful use of sentences* in situation.
- Requires knowledge of world (knowledge that extends beyond the contents of text)
- **Example:** *John saw Mary in a garden with a cat*

The Challenges of NLP

- Related to the *representation* and *interpretation* like human.

1. Representation of meaning of a sentence, meaning of words appearing in it.

- Meaning of word and its use in language.

  Example: *'I like Ice cream'* instead if we use *'Ice cream like I'*

- *Words as well as their syntactic and semantic relation that gives meaning to a sentence*

  Example: *'Kabir and Ayan are Married'*
  'Kabir and Suha are Married'

Contd...

- Language keeps evolving
    - New words are added continually
    - Existing words are introduced in new context.
    - **Example: 9/11** => Terrorist act on World Trade Centre in USA in 2004

2.Machine must rely on *word contexts* to learn the meaning of specific word in a message.

- Context depends on co-occurring words (occurring before or after word)
- Word Frequency in a particular sense also affects its meaning
- **Example:** *'while'* as conjunction or as *'a short interval of time'*

[Note: Example :You can go swimming *while* I am having food.
Once in a *while* it happened so.]

Contd...

3.*Idioms, metaphor and ellipses* add to complexity in *identifying meaning* of written text.

Example: Meaning of *'The old man finally kicked the bucket'* has nothing to do with words *'kick'* and *'bucket'.*

[Note:
Example:Idiom-it's piece of cake-it's easy
Metaphor- Time is money-something is referred to something
Ellipses-were you thinking about me today…?]

Contd...

4.*Quantifier scoping (the,each,etc)* is not clear and poses problem in automatic processing.

5.*Ambiguity of natural languages* is another difficulty.

- **Ambiguity at word level:** we can identify words that have multiple meanings associated with them.
- **Example:** *‘can’* (as verb in *‘can play’* / as noun- as ‘*Empty can*’ means that *container*)
- *‘bank’* (financial institution or as river side)
- *‘bat’* (mammal or playing equipment)

Contd...

- **Structure Ambiguity-Ambiguity at sentence level:**None of the words are ambiguous, but the sentence is.
  - **Example:** *'Stolen rifle found by the tree'*

6. Incorporating *contextual and world knowledge* (culture, language, traditions, …) is greatest difficulty in language computing.
   - **Example:** *'Taj'* means a monument, a brand of tea, or a hotel to an Indian but need not/may not be so for non-Indian.

Contd...

What challenges makes NLP difficult?

- Writing *grammar rules* and the grammar itself for describing the structure of a sentence is complex
- *Non-grammatical sentences* could be interpreted as a semantically correct sentence by humans, which *machines cannot.*
- It is almost impossible for grammar to *capture the structure of all* and *only meaningful* text.

Language and Grammar

- Automatic processing of language requires the *rules* and *exceptions* of language to be explained to the computer.

- Grammar *defines* language

- Grammar consists of a *set of rules* that allow us to *parse* and *generate* sentences of a language.

Language and Grammar

- **Main Hurdle in Language Specifications are:**
  - Constantly *changing nature* of natural languages
  - Presence of a large number of *hard-to-specify* exceptions

Language and Grammar

- **Several efforts made to develop language specification grammars are:**
  - Transformational Grammar (Chomsky, 1957)
  - Lexical Functional Grammar (Kaplan and Bresnan, 1982)
  - Government and Binding (Chomsky, 1981)
  - Generalized phrase structure grammar (Derivation)
  - Dependency grammar (Relationships)
  - Paninian grammar (Relationships)
  - Tree-adjoining grammar (Joshi, 1985)

Transformational Grammar-Noam Chomsky

- Proposed by Noam Chomsky in 1957.
- Hierarchy of formal grammar based on level of complexity
- Grammar uses *phrase structure rules* (Rewrite rules)
- General framework - *Generative grammar*
- Any grammar that uses a set of rules to specify or generate all and only grammatical (well-formed) sentences in a language.

Transformational Grammar-Noam Chomsky

- In transformational grammar ,each sentence in a language has *two levels* of representations:

1. *Deep structure*- represents meaning
2. *Surface structure*-an utterance[words we communicate]

- Mapping from deep to surface structure is carried out by transformations

Transformational Grammar-Noam Chomsky

In sentence we can identify following :

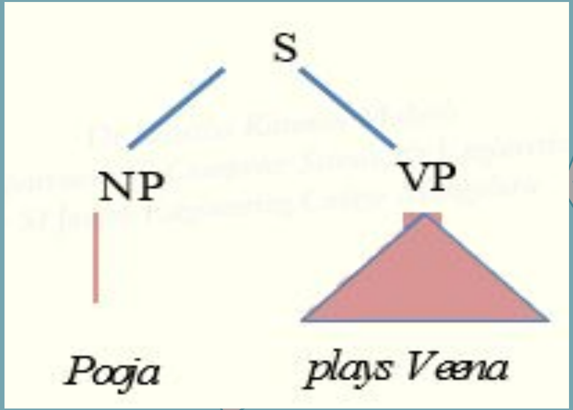
- **Example 1:** Radha went to market
  Here, Radha is *Noun Phrase* and went to to market is *Verb Phrase*

- **Example 2**: The sisters went to market
  Here, The sisters is *Noun Phrase* and went to to market is *Verb Phrase*

- **Example 3 :** The sisters who stays in hostel went to market
  Here, The sisters who stays in hostel is *Noun Phrase* and went to to market is *Verb Phrase*

Transformational Grammar-Noam Chomsky

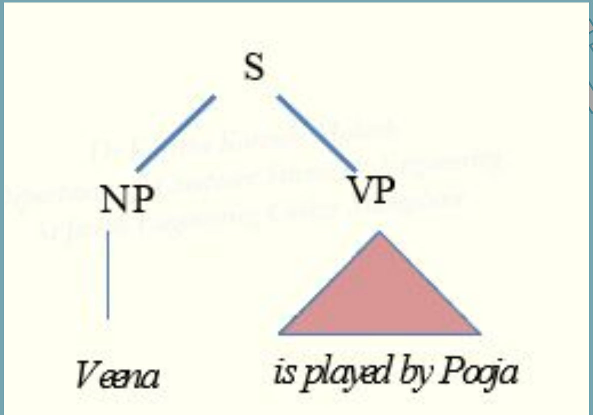
- **Example :** Pooja plays Veena
- *Surface structure* can be represented as:



- Pooja is *Noun* and Veena is object.

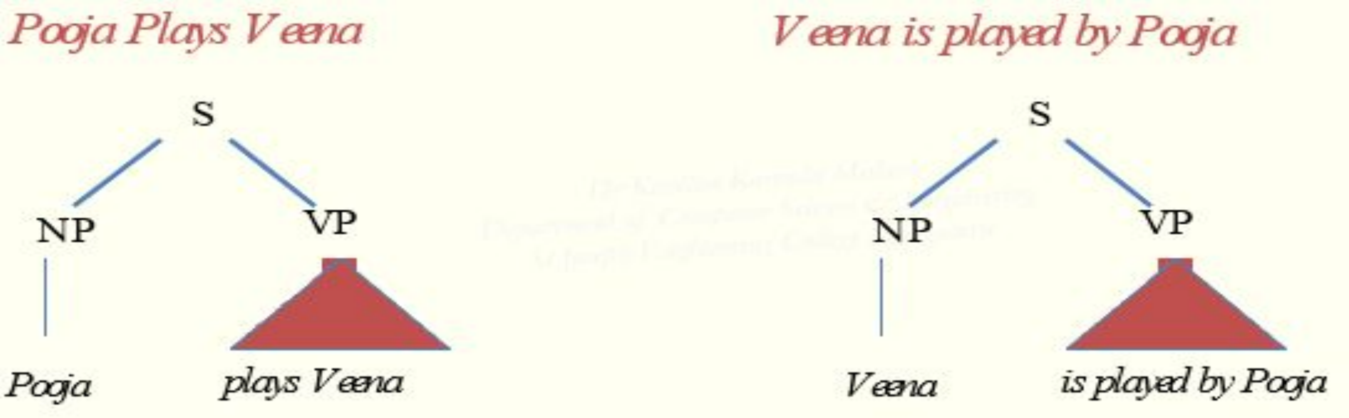
Transformational Grammar-Noam Chomsky

- **Example :** Veena is played by pooja
- *Surface structure* can be represented as:



Transformational Grammar-Noam Chomsky

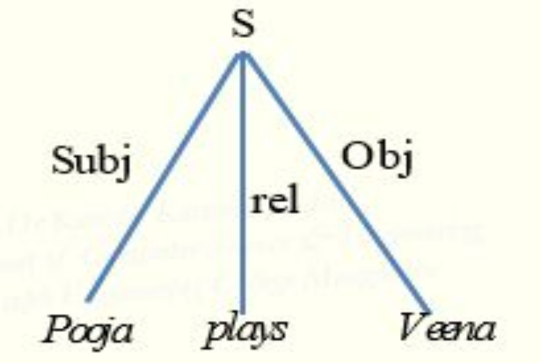
- Deep structure can be transformed in a number of ways to yield *different surface-level representations* having *same meaning*.
- Sentences with different surface-level representations having same meaning share a *common deep-level representation.*
- **Example:**



Transformational Grammar-Noam Chomsky

- **Example :** Pooja plays Veena
  Veena is played by pooja

- *Deep structure* can be represented as:



- Here, *Deep subject* is Pooja

- *Deep object* is Veena.

In Grammar-Subject

- Noun or noun phrase,pronoun that does the action of verb

  Example : *I* love Ice cream.

- It can be a group of words

  Example : *The doctor* went to clinic .

- May include a noun and all the words that are used to *add extra information* to that noun

  Example : *The women in front of office* selling variety of flowers.

In Grammar-Subject

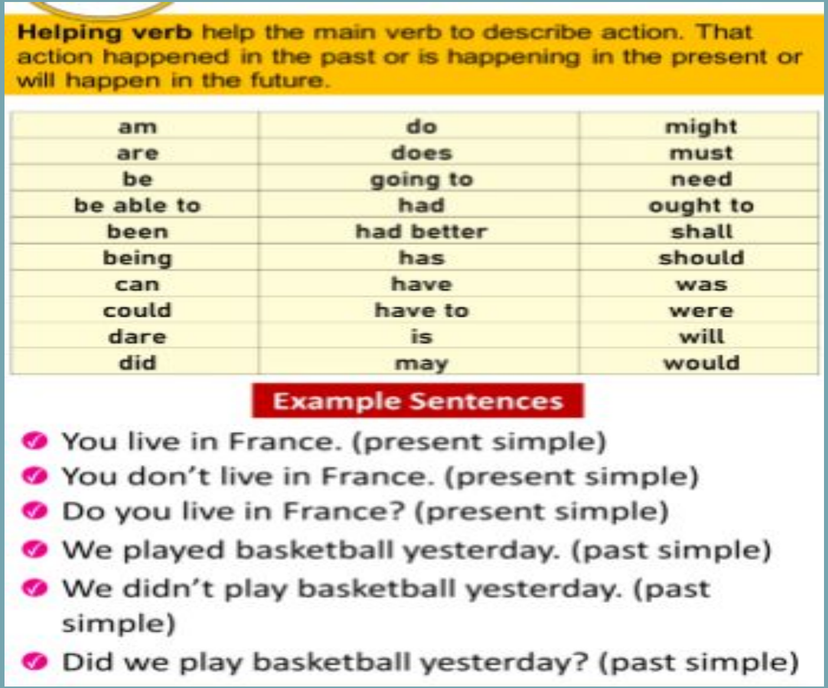
- Can include *two or more nouns* that each have groups of words giving us extra information.

Example:

- *The man whose son I met in high school and the manager of the office where shaw works* played cricket.

In Grammar-Verb

- **Auxiliary/Helping Verbs :**

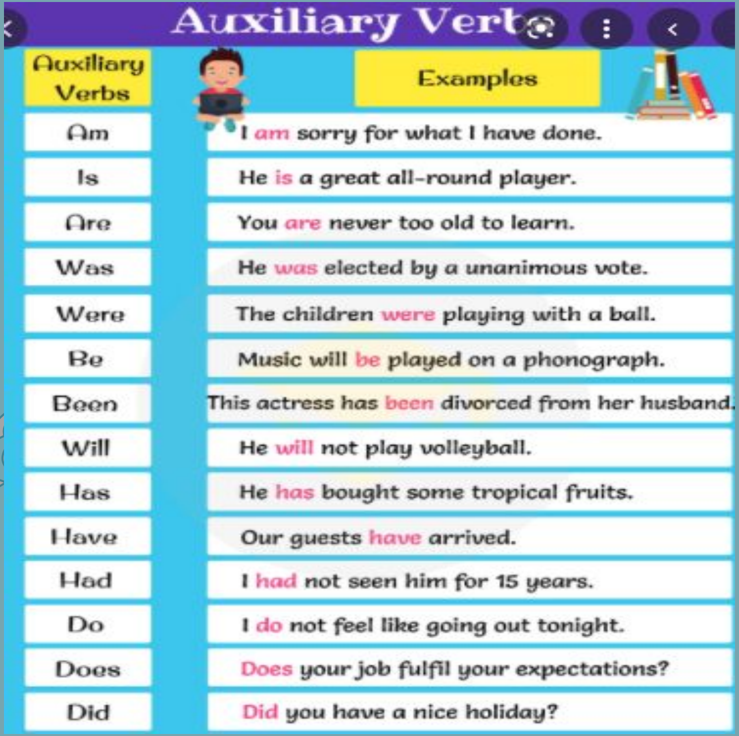


Helping verb help the main verb to describe action. That action happened in the past or is happening in the present or will happen in the future.

am	do	might
are	does	must
be	going to	need
be able to	had	ought to
been	had better	shall
being	has	should
can	have	was
could	have to	were
dare	is	will
did	may	would

Example Sentences

- You live in France. (present simple)
- You don't live in France. (present simple)
- Do you live in France? (present simple)
- We played basketball yesterday. (past simple)
- We didn't play basketball yesterday. (past simple)
- Did we play basketball yesterday? (past simple)

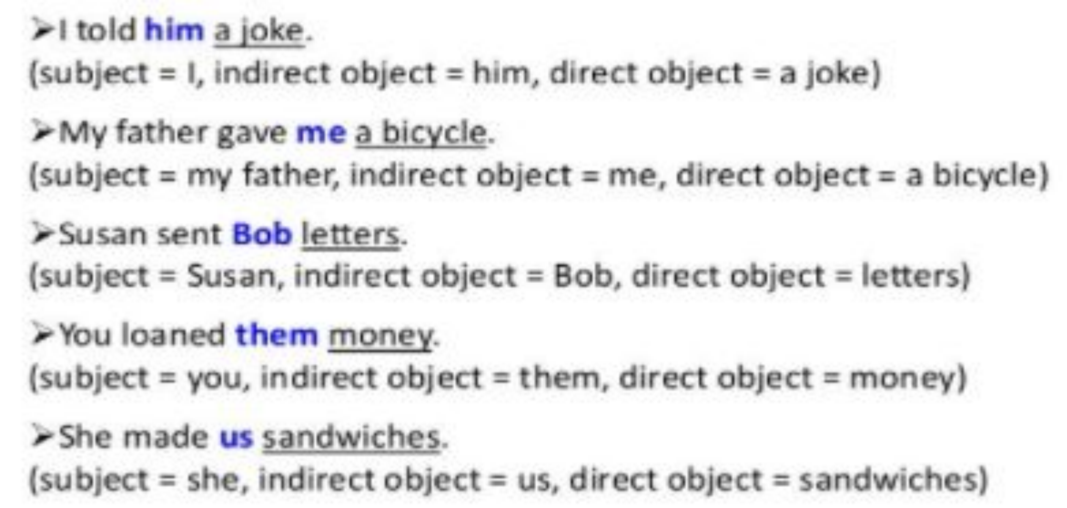


Auxiliary Verbs

Auxiliary Verbs	Examples
Am	I am sorry for what I have done.
Is	He is a great all-round player.
Are	You are never too old to learn.
Was	He was elected by a unanimous vote.
Were	The children were playing with a ball.
Be	Music will be played on a phonograph.
Been	This actress has been divorced from her husband.
Will	He will not play volleyball.
Has	He has bought some tropical fruits.
Have	Our guests have arrived.
Had	I had not seen him for 15 years.
Do	I do not feel like going out tonight.
Does	Does your job fulfil your expectations?
Did	Did you have a nice holiday?

In Grammar-Object

- The entity that is acted upon by the subject.
- It can be a noun, a noun phrase, a pronoun or a longer complex object, which is modified (respect to complex subject).
- **Example:**



➢I told **him** a joke.
(subject = I, indirect object = him, direct object = a joke)

➢My father gave **me** a bicycle.
(subject = my father, indirect object = me, direct object = a bicycle)

➢Susan sent **Bob** letters.
(subject = Susan, indirect object = Bob, direct object = letters)

➢You loaned **them** money.
(subject = you, indirect object = them, direct object = money)

➢She made **us** sandwiches.
(subject = she, indirect object = us, direct object = sandwiches)

Transformational Grammar-Noam Chomsky

- 3 components of Transformational Grammar are:
    - *Phrase structure grammar*
    - *Transformational rules*
    - *Morphophonemic rules*

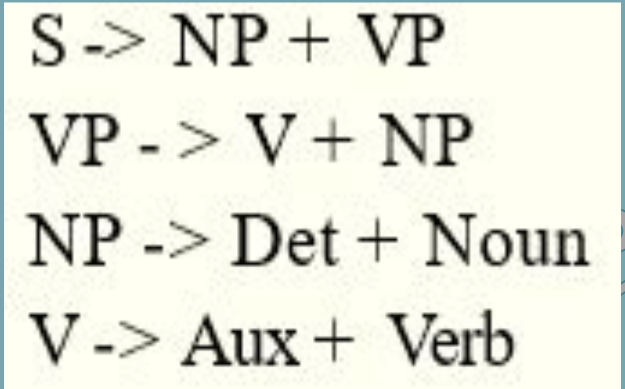
- Each component consists of a set of rules

[Note: morphology- internal construction of words.
Example: sleep-slept, bind-bound
Phonemes -Sound in a specific language p,b,d,t in bat, bad etc.]

Transformational Grammar-Noam Chomsky

Phrase structure grammar:

- Consists of *rules* that generate natural language sentence and assign a structural description to them
- **Consider the Rules:**

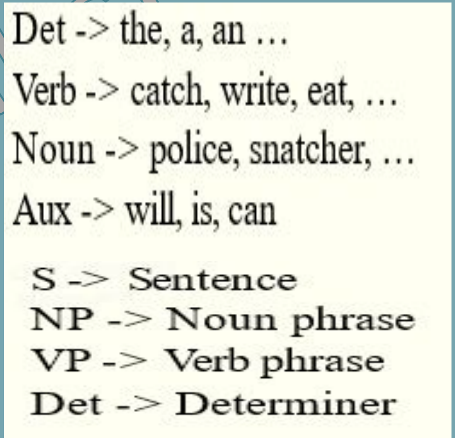


S -> NP + VP

VP - > V + NP

NP -> Det + Noun

V -> Aux + Verb



Det -> the, a, an …

Verb -> catch, write, eat, …

Noun -> police, snatcher, …

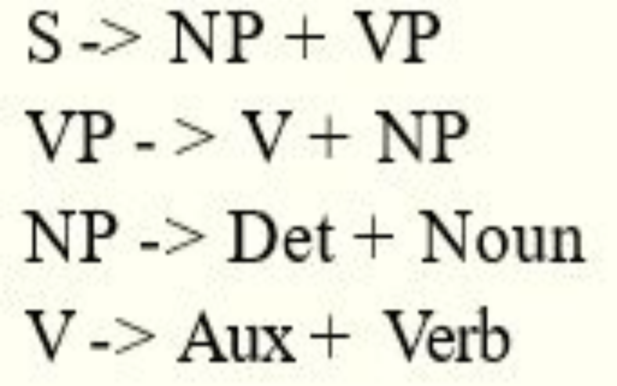
Aux -> will, is, can

S -> Sentence
NP -> Noun phrase
VP -> Verb phrase
Det -> Determiner

Transformational Grammar-Noam Chomsky

Phrase structure grammar:

- Sentences generated using these *rules* are termed *grammatical*.
- Structure assigned by the grammar is a *constituent structure analysis* of the sentence.



S -> NP + VP

VP - > V + NP

NP -> Det + Noun

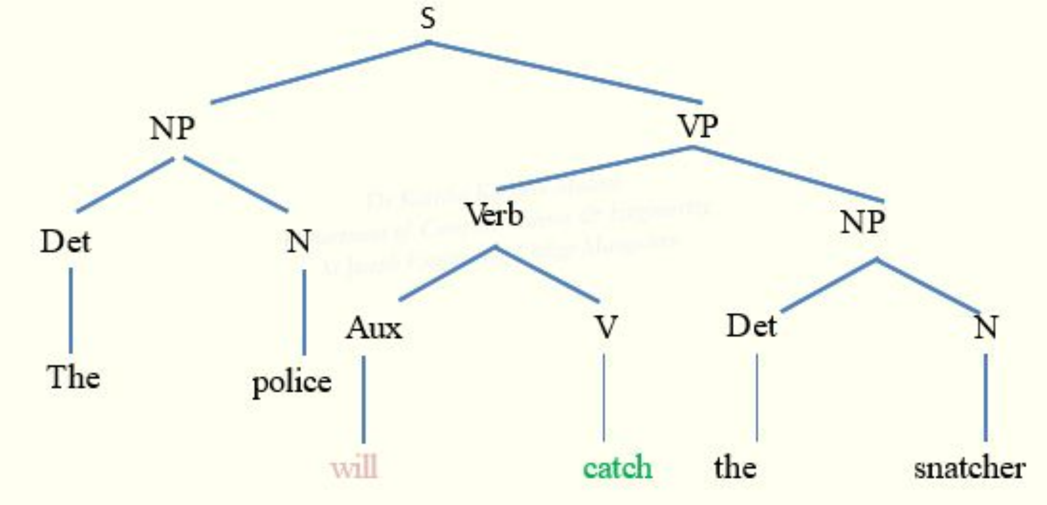
V -> Aux + Verb

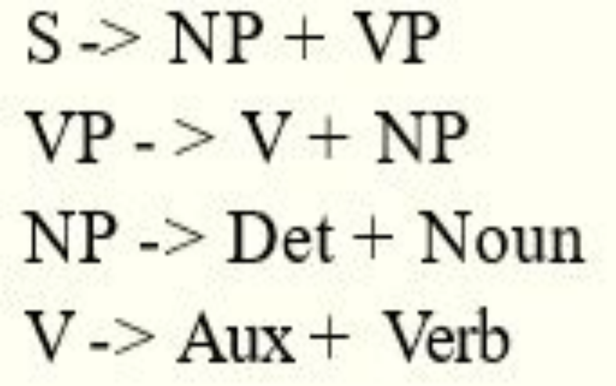
Transformational Grammar-Noam Chomsky

Phrase structure grammar:

Example: *The police will catch the snatcher*

- Parse structure of sentence can be represented as:





S -> NP + VP

VP - > V + NP

NP -> Det + Noun

V -> Aux + Verb

Transformational Grammar-Noam Chomsky

Transformational rules:

- Set of rules, which *transform* one phrase-marker (underlying) into another phrase-marker (derived).

- Heterogeneous unlike phrase structure rules and may have *more than one symbol on their left hand side*

- Rules are used to transform *one surface representation into another.*

  Example: Active sentence into passive sentence

Transformational Grammar-Noam Chomsky

Transformational rules:

- Example: Active sentence into passive sentence

Active sentence: *The police will catch the snatcher*

Passive Sentence: *The snatcher will be caught by the police*

Transformational Grammar-Noam Chomsky

Transformational rules:

- *The police will catch the snatcher*   *The snatcher will be caught by the police*

⇩

$$NP_1 - \text{Aux} - V - NP_2 \longrightarrow NP_2 - \text{Aux} + \textit{be} + \textit{en} - V - \textit{by} + NP_1$$

- Input having the structure NP_1 - Aux – V – NP_2 can be transformed to NP_2 – Aux + – V – *by*+ NP_1

- This Involves: addition of strings *'be'*, *'by'* and *'en'*

Transformational Grammar-Noam Chomsky

Transformational rules:

- The passive transformation rules will convert the sentence

  “The police will catch the snatcher ”

- The + culprit + *will* +*be* + *en+ catch* + *by +* the + police
- Another transformational rule will *re order* of the constituents of a sentence as: en+ catch to catch+en
- *Morphophonemic rules* will convert as: catch+ en to caught

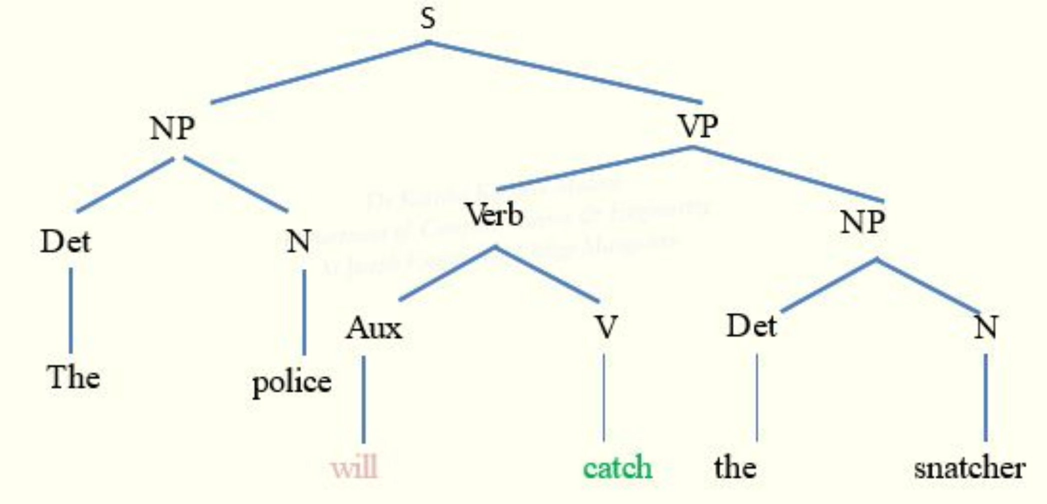
Note: *Here, V is a verb ,which is used in forming the tenses, moods, and voices of other verbs.*
In English be, do, and have - Primary auxiliary verbs;
can, could, may, might, must, shall, should, will and would. - Modal auxiliary verbs

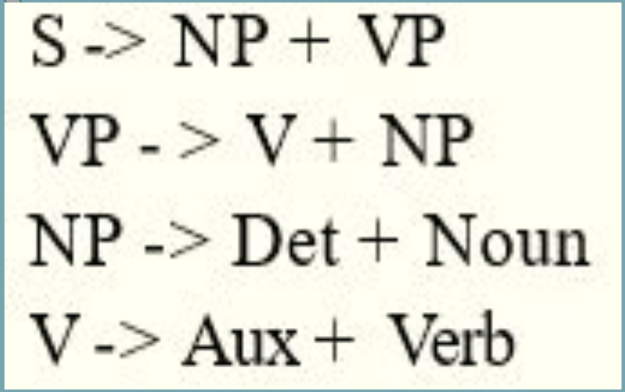
Transformational Grammar-Noam Chomsky

Transformational rules:

The police will catch the snatcher

- Parse structure of sentence can be represented as:





S -> NP + VP

VP - > V + NP

NP -> Det + Noun

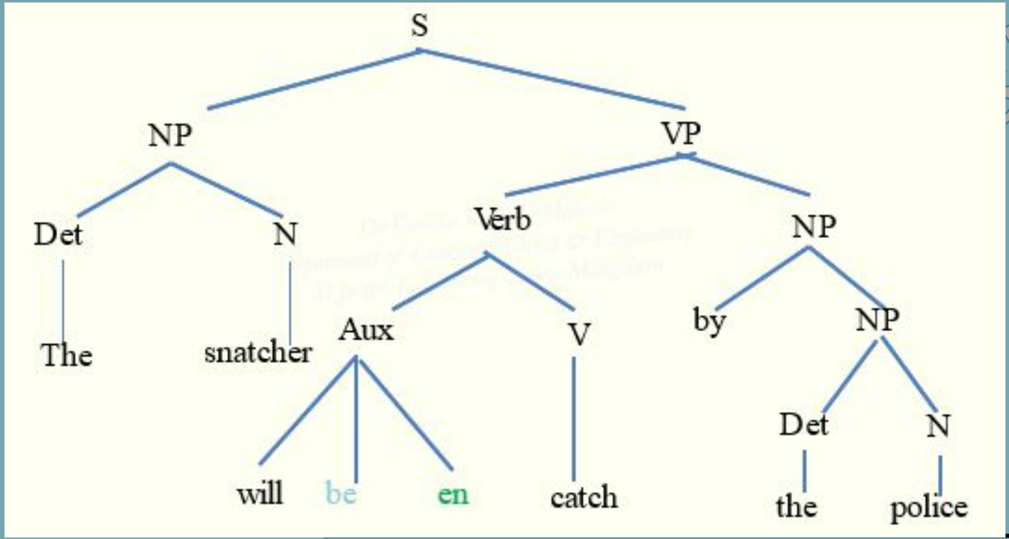
V -> Aux + Verb

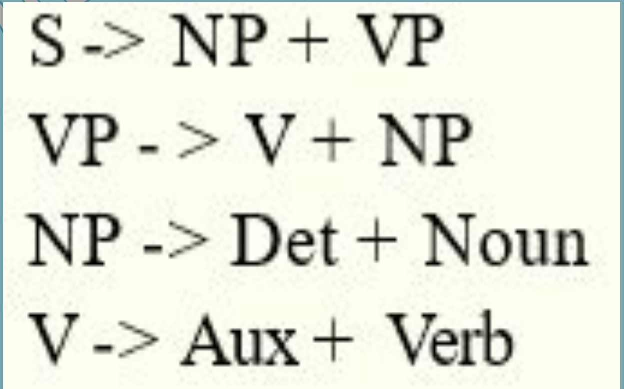
Transformational Grammar-Noam Chomsky

Transformational rules:

The snatcher will be caught by the police

- Parse structure of sentence can be represented as





S -> NP + VP

VP - > V + NP

NP -> Det + Noun

V -> Aux + Verb

Processing Indian Languages:

- There are number of *differences* between *Indian Languages* and *English*. They are:

a) Unlike English, Indic scripts have a non linear structure.

b)Unlike English, Indian Languages are Subject Object Verb as default sentence

Eg: Usne Khaana Khaaya

c)Spelling standardization is more subtle(delicate) in Hindi than English.

d) Indian languages have a relatively rich set of morphological (internal structure of words) variants

Eg: Plural & possessive forms of nouns-computer,computers,computer's)
Comparative & superlative form of adjective-good,better,best)

Processing Indian Languages:

e)Indian languages make extensive & productive use of complex predicates(CP)

Eg: Ram had let sham cut the plant

f) Indian languages use post-position(karakas) case markers instead of prepositions

Eg:Raam could not hitch the cart (English)
Raam *se* gaadi nahi roka saka(Hindi)

g)Indian languages use verb complex consisting of sequence of verbs.

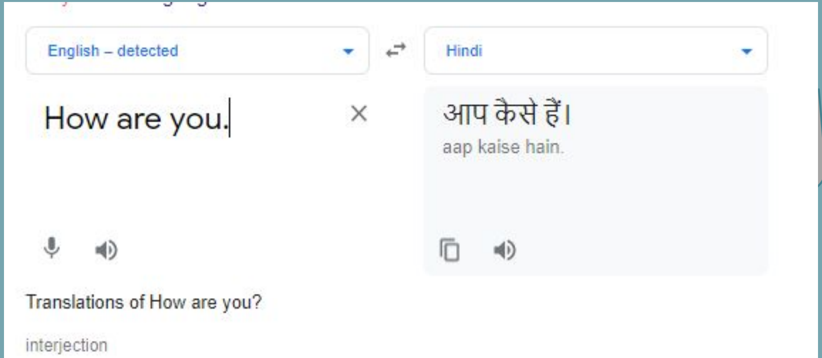
Eg:Ga raha hai - singing
Khel rahi hai-playing

g)Urdu is closely related to Hindi with respect to phonology,morphology,syntax.

NLP Applications

1. Machine Translation

- *Automatic translation* of text from one *human language* to *another*
- This translation requires :
  - Understanding of words, phrases
  - Grammar of two languages
  - Semantics and World knowledge



English – detected

Hindi

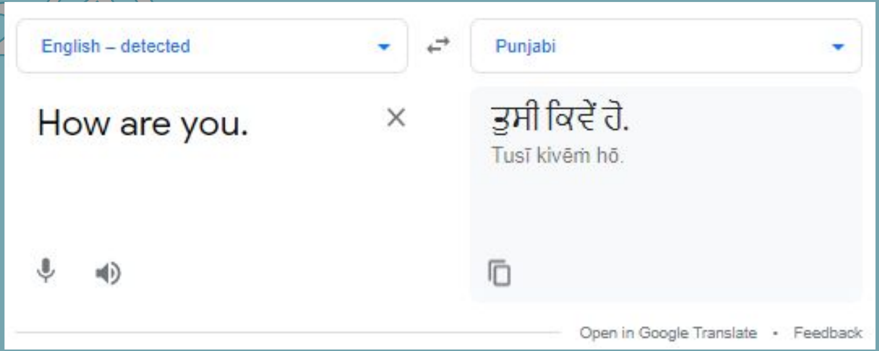
How are you.

आप कैसे हैं।

aap kaise hain.

Translations of How are you?

interjection



English – detected

Punjabi

How are you.

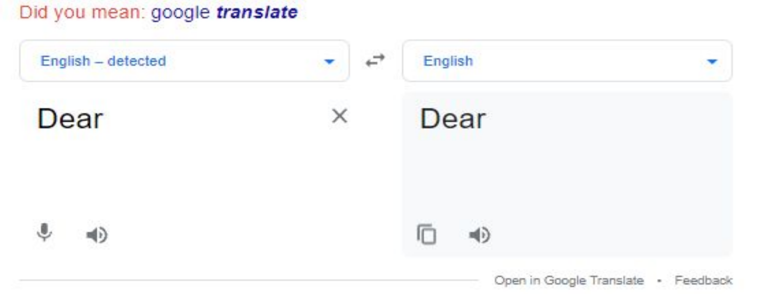
ਤੁਸੀ ਕਿਵੇਂ ਹੋ.

Tusī kivēṁ hō.

NLP Applications

2. Speech Recognition

- *Mapping* acoustic *speech signals* to a *set of words*.It means translation of spoken language to text.
- Difficulties in Recognition is wide variation in pronunciation of words
  - Homonyms -- Same spelling/pronunciation, but different meaning. Eg: deer ,dear
  - Acoustic ambiguities Eg: in the rest , interest



Did you mean: google ***translate***

English – detected ⇄ English

Dear → Dear

NLP Applications

3. Speech Synthesis

- *Automatic production of speech* (utterance of natural language sentences)

- Such systems can
  - Read out your mails on telephone
  - Read out storybook for you.

- Text has to be processed,So NLP is important component of speech synthesis.

NLP Applications

4. Natural Language interfaces to databases

- Querying structured database using Natural language sentences.

- Processes natural language sentences and returns the result to the user

NLP Applications

5. Information Retrieval(IR):

- Identifying documents relevant to user's query
- NLP techniques found useful in:
  - *Indexing*

    -Stop word elimination- Eg: a,the,are,etc

    - Stemming- Eg: eating,eaten,eats=> eat

    - Phrase extraction
  - *Word sense disambiguation -* Identifying which sense of a word is used in a sentence Eg :*Apple* is fruit or organization

NLP Applications

6. Information Extraction (IE):

- *Captures and outputs* factual information within a document.
- Similar to IR systems, in that, it responds to user’s information need
- In Information Retrieval it needs:
    - Information need is expressed as a keyword query
    - System identifies a subset of documents in a large repository of text database
- Information Extraction needs:
    - Information need is specified as pre-defined database schemas or templates
    - Identifies a subset of information within a document that fits the pre-defined templates

Eg: In Library scenario, IE identifies a subset of information within a document it fits.

NLP Applications

7. Question Answering(QA):

- *Given a question* and a *set of documents*, attempts to find the precise answer / precise portion of text in which the answer appears
- In IR system:
  - Returns whole document that seems relevant to user's query
- In QA-Content to be extracted is *unknown*.
- Requires more NLP than IR system or IE system because:
  - Precise analysis of questions and *portions of texts*
  - *Semantic and background knowledge* to answer specific type of questions.

NLP Applications

8. Text Summarization:

- Deals with the *creation of document summaries*
- Involves syntactic,semantic and discourse level processing of text.

Information Retrieval (IR)

- *Information* refers to "*subject matter*" or "*content*" of some text.
- *Retrieval* referees to *accessing information from memory*.
- Information needs to be expressed in the form of query.
- IR Deals with: organization, storage, retrieval & evaluation of information relevant to the query.
- IR incorporates with *different types of information system* such as:

  1. Database management system
  2. Bibliography text retrieval systems
  3. Question answering systems
  4. Search engines

Information Retrieval (IR)

- *Accessing large text collection* can be classified into 2 categories:

1. Construct Topic hierarchy:

   Eg: *Yahoo* -Helps user locate documents of interest manually by traversing the hierarchy .Cost ineffective and inapplicable due to rapid growth of documents on the web

2. Rank the relevant documents according to the relevance.

Information Retrieval (IR)

- **Major *issues* in Information Retrieval (Siddiqui 2006)**

1. Choosing a *representation of the document*:
    - Most human knowledge difficult use in representation
    - Creates problem on keyword representation models when information is *polysemy,homonymy & synonymy*

Note:

- polysemy -lexeme with multiple meaning

  Eg: Head-Part of the body / Person in charge of organization
- homonymy -Ambiguity in which words appear same have unrelated meanings lexeme with multiple meaning

  Eg: Dear/Deer
- synonymy - synonyms-different meanings

  Eg: Give-provide,accord,offer,allow

Information Retrieval (IR)

- **Major *issues* in Information Retrieval (Siddiqui 2006)**

2. Inappropriate characterization of *Queries* by the user.

   Leads to :

   - ○ Lack of Knowledge of the subject / not clear natural language
   - ○ Most human knowledge difficult use in representation

3. Matching of query with respect to the document when there is similarity in document.
4. Evaluation of performance of IR system :
   - ○ Effectiveness of the system.
   - ○ Recall and precision -measures used.

Question Bank- Chapter 1

1.What is NLP?Explain origin of NLP and challenges of NLP-15M

Or What i s NLP?How it has been originated and what are the challenges of NLP

2.List and explain different levels of text analysis

Or Explain semantic and syntactic levels of analysis.-10M

Or what is NLP?List and explain different levels of processing involved in it.

3. Explain the challenges of NLP.-8M

4. Explain in detail about Transformational grammar.-8M

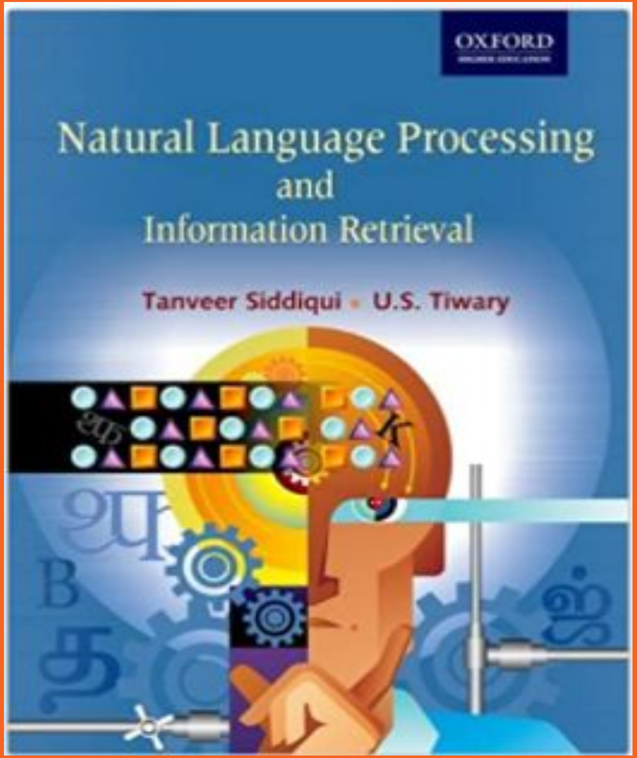
5.Explain in detail applications of NLP. -8M

6.Explain How to process Indian Languages.-5M

7.What is informational retrieval? Write a note on Information Retrieval issues.-5M

Chapter-2

Language Modelling

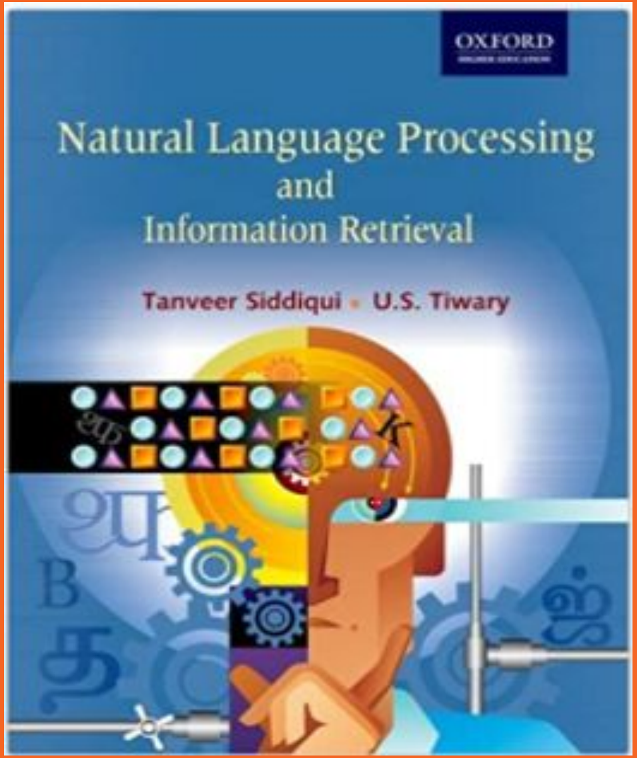


Chapter 2

Language Modelling

○Various Grammar Based Models
Such as:
- Generative grammars
- Hierarchical grammar.
- Paninian Framework
- Karaka Theory

○Statistical Language Model
- n-gram model
- Add-one Smoothing
- Good-Turing Smoothing
- Caching Techniques



Language Modelling

- Model refers to “*entity*” or “*process*”.
- Language model refers to *description of language*.
- Natural language is a complex entity and in order to process it through a computer based programs we need to build a representation (model) of it. It is known as *Language modelling*
- *2 Approaches* for language Modelling are :
  a) Define a grammar that can handle the language-*grammar based language model*
  b)To capture the pattern in a grammar language statistically - *Statistical based language model.*

Grammar Based Language Model

- Uses the *grammar of a language* to create its model.
- Attempts to represent the *syntactic structure* of language.
- Grammar consists of *hand-coded rules* defining the structure and ordering of various constituents appearing in a linguistic unit (phrase, sentence..)

  Eg: Sentence consists of noun phrase and verb phrase.
- Grammar based approach attempts to utilize this structure and relationships between these structures.

Statistical Language Modelling (SLM)

- Statistical Language model is created using a *corpus*, sufficiently large to capture language irregularities.
- SLM is one of the fundamental task in *NLP applications* such as speech recognition, spelling correction, handwriting recognition, Machine Translation.
- Currently has applications in Information Retrieval, Text summarization, question answering etc
- Most popular models are *n-gram models*.

Various Grammar - Based Language Models

- **Various Grammar-based Language Models are:**
  - Transformational Grammar (Chomsky 1957)
  - Generative Grammars
  - Hierarchical Grammar
  - Paninian Grammar

- **Statistical Language Model**
  - n-gram model

Various Grammar - Based Language Models

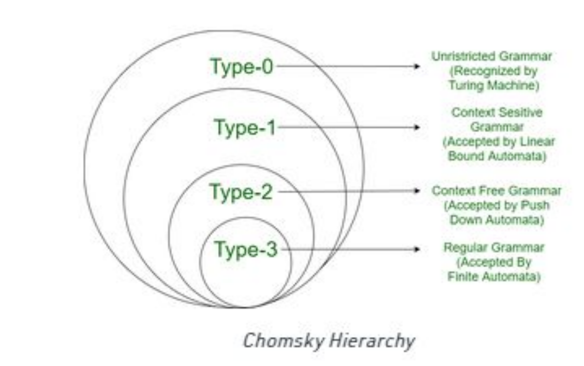
Generative Grammars:

- According to Noam Chomsky(1957), Sentences in a language can be generated, if we know a *collection of words and rules* in that language.
- *Complete set of rules* that can generate possible *sentences* in language.
- This point of view is generative grammar and is used to model a language.
- Language is a relation between *text or sound* and *its meaning*

Various Grammar - Based Language Models

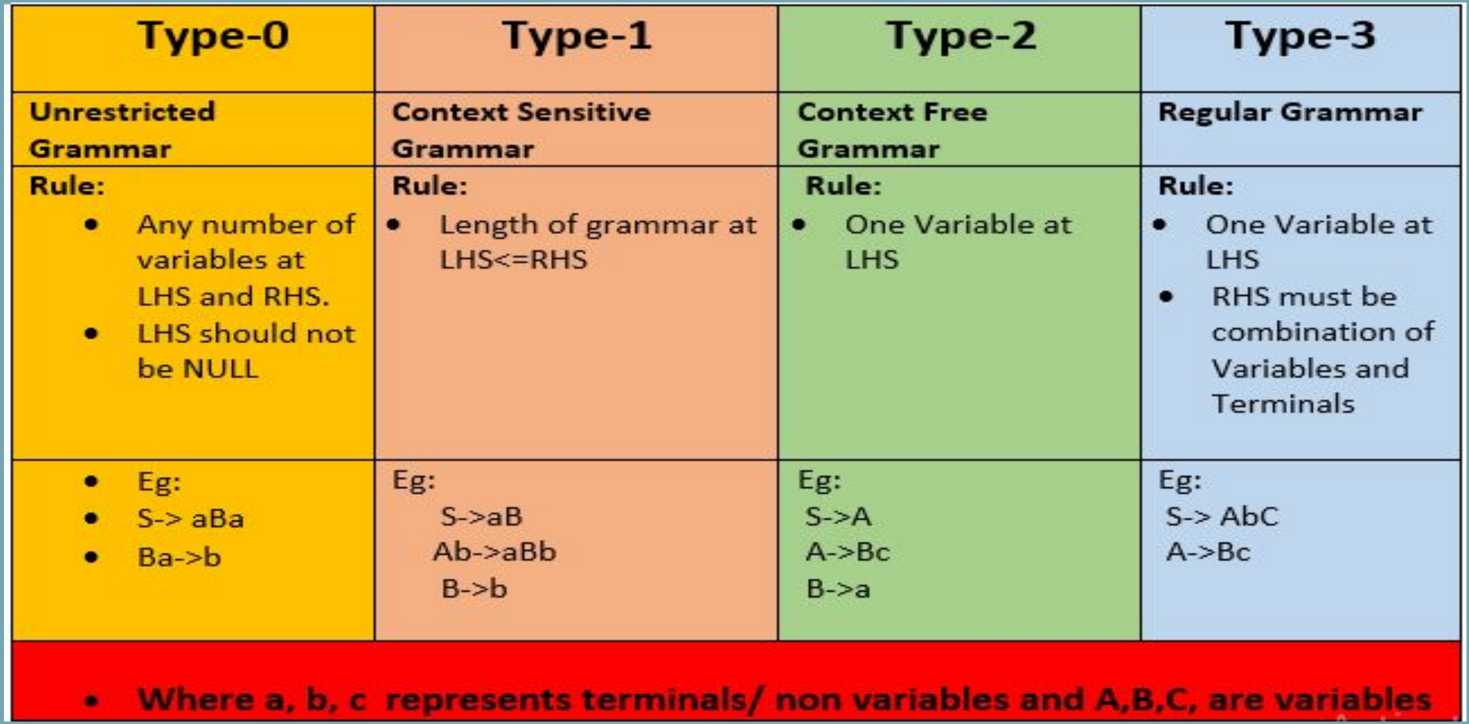
Hierarchical Grammars:

- Includes *classes of grammars* in a hierarchical manner.
- Top layer contains the grammar represented by its *subclasses*
  - Type 0 – Unrestricted grammar (Top most layer)
  - Type 1 – Context-sensitive grammar
  - Type 2 – Context-free grammar
  - Type 3 – Regular grammar


Type-0
Unrestricted Grammar (Recognized by Turing Machine)
Type-1
Context Sesitive Grammar (Accepted by Linear Bound Automata)
Type-2
Context Free Grammar (Accepted by Push Down Automata)
Type-3
Regular Grammar (Accepted By Finite Automata)
Chomsky Hierarchy

Various Grammar - Based Language Models

- **Hierarchical Grammar**



Type-0	Type-1	Type-2	Type-3
Unrestricted Grammar	**Context Sensitive Grammar**	**Context Free Grammar**	**Regular Grammar**
Rule: • Any number of variables at LHS and RHS. • LHS should not be NULL	**Rule:** • Length of grammar at LHS<=RHS	**Rule:** • One Variable at LHS	**Rule:** • One Variable at LHS • RHS must be combination of Variables and Terminals
• Eg: • S-> aBa • Ba->b	Eg: S->aB Ab->aBb B->b	Eg: S->A A->Bc B->a	Eg: S-> AbC A->Bc
• **Where a, b, c represents terminals/ non variables and A,B,C, are variables**			

In English- Nominative,objective,genitive examples

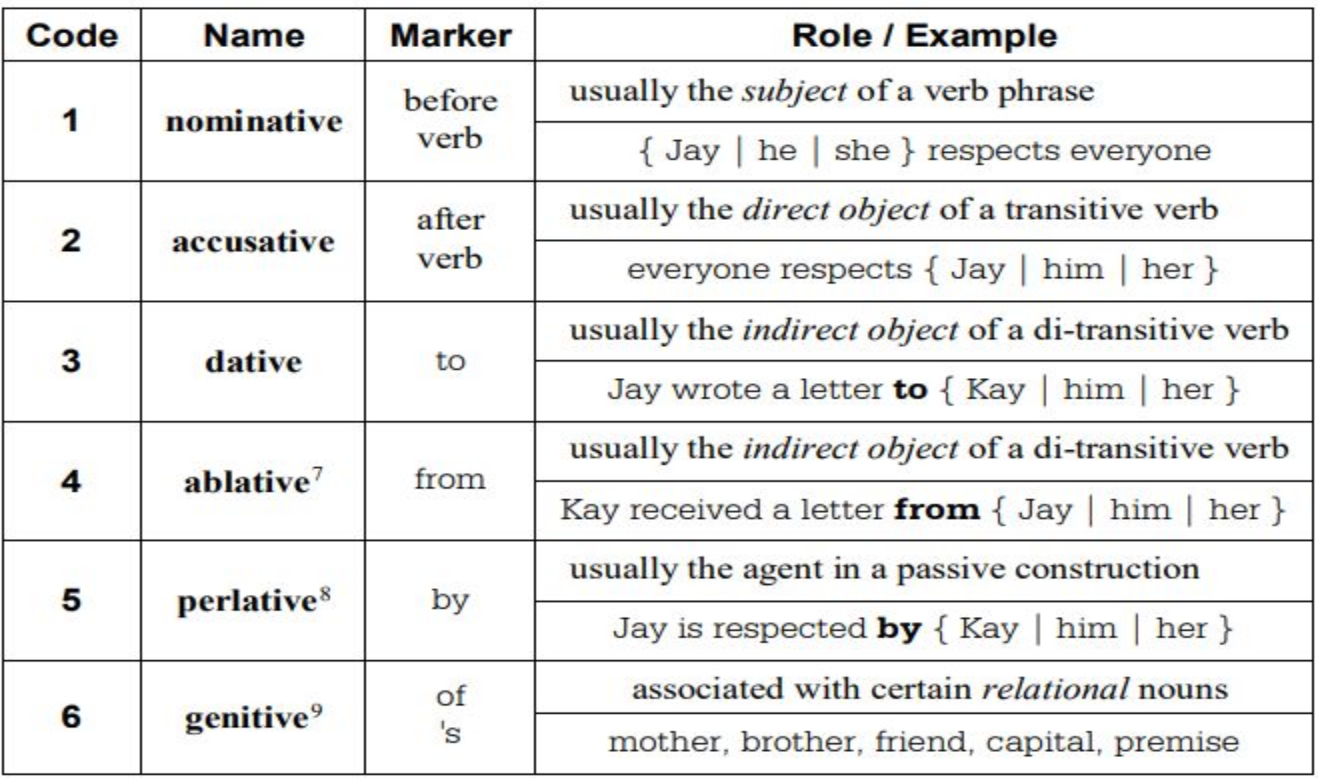
- **Objective Case:**
- When a Noun or a Pronoun is used as the object of a verb it is said to be in the *Objective Case.*
- To find the object in the sentence, put whom or what before the verb and the subject.
- **Example:** *the horse kicked the boy*, the *subject is the horse* and the answer to the question whom did the horse kick? is *the boy*. Hence in the above sentence the ***noun** boy is the object* and it is said to be in *Objective Case.*

In English- Nominative,objective,genitive examples

- **Genitive Case:**
- *When a noun or pronoun shows possession, it is said to be in the Possessive case or genitive case.*
- Let us understand the possessive case with these examples: –The (‘) used to show possession is called an apostrophe.
- **Example:**

1. Shirley's bag is on the table.
2. The dog bit the cat’s tail.

In English- case marker and roles



Code	Name	Marker	Role / Example
1	**nominative**	before verb	usually the *subject* of a verb phrase
			{ Jay | he | she } respects everyone
2	**accusative**	after verb	usually the *direct object* of a transitive verb
			everyone respects { Jay | him | her }
3	**dative**	to	usually the *indirect object* of a di-transitive verb
			Jay wrote a letter **to** { Kay | him | her }
4	**ablative**[7]	from	usually the *indirect object* of a di-transitive verb
			Kay received a letter **from** { Jay | him | her }
5	**perlative**[8]	by	usually the agent in a passive construction
			Jay is respected **by** { Kay | him | her }
6	**genitive**[9]	of 's	associated with certain *relational* nouns
			mother, brother, friend, capital, premise

Some Important Features of Indian Languages

- Indian Languages have traditionally used *oral communication* for knowledge propagation.
- Oral tradition given rise to *morphologically* rich language.
- *Free* word-order.
- Languages like *sanskrit, hindi* has the flexibility to allow word groups representing subject, object and verb group to occur in *any order*.

Some Important Features of Indian Languages

- In others, like Hindi, we can change the position of subject and object
- **Example:**

a) माँ बच्चे को खाना देती है ।

Maan Bachche ko khanaa detti hai

Mother child to food give -(s)

Mother gives food to the child

b) बच्चे को माँ खाना देती है।

Bachche ko Maan khanaa detti hai

Child to mother food give -(s)

Mother gives food to the child

Some Important Features of Indian Languages

- In English, *auxiliary verb follow the main verb*
- In Hindi, they remain as *separate words*
- In South Indian languages (Dravidian) they combine with the main verb.
- **Example:**

खा रहा है ।

khaa raha hai

eat-ing -(s)

eating

करता रहा है।

kartaa rahaa hai

Doing been has -(s)

Has been doing

Some Important Features of Indian Languages

- In Hindi, some *verbs (main)* eg: लेना , देना also combine with *other verbs (main)* to change the aspect and modality of the verbs.
- **Example:**

उसुने खाना खाया।

Usne khanaa khaaya

He (Subj) food ate

He ate food

वह चला।

He moved

उसुने खाना खा लिया।

Usne khaanaa kha liyaa

He (Subj) food eat taken

He ate food (completed action)

वह चला दिया।

He move given

He moved (started the action)

Some Important Features of Indian Languages

- In Indian languages, the *nouns* are followed by *post- positions* instead of prepositions.
- Generally remain as *separate words* in *Hindi*, except in case of pronouns.

- **Example:**

रेखा के पिता

Rekha ke pita

Rekha of father

Father of Rekha

उसके पिता (pronoun)

Her (His) father

Some Important Features of Indian Languages

- Among Indian languages all *features* are *not the same*.
- As seen before, Verb groups are formed *differently* in Indo-Aryan and Dravidian languages
- Sanskrit is different from other Indian languages
- Has 5 tenses and 3 numbers(singular,plural,dual) and one time-aspect in each tense.

  Example: Translation of He goes and He is going are same in sanskrit.

Some Important Features of Indian Languages

- Hindi is unique- *No Neuter gender*
- All nouns are categorized as feminine (-gender female Eg:woman, hen)or masculine(gender male Eg:man, rooster).
- Neuter gender refers to (-no gender Eg: doctor,chicken)
- Verb form must have a gender agreement with the subject (sometimes the object)
- **Example:**

ताला खो गया ।
Taala *kho gayaa*
Lock lose (past)
The lock was lost

चाभी खो गई ।
Chaabhii *kho gayee*
Key lose (past)
The key was lost

Paninian Framework

- Paninian Grammar (PG) written by *Panini* in 500 BC in Sanskrit (Asthadhyayi).
- This framework can be used for *other Indian languages* and some Asian language as well.
- Asian languages are SOV(Subject-Object-Verb) rich ordered than English.
- Takes advantage of rich inflections that provide syntactic and semantic cues for language analysis and understanding.

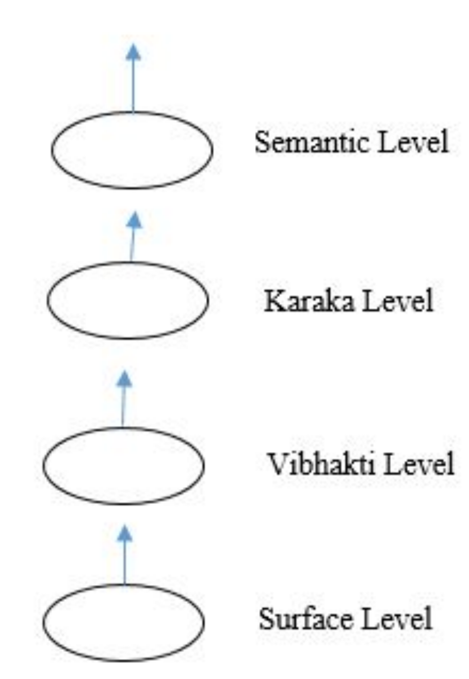
Layered Representation in PG

Paninian Grammar

- *Syntactico-semantic* – one can go from surface layer to deep semantics by passing through *intermediate layers*
- Layers can be represented as:
  - Semantic level
  - Karaka Level
  - Vibhakti level
  - Surface level

Layered Representation in PG

- ***Surface level*** is a written form of sentence
- ***Vibhakti level***
- Literally refers to inflection
- Here, Vibhakti (word level) refers to word (noun,verb or other) groups based on case endings/post positions/compound verbs (eg: follow up, get down) /main and auxiliary verbs.
- Word groups rely on *various kinds of markers* which are language specific. All *Indian languages* can be represented at Vibhakti level



Layered Representation in PG

- ***Karaka level (pronounced as Kaaraka)***
- Literally means *case*(is a grammatical function of noun/pronoun).
- Karaka relations are based on the *way the word groups participate in the activity* denoted by the *verb group*.
- *Complexities* arise in this level due to absence of inflections, multiple categories of words,multiple meanings and large number of exceptions .
- Implementation difficult.

Layered Representation in PG

- ***Semantic level***
- It represents what the speaker has in mind.
- Purpose of language is to *communicate* between one human and another.
- Resolution of ambiguities is left to the listener.
- *Number of semantic levels are not particular.*
- Here, multiple-meaning texts are abundant(plenty) in Indian literature as seen in hundreds of interpretations of the epics.

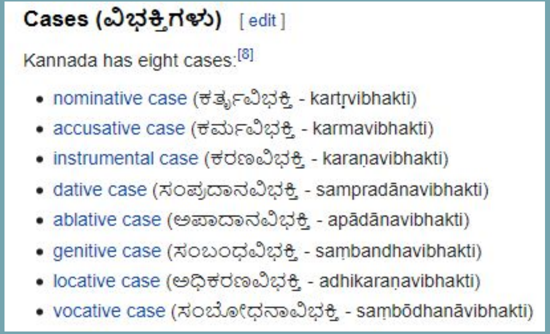
Karaka Theory

- Central theme of PG framework.
- Karaka relations are assigned *based on roles* played by various *participants in the main activity.*
- Roles are reflected in the case markers and post-position markers.
- Karaka relations are similar to the *case relations* in English
- But Defined in a different manner.
- Takes advantage of richness of the *case endings* found in Indian languages.

Karaka Theory

- Various karakas:
  - Karta (Subject)
  - Karma (Object)
  - Karana (Instrument)
  - Sampradan (Beneficiary)
  - Apaadaan (Separation)
  - Adhikaran (Locus)

For Reference: Cases in Kannada Grammar

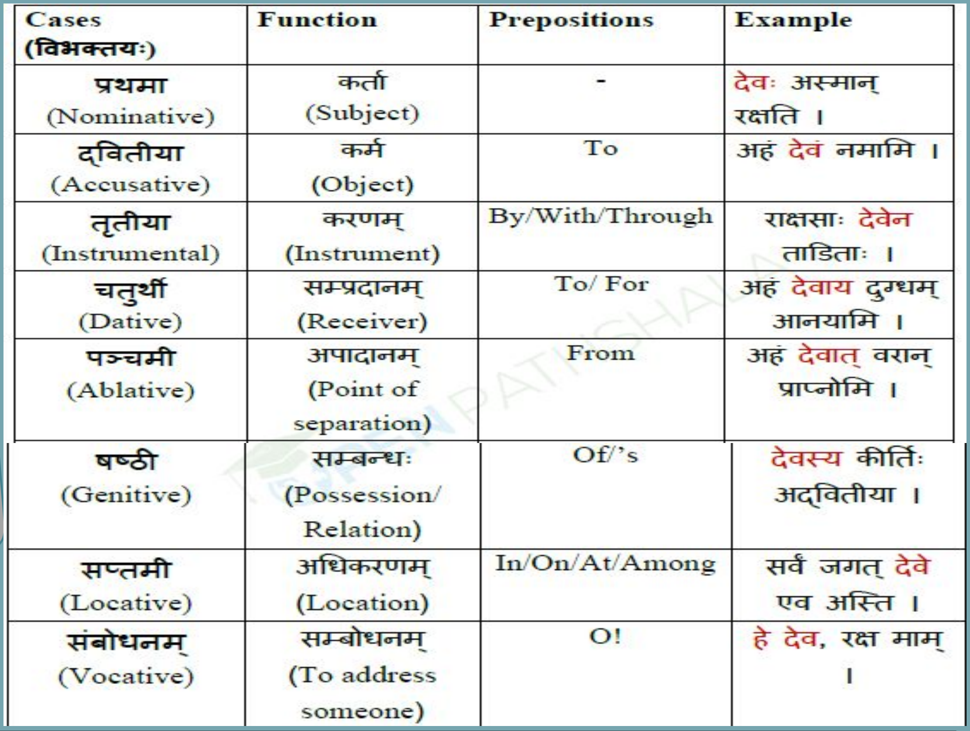


Cases (ವಿಭಕ್ತಿಗಳು) [edit]

Kannada has eight cases:[8]

- nominative case (ಕರ್ತೃವಿಭಕ್ತಿ - kartṛvibhakti)
- accusative case (ಕರ್ಮವಿಭಕ್ತಿ - karmavibhakti)
- instrumental case (ಕರಣವಿಭಕ್ತಿ - karaṇavibhakti)
- dative case (ಸಂಪ್ರದಾನವಿಭಕ್ತಿ - sampradānavibhakti)
- ablative case (ಅಪಾದಾನವಿಭಕ್ತಿ - apādānavibhakti)
- genitive case (ಸಂಬಂಧವಿಭಕ್ತಿ - saṃbandhavibhakti)
- locative case (ಅಧಿಕರಣವಿಭಕ್ತಿ - adhikaraṇavibhakti)
- vocative case (ಸಂಬೋಧನಾವಿಭಕ್ತಿ - saṃbōdhanāvibhakti)

For reference: Cases in English and Hindi Grammar



Cases (विभक्तयः)	Function	Prepositions	Example
प्रथमा (Nominative)	कर्ता (Subject)	-	देवः अस्मान् रक्षति ।
द्वितीया (Accusative)	कर्म (Object)	To	अहं देवं नमामि ।
तृतीया (Instrumental)	करणम् (Instrument)	By/With/Through	राक्षसाः देवेन ताडिताः ।
चतुर्थी (Dative)	सम्प्रदानम् (Receiver)	To/ For	अहं देवाय दुग्धम् आनयामि ।
पञ्चमी (Ablative)	अपादानम् (Point of separation)	From	अहं देवात् वरान् प्राप्नोमि ।
षष्ठी (Genitive)	सम्बन्धः (Possession/ Relation)	Of/'s	देवस्य कीर्तिः अद्वितीया ।
सप्तमी (Locative)	अधिकरणम् (Location)	In/On/At/Among	सर्वं जगत् देवे एव अस्ति ।
संबोधनम् (Vocative)	सम्बोधनम् (To address someone)	O!	हे देव, रक्ष माम् ।

Karaka Theory

- Let us consider an example of various Karaka relations in a sentence:

माँ बच्ची को आंगन में हाथ से रोटी खिलाती है।

Maan Bachchi ko aangan mein haath se rotti khilathi hei

Mother child-to courtyard-in hand-by bread feed -(s)

The mother feeds bread to the child by hand in the courtyard.

Karaka Theory-Karta

माँ बच्ची को आंगन में हाथ से रोटी खिलाती है।

Maan Bachchi ko aangan mein haath se rotti khilathi hei

Mother child-to courtyard-in hand-by bread feed -(s)

The mother feeds bread to the child by hand in the courtyard.

- The first important Karaka is *Subject* called *'Karta'* in PG.
- Karta is defined as the *noun group* - most independent in Hindi
- Has *'ne'* or *'ϕ'* case marker.
- *Maan (mother)* is the Karta.
- Karta is an independent entity in the activity denoted by the main verb

Karaka Theory-Karta versus Agent role

- The Concept of Karta is different from the *'agent'* concept in the sense that *Karta can also take up the role of experiencer.*
- **Example:**

मुजुसे रहा न गया।

Mujhse rahaa na gayaa

Me hold-not -passive

I could not hold by myself.

Karaka Theory-Karma

माँ बच्ची को आंगन में हाथ से *रोटी* खिलाती है।
Maan Bachchi ko aangan mein haath se *rotti* khilathi hei
Mother child-to courtyard-in hand-by *bread* feed -(s)
The mother feeds *bread* to the child by hand in the courtyard.

- Similar to *object.*
- It is the *locus(adhikaran)* of the *result of the activity*.
- *Rotii (bread)* is the Karma.
- When *Karta* is the experiencer, it (she) is also the *locus of the result.*
- Generally has *‘ϕ’* or *‘KO’* case marker.

Karaka Theory-Karan

मॉँ बच्ची को आंगन में *हाथ* से रोटी खिलाती है।

Maan Bachchi ko aangan mein *haath se* rotti khilathi hei

Mother child-to courtyard-in *hand-by* bread feed -(s)

The mother feeds bread to the child *by hand* in the courtyard.

- Another Karaka relation is *'Karan' (instrument)*, essential for action to take place.
- *Is a noun group* through which the goal is achieved.
- *Haath (hand)* is the Karan.
- Has the case marker *'dwara' (by)* or *'se'.*

Karaka Theory-Sampradan

माँ *बच्ची को* आंगन में हाथ से रोटी खिलाती है।
Maan *Bachchi ko* aangan mein haath se rotti khilathi hei
Mother *child-to* courtyard-in hand-by bread feed -(s)
The mother feeds bread *to* the *child* by hand in the courtyard.

- *'Sampradan'* is the *beneficiary/recipient* of the activity.
- *bachchi (child)* is the Sampradan.
- Takes the case marker *'ko' (to)* or *'ke liye' (for).*

Karaka Theory-Apaadaan

माँ ने *थाली* से खाना उठाकर बच्चों को दिया।

Maan ne *thaali* se khana uthakar bachche ko diyaa

Mother-karta *plate from* Apaadan food taking -up child-to gave.

The mother gave food to the child taking it up *from* the *plate.*

- ‘Apaadaan’ denotes *source of activity*.
- A *noun* denoting the *point of separation* for *a verb expressing an activity* which involves movement away from is apaadaan.
- The *marker* is attached to the *part that serves as a reference point* (being stationary).
- *‘Thaali’* is the Apaadaan.

Karaka Theory-Adhikaran

माँ बच्ची को *आंगन* में हाथ से रोटी खिलाती है।

Maan Bachchi ko *aangan* mein haath se rotti khilathi hai

Mother child-to *courtyard*-in hand-by bread feed -(s)

The mother feeds bread to the child by hand in the *courtyard*.

- *'Adhikaran'* denotes the *locus (support in space or time)* of Karta or Karma.
- *'aangan'* (courtyard) is the *Adhikaran.*
- **Note:** All the six relations are not sufficient to capture all possible relations.It also needs 'Sambandh' (relation) and 'Tadarthya' (purpose) have also been used.

Issues in Paninian Grammar (PG)

- 2 problems challenging linguists are:
  - *Computational implementation* of PG
  - *Adaptation of PG to Indian*, and *other similar languages*.

- Multi layered implementation - rules arranged in multiple layers Different-karaka chart rules
- *Mapping* of Vibhakti(case markers and post positions) and semantic relation (w.r.to verb) is not one to one.
- That is, *2 vibhakti* can represent the *same* relation or the same vibhakti can represent different relations in different contexts.

Example:Accusative and dative case uses ***to*** as prepositions (From karaka chart)

For Your Information:Before Statistical Language Modelling(SLM)

- Language modelling uses various statistical and probabilistic techniques to determine the *probability of a given sequence of words* occurring in a *sentence*
- In any sentences word should be arranged in order.
- Language model has 2 probabilities:joint probability and *conditional probability*
- N-gram uses conditional probability
- For Example: Sentence= She is dead
- Conditional probability using chain rule,For 3 words:

$$P(x, y, z) = P(x)P(y/x)P(z/x, y)$$
$$P(she, is, dead) = P(she)P(is/she)P(dead/she, is)$$

For Your Information:Before Statistical Language Modelling(SLM)

- An N-gram means a *sequence of N words*.
- In previous example, “P(she) -unigram” “P(is/she)” is a 2-gram (a bigram), “ P(dead/she,is) ” is a tri-gram.
- NLP includes n-grams in variety of applications.
- Examples are: auto completion of sentences (such as the one we see in Gmail these days), auto spell check and to a certain extent, we can check for grammar in a given sentence.
- Example: “Thank you so much for your ”. Now we know that the next word is “help” with a very high probability.
- But how will the system know that?

For Your Information:Before Statistical Language Modelling(SLM)

- Train the model with a huge *corpus of data*.
- NLP model will find “probability” of the occurrence of a word after a certain word.
- Improve the predictions of auto completion systems.
- We can use NLP and n-grams to train voice-based personal assistant bots.
- For example, a bot will be able to understand the difference between sentences “what’s the temperature?” and “set the temperature.”

For Your Information:Before Statistical Language Modelling(SLM)

- In Markov assumption it says:
- For Example: Sentence= Good to learn NLP

  P(NLP/Good to learn) ~ P(NLP/learn)~ P(NLP/ to learn)

- Here, if number of word increases it leads to complex to represent
- We have N-gram model.
- Where n- represents :1,2,3,4……..n

Statistical Language Model (SLM)

- Probability distribution, P(s) over all possible word sequences (or words, sentences, paragraphs,documents or spoken utterance).

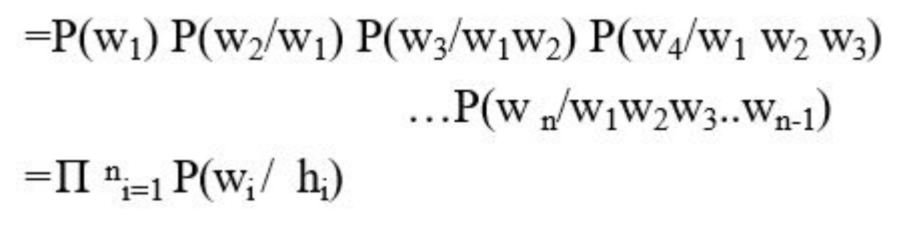
- Dominant approach in statistical language modelling is the ***n-gram model.***

- *Goal of n-gram model is:* Estimate the *probability (likelihood) of a sentence.*

n-gram Model

- Probability estimation in n-gram is achieved by Decompose sentence probability into a *product of conditional probabilities* using the *chain rule* as below:

$$P(s) = P(w1, w2, w3, \ldots\ldots wn)$$



$$=P(w_1)\,P(w_2/w_1)\,P(w_3/w_1w_2)\,P(w_4/w_1\ w_2\ w_3)\ \ldots P(w_{\,n}/w_1w_2w_3..w_{n-1})$$

$$=\Pi\ {}^{n}{}_{i=1}\,P(w_i/\ h_i)$$

- Where h$_i$ history of word w$_i$ defined as: w$_1$ w$_2$ .... w$_{i-1}$

n-gram Model

- To calculate *sentence probability*, need to calculate the probability of a word, given the sequence of words preceding it.
- ***n-gram model*** simplifies this task by *approximating* the probability of a word given all the previous words by the *conditional probability given previous n-1 words only.*

$$P(w_i / h_i) \approx P(w_i / w_{i-n+1} \ldots w_{i-1})$$

- Thus, n-gram model calculates probability by looking at the previous n-1 words only

n-gram Model

- A model that limits the history to the previous one word only is termed as :**Bi-gram (n=1) model**
- Using Bi-gram the *probability* of a sentence can be calculated as :

$$P(s) \approx \Pi\ ^{n}{}_{i=1}\ P(w_i / w_{i-1})$$

- A model that conditions the probability of a word to the previous two words, is termed as :**tri-gram (n=2) model**
- Using tri-gram the *probability* of a sentence can be calculated as :

$$P(s) \approx \Pi\ ^{n}{}_{i=1}\ P(w_i / w_{i-2} .w_{i-1})$$

n-gram Model

- **Example:**
- Consider the sentence :

  The Arabian knights are the fairy tales of the east

- The bi-gram approximation of:

  P(*east/The Arabian knights are the fairy tales of the*)
  is P(*east/the*)

- The Trigram approximation of:

  P(*east/The Arabian knights are the fairy tales of the*)
  is P(*east/of the*)

n-gram Model

- *N-gram model adds <s>* - A special word (pseudo word) introduced to mark the *beginning* of the sentence in *bi-gram estimation.*
- Probability of *first word* is conditioned on *<s>.*
- In *trigram estimation, 2 pseudo words* are introduced *<s1>* and *<s2>*

n-gram Model

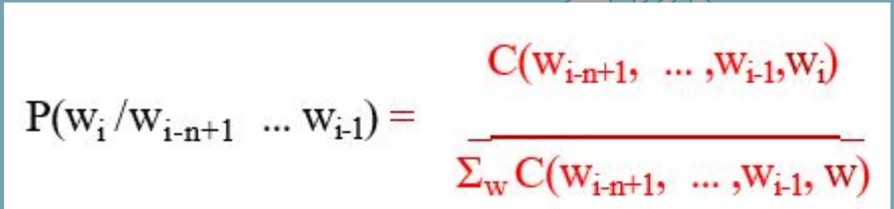
- *Estimating* the *probabilities* is done by training the n-gram model on the training corpus

$$P(w_i / h_i) \approx P(w_i / w_{i-n+1} \ldots w_{i-1})$$

- Using Maximum Likelihood Estimation (MLE), *count* a *particular n-gram* in the training corpus and *divide it* by the *sum of all n-grams* that *share the same prefix.*

n-gram Model

- So, consider the probability of word w_i with respect to its previous words can be denoted as : $P(w_i /w_{i-n+1} \ ... \ w_{i-1})$
- To calculate probability, count the words as below:
- *Count* a particular *n-gram in the training corpus* and divide it by the *sum of all n-grams that share the same prefix*.

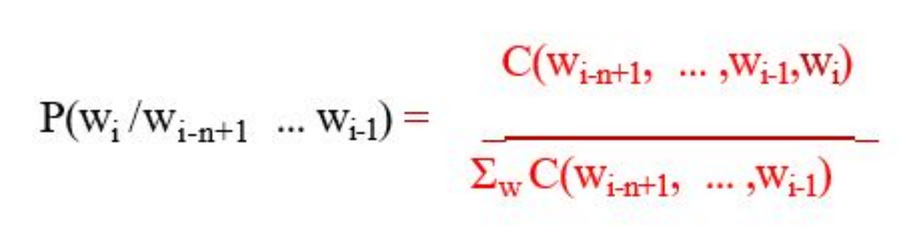


$$P(w_i / w_{i-n+1} \ ... \ w_{i-1}) = \frac{C(w_{i-n+1}, \ ... ,w_{i-1},w_i)}{\Sigma_w C(w_{i-n+1}, \ ... ,w_{i-1}, w)}$$

Note: w represent word in the corpus

n-gram Model

- Here, denominator *Sum of all n-grams that share first n-1 words* is equal to the *count of the common prefix* $w_{i-n+1}, \ldots, w_{i-1}$
- To calculate the probability of word can be written as below:



$$P(w_i / w_{i-n+1} \ldots w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{\Sigma_w C(w_{i-n+1}, \ldots, w_{i-1})}$$

- The model take Training set (T)data and calculate probability for each word based on bigram /trigram/n-gram. Then probability for Test sentence (s) will be calculated.

n-gram Model

- **Example:**
- *Training Set (T):*

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 given *Test sentence(s)* using the bi-gram model:

The Arabian knights are the fairy tales of the east.

n-gram Model

- Initially, for bi gram model, *Training Set (T)* should be considered with special word <s>:

<s>The Arabian Knights

<s>These are the fairy tales of the east

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

n-gram Model

- Find the probability of the Test sentence (s) :

Test sentence (s):
<s>The Arabian knights are the fairy tales of the east.

- The probability of Test sentence (s) can be calculated using bigram model:

P(s) = P(The/<s>) X P(Arabian/the) X P(knights/Arabian) X P(are/knights) X P(the/are) X P(fairy/the) X P(tales/fairy) X P(of/tales) X P(the/of) X P(east/the)

n-gram Model

- For bi gram model, *Training Set (T)* :

<s>The Arabian Knights

<s>These are the fairy tales of the east

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

n-gram Model

Test sentence (s): The Arabian knights are the fairy tales of the east.

P(s) = P(The/<s>) X P(Arabian/the) X P(knights/Arabian) X P(are/knights) X P(the/are) X P(fairy/the) X P(tales/fairy) X P(of/tales) X P(the/of) X P(east/the)

P(*the* /<s>) =2/3=0.67
P(*Arabian* / *the*) = 2/5 = 0.4
P(*knights* / *Arabian*) =2/2 =1
P(*are* /*knights*) = ½ =0.5
P(*the* /*are*) = ½ = 0.5
P(*fairy* / *the*) = 1/5 =0.2
P(*tales* / *fairy*) =1/1 =1

P(*of* / *tales*) =1/1=1
P(*the* / *of*) =2/2 =1
P(*east* / *the*) = 1/5 =0.2

Note:
To compute probability from the corpus:
P(*the* /<s>)= C(<s>/the)/C(<s>)=⅔=0.67
P(*Arabian* /*the*)=C(the/Arabian)/C(the)=2/5=0.4
P(*are* /*knights*)=C(knights/are)/C(knights)=½=0.5
P(*of* /*tales*)=C(tales/of)/C(tales)=1/1=1
P(*are* /*knights*)=C(knights/are)/C(knights)=½=0.5

P(s) =0.67 * 0.4* 1 *0.5 * 0.5*0.2 * 1 * 1 * 1 * 0.2=0.00268

n-gram Model

- Each probability must less than 1
- Multiplying probabilities might cause numerical underflow (in long sentences).
- To avoid this, calculations should be made in *log space.*
- Where, estimate the probability of a sentence by *adding log of individual* probabilities and take *antilog* of the sum.

Problem of n-gram Model

- ***Data sparseness problem :*** n-gram that *does not occur* in the training data is assigned *zero probability*.
- For the Training Set:

  <s>The Arabian Knights

  <s>These are the fairy tales of the east

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

- If the Test sentence , s: Arabian knights

  P(s) = P(Arabian/<s>) X P(knights/Arabian)

  P(s)= (0/3) * (2/2)

  P(s)=0* 1=0

Problem of n-gram Model

- ***Data sparseness problem:*** n-gram that *does not occur* in the training data is assigned *zero probability.*
- Assumption: *Probability of occurrence of a word depends only on the preceding word (or preceding n-1 words),* which is not necessarily true.
- There are several *long distance dependencies* in Natural Language sentences.
- The n-gram model *fails* to capture this.

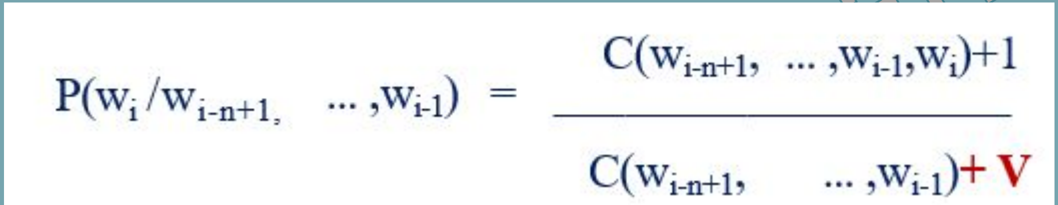
Note: Example : Long-distance dependencies:the man that I saw yesterday after lunch went fishing (the man _ went fishing)

Handling problem of n-gram Model

- ***Handling Data Sparseness*** **by** ***Smoothing Techniques.***
- The smoothing technique " Task of *re-evaluating zero probability or low-probability n-grams* and *assigning* them *non-zero values*"
- Smoothing makes the distributions more uniform by moving the extreme probabilities towards the average
- Smoothing can be done by:
  - Add-one Smoothing
  - Good-Turing Smoothing
  - Caching Technique

Smoothing Techniques for n-gram model

- **Add-one smoothing:**
- *Adds* a value of *one* to *each n-gram frequency* before normalizing them into probabilities.



$$P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)+1}{C(w_{i-n+1}, \ldots, w_{i-1})+V}$$

- Here, ***V***- *vocabulary size* is i.e., size of the set of all the words being considered.

Smoothing Techniques for n-gram model

- **Add-one smoothing:**
- *Adds* a value of *one* to *each n-gram frequency* before normalizing them into probabilities.
- For the Training Set:

  <s>The Arabian Knights

  <s>These are the fairy tales of the east

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

- If the Test sentence , s: Arabian knights
- By applying Add-0ne smoothing

  P(s) = P(Arabian/<s>) X P(knights/Arabian)

  P(s)= (0+1/3+ 14) * (2+1/2+ 14)

  P(s)=(1/17)*(3/16)=0.011029

Smoothing Techniques for n-gram model

- **Add-one smoothing:**
- Not a good smoothing technique
- *Assigns same probability* to all *missing n-grams*, even though some of them could be more intuitively appealing.
- *Variance* of counts produced by the add-one smoothing is *worse* than the unsmoothed MLE methods (Gale and Church, 1994).
- *It Shifts* too much of *probability mass* towards the *unseen n-grams* (n-grams with 0 probabilities) as number is quite large.

Smoothing Techniques for n-gram model

- **Good-Turing smoothing (Good 1953):**
- It attempts to *improve* the situation by looking at the *number of n-grams* with a *high frequency* in order to estimate the *probability mass.*
- That needs to be *assigned to missing or low frequency n- grams*

Smoothing Techniques for n-gram model

- **Good-Turing smoothing (Good 1953):**
- *Adjust* the *frequency* ***f*** *of an n-gram* using the *count of n-grams* having a frequency of occurrence ***f+1***.
- *Converts* the frequency of an *n-grams* from *f* to f^* using the following expression:

$$f^* = (f+1)\, n_{f+1} / n_f$$

- n_f - *number of n-grams* occurring exactly *f times* in the training corpus.

Smoothing Techniques for n-gram model

- **Good-Turing smoothing (Good 1953):**
- Let,**For Example :** In corpus,if the highest frequencies are:

  Number of n-grams occurring 4 times – 25,108

  Number of n-grams occurring 5 times - 20,542

- Then the smoothed count for 5 will be:

$$f^{*} = (f+1)\, n_{f+1} / n_f$$

$$f^{*} = 20542 * 5 / 25108 = 4.09$$

Note: This says the probability mass assigned to missing or low-frequency n-grams

Smoothing Techniques for n-gram model

- **Caching Technique:**
- Another improvement over n-gram model is *caching*
- *Frequency* of n-gram is *not uniform* across the text segments or corpus.
- For example: In this section , the frequency of word **‘n-gram’** is high but rare in other section
- But Basic n-gram model *ignores* this sort of variation *in n-gram frequency*.

Smoothing Techniques for n-gram model

- **Caching Technique:**
- Cache model *combines* the most recent n-gram frequency with the standard n-gram model to *improve* the *performance* locally.
- The underlying *assumption* here is that: *Recently discovered words are more likely to be repeated*.

Question Bank- Chapter 2

1.Difference between grammar based model and statistical based model-5M

2.Write a note on generative grammars and hierarchical grammar-8M

3.Explain layered representation of Paninian Grammar (PG)-6M

4. Explain karaka theory of Paninian grammar. Identify different Karaka's in the following sentence in Hindi language: Maan Bachchi ko aangan mein haath se roti khilathi hain' -8M

5..Explain n-gram modelling of natural languages.Find the probability of the test sentence S2 in the following training set:-8M

S1: The Arabian knights

S2: These are the fairy tales of the east

S3:The stories of the Arabian Knights are translated in many languages.

6.Determine the probability for training set corpus, predict the test case sentences probability and choose the highest probability sentence among two by the use of bigram model -10M

Training Set:

<s> I am Chintu</s>
<s> I like College </s>
<s>Do Chintu like College </s>
<s>Chintu I am </s>
<s>Do I like Chintu </s>
<s>Do I like College </s>
<I do like Chintu </s>

Question Bank- Chapter 2

Testing Set:

I like College.

Do I like Chintu

7.Determine the probability of words and sentence using add one smoothing.

Training Set:

I love India

India is my country

I feel proud

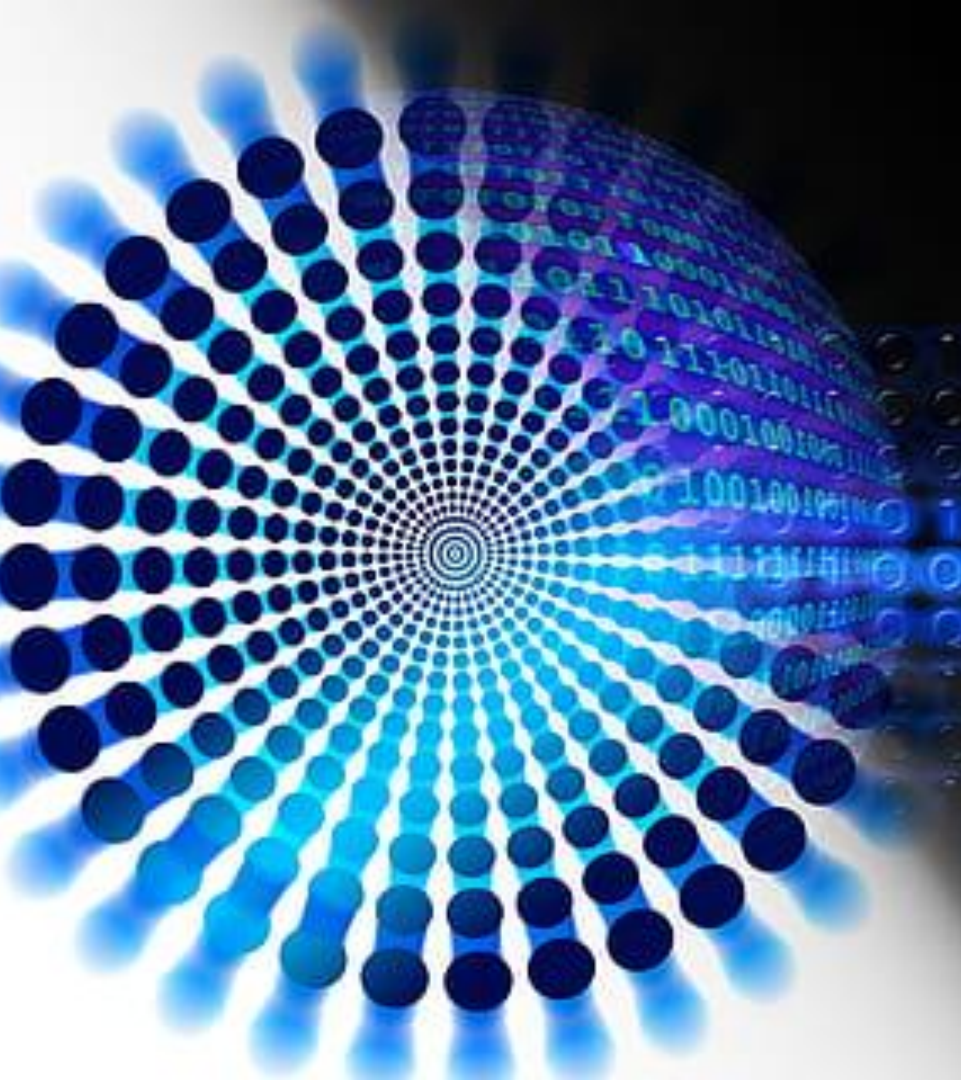
Testing Set:

I love india and feel proud

my country

8.What is statistical language model and explain features of n-gram model-6M

9.Explain statistical language model-10M



Thank You