

Natural Language Processing

Dr. Yashodhara Haribhakta

Department of Computer Engg. & I.T.,
College of Engineering Pune

Email: ybl.comp@coep.ac.in

Theory Assessment

T1/Quiz/Surprise test – 20 marks

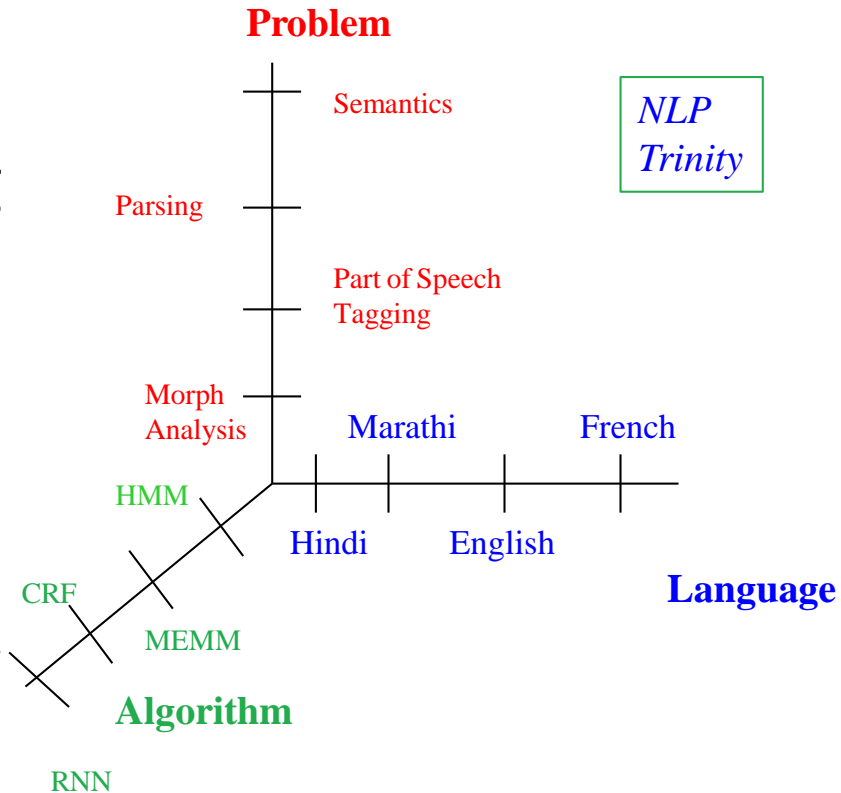
Mini-project – 20 marks

EndSem – 60 marks

Minimum marks – 40 marks for passing in this subject.

Course Objectives

- Introduce the fundamental concepts and techniques of natural language processing (NLP) by studying the phonological, morphological, syntactic and semantic processing.
- To gain an in-depth understanding of algorithms available for the processing of linguistic text information and the underlying computational properties of natural languages .



Why do we need to study NLP?

Introduction

- What is NLP?

Introduction

- What is NLP?
 - Processing text data so that able to infer some information which is useful.

Introduction

- What is NLP?
 - Processing text data so that able to infer some information which is useful.
- What is the main goal of NLP?

Introduction

- What is NLP?
 - Processing text data so that able to infer some information which is useful.
- What is the main goal of NLP?
 1. Fundamental and Scientific Goal
 - Deep Understanding of natural language

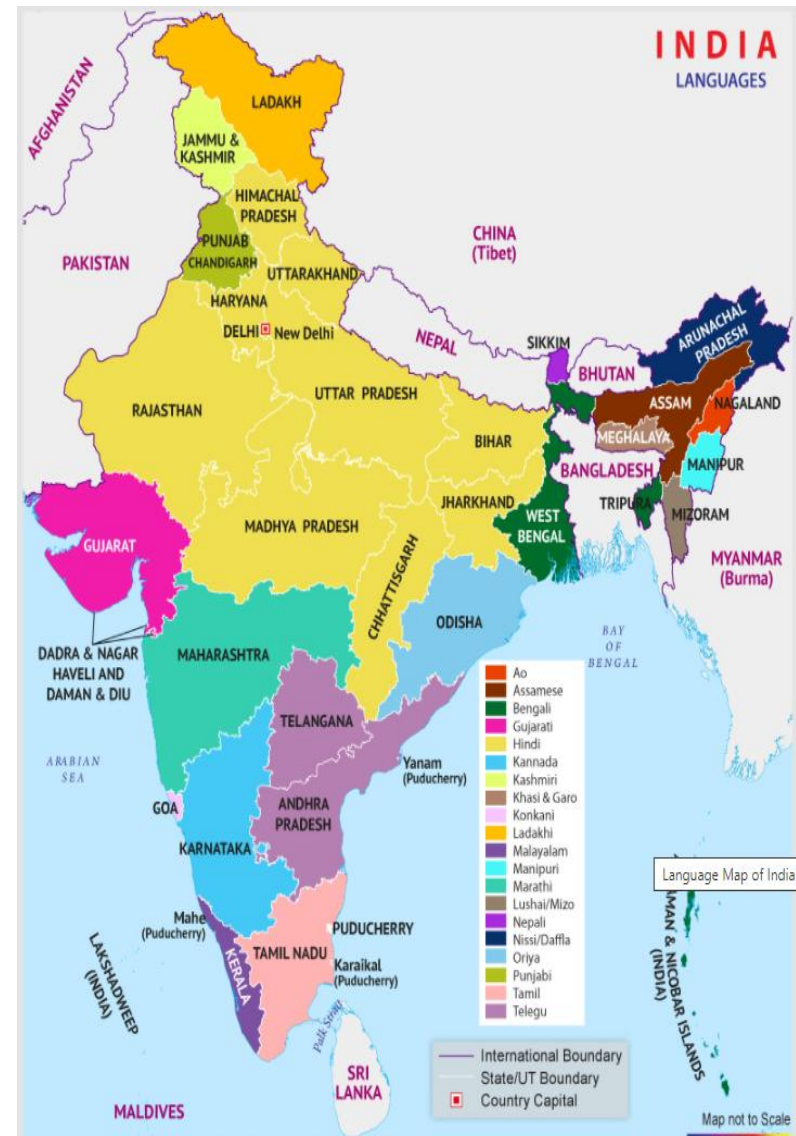
Introduction

- What is NLP?
 - Processing text data so that able to infer some information which is useful.
- What is the main goal of NLP?
 1. Fundamental and Scientific Goal
 - Deep Understanding of natural language
 2. Practical and Engineering goal
 - Design, implement and test systems that process natural language for practical applications

Language Families

Multilinguality: Indian situation

- Language families
 - Indo Aryan
 - Dravidian
 - Austro-Asiatic
 - Tibeto-Burman
- Languages that are ranked within 20 in the world in terms of the populations speaking them, are
 - Hindi : 3rd (~350 milion)
 - Bangla: 7th (~230 million)
 - Marathi: 15th (~84 million)



***Natural Language Processing:
Background & Relevance in Indian
Scenario***

Background: Indian Context

- India is a multi-lingual country with great linguistic and cultural diversities
- 22 official languages mentioned in the Indian constitution
- However, Census of India in 2011 reported-
 - **121 major languages**
 - **1,599 other regional languages**
 - **2,371 scripts**
 - **30 languages** are spoken by more than **one million native speakers**
 - **121** are spoken by more than **10,000 people**
- **20%** understand English
- **80%** cannot understand

Background

- Phenomenal growth in the number of internet users, social media (*Facebook, Twitter* etc.)
- Increasing tendency of using Indian language contents for exchanging information
- **Digital divide** cannot be tackled unless citizens are given flexibility in **communicating in their own languages**

Natural Language Processing (NLP) that deals with developing theories and techniques for effective communication in human languages play an important role towards creating this digital society

Motivation

TDIL: MeiT, Govt. of India

TDIL : Technology Development for Indian Languages Programme initiated by the Ministry of Electronics & Information Technology, Govt. of India

Objective:

- objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier;
- creating and accessing multilingual knowledge resources; and
- integrating them to develop innovative user products and services.

TDIL: Some major Machine Translation Projects

1. **Development of English to Indian Language Machine Translation System (Anuvadaksh):** Translator for English to Hindi/ Marathi/ Bangla/ Oriya/ Tamil/ Urdu/ Gujrati/ Bodo
2. **Development of English to Indian Language Machine Translation System with Angla-Bharti Technology:** ANGLABHARTI represents a machine-aided translation methodology specifically designed for translating English to Indian languages, like, English to Bangla/ Punjabi/ Malaylam/ Urdu/ Hindi/ Telugu
3. **Development of Indian Language to Indian Language Machine Translation System (Sampark)-** 18 pairs of languages, like, -Hindi to Bengali, Bengali to Hindi, Marathi to Hindi, Hindi to Marathi, Hindi to Punjabi, Punjabi to Hindi, Hindi to Tamil, Tamil to Hindi, Hindi to Kannada, Kannada to Hindi, Hindi to Telugu, Telugu to Hindi, Hindi to Urdu, Urdu-Hindi, Malaylam to Tamil, Tamil to Malaylam, Tamil to Telugu, Telugu to Tamil

TDIL: Some major initiatives

- Development of Cross-Lingual Information Access (CLIA)
 - Assamese, Bengali, Hindi, Oriya, Punjabi, Tamil, Telugu, Marathi, Gujarati
- Development of Robust Document Analysis & Recognition System for Indian Languages (OCR)-14 languages
 - Assamese, Bengali, Devanagri, Gujarati, Gurumukhi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Tamil, Telugu, Tibetan, Urdu
- Development of Text to Speech System in Indian Languages
- Development of Automatic Speech Recognition System in Indian Languages
- Development of Sanskrit Machine Translation System
- *Development of Hindi to English Machine Translation in Judicial Domain*

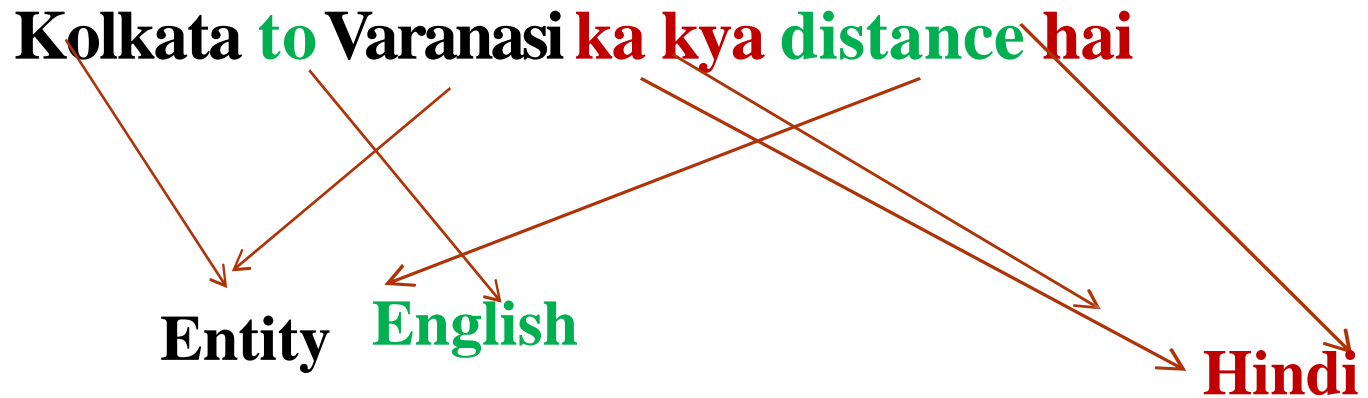
Languages and the Institutes working on different language

Language	Institute
Assamese	Guwahati University , Guwahati , Assam
Bengali	Indian Statistical Institute , Kolkata , West Bengal
Bodo	Guwahati University , Guwahati , Assam
Gujarati	Dharamsinh Desai University , Nadiad , Gujarat
Hindi	IIT Bombay , Mumbai , Maharashtra
Kannada	Mysore University , Mysore , Karnataka
Kashmiri	Kashmir University , Srinagar , Jammu and Kashmir
Konkani	Goa University , Taleigao , Goa
Malayalam	Amrita University , Coimbatore , Tamil Nadu
Marathi	IIT Bombay , Mumbai , Maharashtra
Meitei	Manipur University , Imphal , Manipur
Nepali	Assam University , Silchar , Assam
Oriya	Hyderabad Central University , Hyderabad , Andhra Pradesh
Punjabi	Thapar University and Punjabi University , Patiala , Punjab
Sanskrit	IIT Bombay , Mumbai , Maharashtra
Tamil	Tamil University , Thanjavur , Tamil Nadu
Telugu	Dravidian University , Kuppam , Andhra Pradesh
Urdu	Jawaharlal Nehru University , New Delhi

Code Mixing

Code-Mixing

- Code- mixing refers to the mixing of two or more languages or language varieties in speech/text



Code-Mixing in MyGov.in: Few Examples

- *Sir ji aapka ye abhiyan acha ha isse naye bharat ka nirman hoga maine apne school ke student ke sath milkar hospital ki safai ki and jagrukta rali nikali jisse log gandagi kam failaye.*
- *Aaj her school main swachta abhiyan honi chye we do it*
- *india ko clean rakhne ke lie gandgi karne walo pe penalty lagani chahiye jo kaam das sal me hoga penalty lagane ke bad wo kuch hi dino me ho jaega*
- *Modi sir swachh bharat m aapke bjp poltician photo click krawane k liye safai krte h sathinye neta sirf pik click krte h bs.*

NLP: In Governance

- **Uses of NLP in Government Websites**
 - Making e-governance related information to be available in multiple languages
- **Natural Language Generation in e-Governance**
 - Chatbot
 - E.g. farmer can not read or write, but with the multilingual support and NLP generation, s/he can communicate the query in any language and get it resolved



NLP: In Healthcare

- **NLP in Healthcare**

- the healthcare system is to provide better and 24/7 Electronic health record experience
- for doing Predictive analytics, Prescriptive analytics
- Patients can interact in his/her own language
- Easier for a patient to understand health status

- **Identification of the patients which require Improved Care Coordination**

- Automated detection of cancer, detection of the root causes related to any disorder are some of the examples

NLP: In Finance

- **Credit Scoring Method**
 - **Estimate risk factor of giving loan with the past histories**
 - **E.g. Lenddo EFL** (with 115 employees), a Singapore-based company developed a software called **Lenddo Score** which uses machine learning and NLP to assess and calculate an individual's creditworthiness.
- **Fraud detection in banking**
- **Stock market prediction-** based on sentiment

NLP: In Business

- **Searching –Autocorrect/Autocomplete**
- **Translation from one language to another language**
- **Survey Analysis**
- **Sentiment Analysis:** Analyzing public opinion
- **Email Filters:** Filtering out irrelevant emails
- **Information Extraction**

NLP: In Other domains

- **National Security**

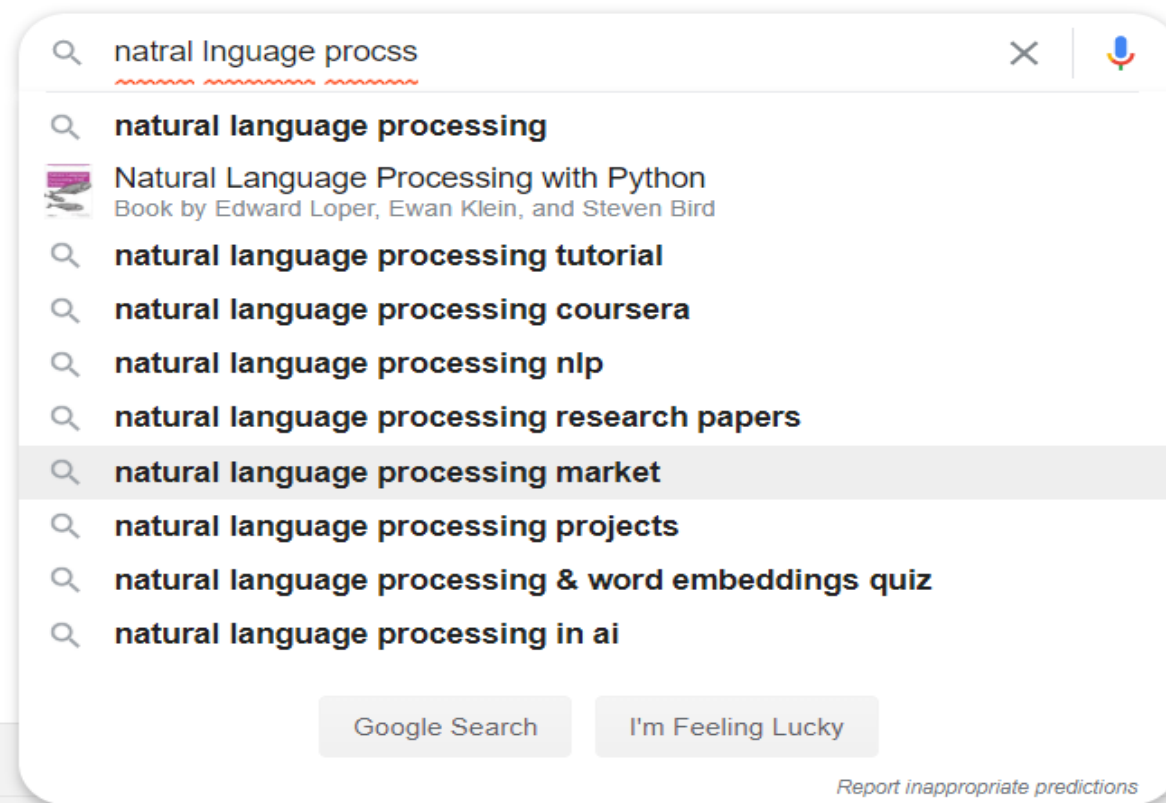
- Sentiment in Cross-border languages
- Hate Speech, Radicalization

- **NLP in Recruitment**

- searching the appropriate applications from the data, and it also can be used for selecting the best applications from the data available





Application







1. Search Autocorrect and Autocomplete




Application


2. Language Translator


 google translate   


 All  Books  Shopping  News  Images  More Settings Tools

About 65,60,00,000 results (0.42 seconds)



English – detected 





Marathi 

i am delivering an
online lecture 

मी एक ऑनलाइन व्याख्यान
देत आहे
Mī ēka ōnalā'ina vyākhyāna dēta āhē

[Open in Google Translate](#) [Feedback](#)

Application

2. Language Translator

The screenshot displays the Google Translate web interface. At the top, the Google logo is on the left, and a search bar contains the text "google translate". Below the search bar, navigation links for "All", "Books", "Shopping", "News", "Images", "More", "Settings", and "Tools" are visible. The search results indicate "About 65,60,00,000 results (0.42 seconds)". The main translation area shows "English – detected" on the left and "Hindi" on the right. The English input text is "i am delivering an online lecture", and the Hindi output is "मैं एक ऑनलाइन व्याख्यान दे रहा हूँ" with a phonetic transcription "main ek onalain vyaakhyaan de raha hoon" below it. Audio playback icons are present for both languages. At the bottom, there are links for "Open in Google Translate" and "Feedback".

Google

google translate

All Books Shopping News Images More Settings Tools

About 65,60,00,000 results (0.42 seconds)

English – detected Hindi

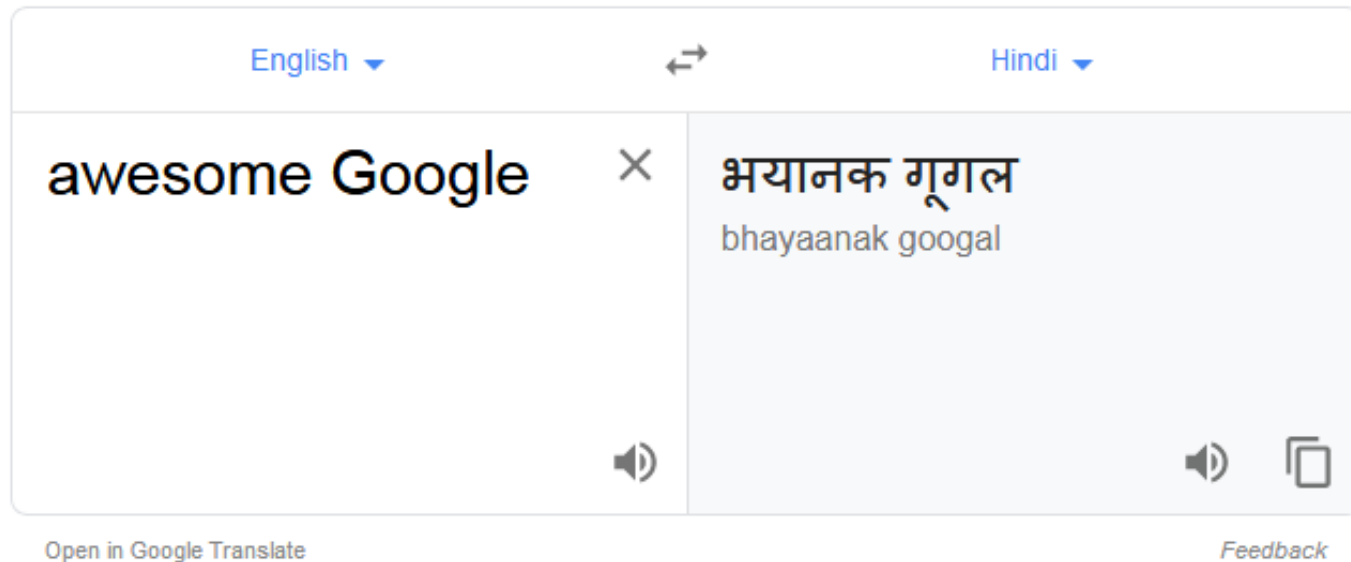
i am delivering an online lecture

मैं एक ऑनलाइन व्याख्यान दे रहा हूँ
main ek onalain vyaakhyaan de raha hoon

Open in Google Translate Feedback

Application

2. Language Translator



Google Translate

<https://translate.google.co.in/>

Google's free service instantly translates words, phrases, and web pages between English and over 100 other languages.

[About Google Translate](#) · [Translate Community](#) · [Download & use Google](#) · [Google](#)

Application

2. Language Translator



google translate



All



Books



Shopping



News



Images



More

Settings

Tools

About 65,60,00,000 results (0.42 seconds)

English – detected ▼



Hindi ▼

google is cool



गूगल शांत है
googal shaant hai



[Open in Google Translate](#)

[Feedback](#)

Application

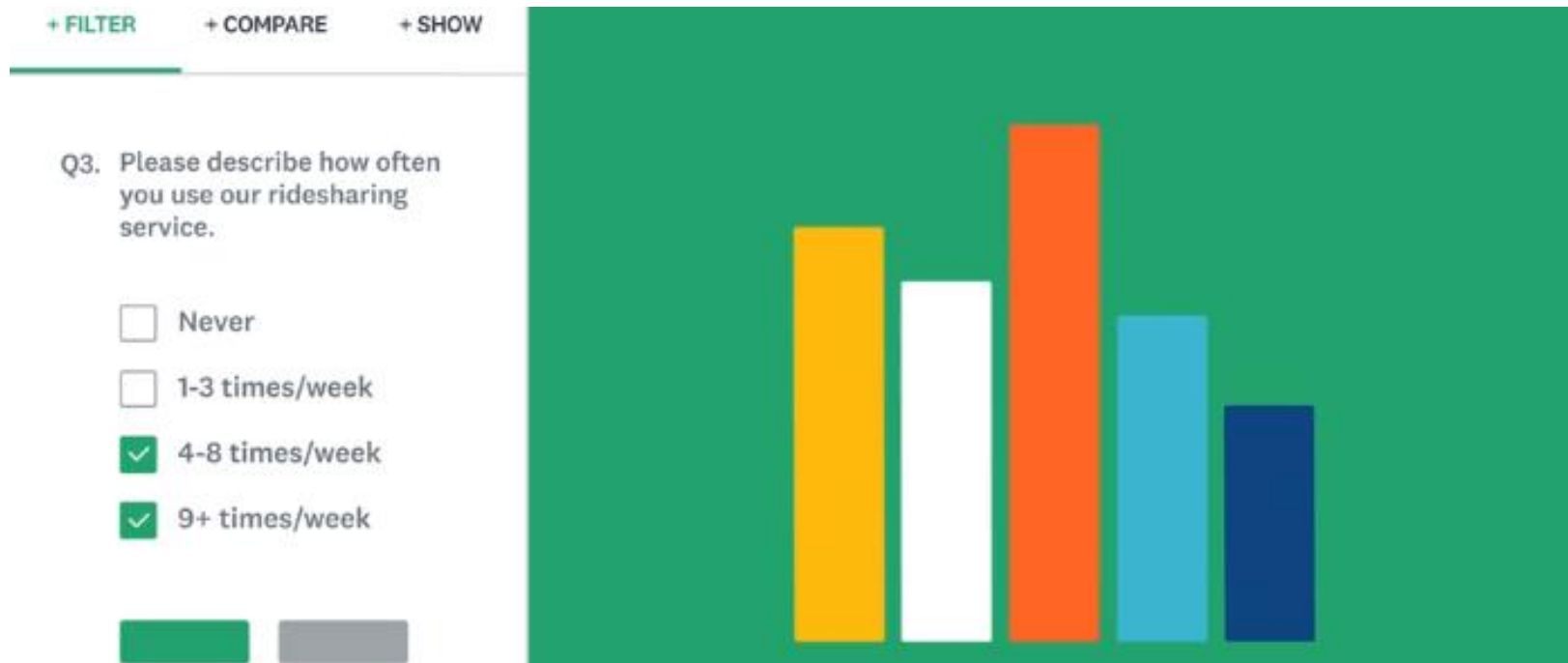
3. Sentiment Analysis



Discovering people opinions, emotions and feelings about
a product or service

Application

4. Survey Analysis



Application

5. Targeted Advertising



Application

6. Hiring and Recruitment

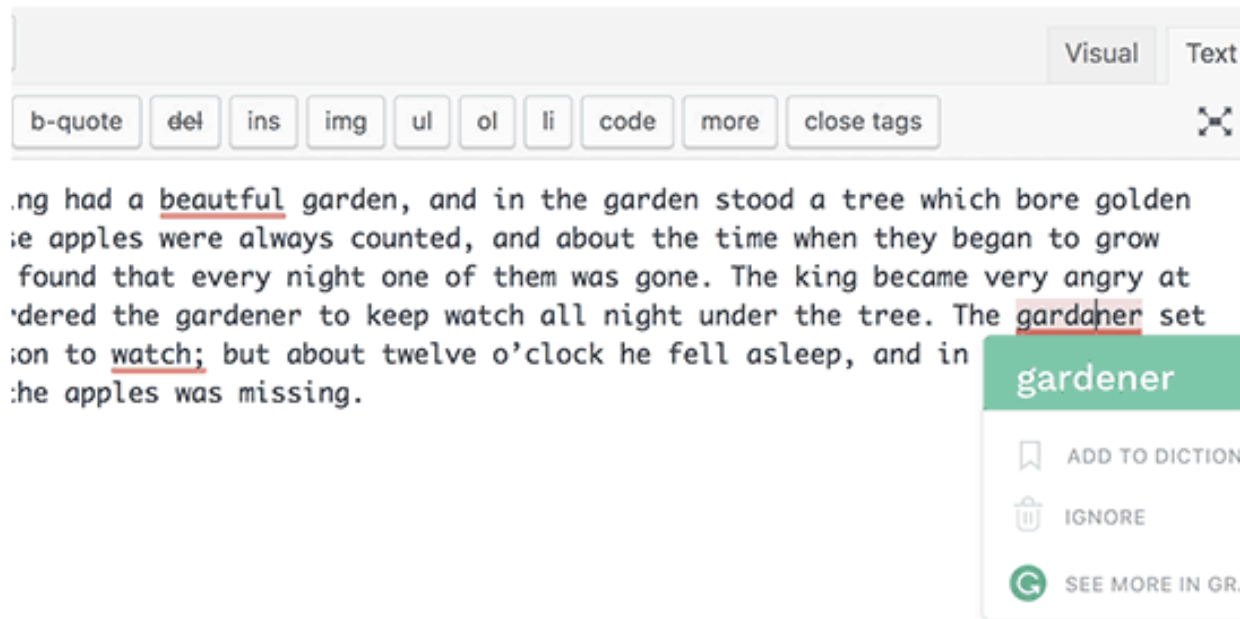


Application

7. Grammar Checkers

1. Grammarly

Grammarly is a popular online grammar checker tool available as a browser addon for Google Chrome, Firefox, and even Microsoft Edge. It checks for grammar and spelling errors as you write your posts.



Application

8. Email Filtering

Have you ever used Gmail ?

The screenshot displays the Gmail web interface. At the top, there's a search bar and a navigation menu. The left sidebar shows the 'Inbox' with 3,888 emails, and other categories like 'Starred', 'Snoozed', 'Important', 'Sent', 'Drafts', 'Categories', 'Social', 'Updates', 'Forums', and 'Promotions'. The main area shows the 'Primary' tab selected, displaying a list of emails from various senders like YouTube Kids, Mynta, Adrian at PylImageSe., ProGrad Junior, online course, Zomato, onlinecourses, and saurabh kumar (Meet.).

Gmail Interface Components:

- Search Bar:** Search mail
- Compose Button:** + Compose
- Left Sidebar:**
 - Inbox:** 3,888
 - Starred**
 - Snoozed**
 - Important**
 - Sent**
 - Drafts:** 25
 - Categories**
 - Social:** 5,441
 - Updates:** 2,129
 - Forums:** 1,732
 - Promotions:** 1,015
- Main Content Area:**
 - Primary Tab:** ResearchGate, YouTube, Linked...
 - Social Tab:** 46 new (ResearchGate, YouTube, Linked...)
 - Promotions Tab:** 44 new (ProGrad Junior, Quora, etc.)

Email List (Primary Tab):

Sender	Subject
YouTube Kids	Celebrate Reading Month with YouTube Kids - Discover read-alongs, early literacy learning videos, and more.
Mynta	Happy Raksha Bandhan! - Presents to celebrate your first friend To unsubscribe from these mailings, you may opt out here. If you would like to exp
Adrian at PylImageSe.	OCR'ing Non-English Languages with Tesseract. - Hi, In today's brand-new tutorial, you will learn how to OCR non-English languages using the Tesser
ProGrad Junior	Your Child's Coding Curriculum - Dear Parent, This is Rajesh, Co-founder of ProGrad Junior. On behalf of myself and the entire team here at ProGrad .
online course	IIT Madras Online Degree Program (BSc in Programming & Data Science) - Tune in to Know more! - Dear Learner, The demand for data science is incr
Zomato	Know whose birthday it is today? 🥳 - ...the treat will be on us anyway! In case you wish to stop receiving emails from Zomato, please unsubscribe he
onlinecourses	Want to specialize? Complete an NPTEL Domain!! - Dear Learner NPTEL Provides 41 Domains across 10 Disciplines which makes learners an expert
saurabh kumar (Meet.	Pune Artificial Intelligence & Deep Learning Online list: "[Meetup Tomorrow] Text Classification and Use-Cases" - saurabh kumar (Co-Organizer) ser

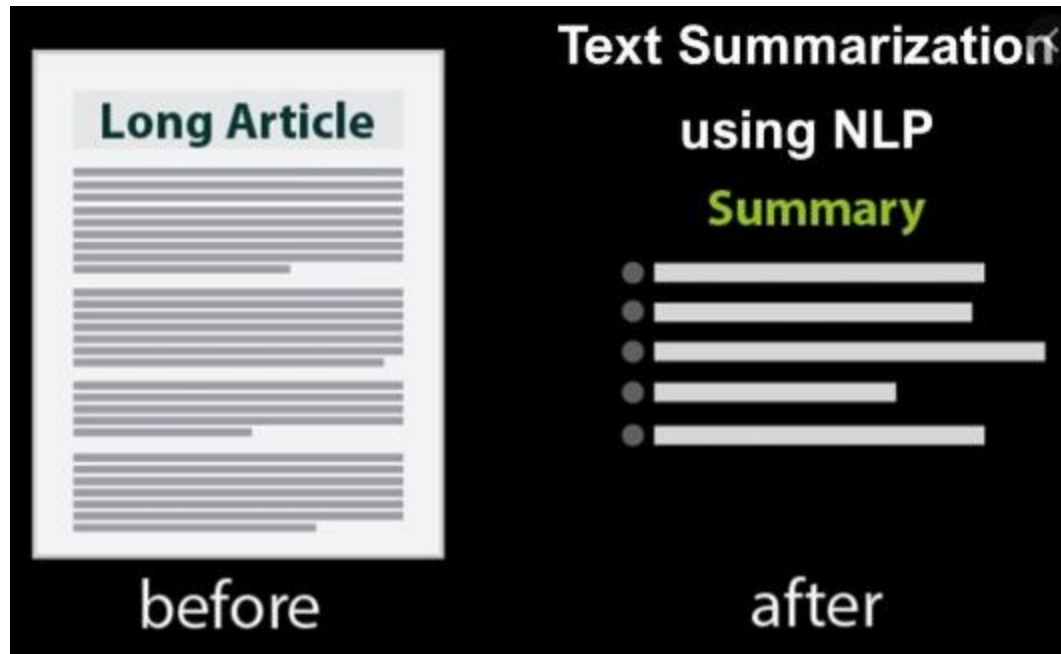
Application

9. Question Answering/ Chatbots



Application

10. Automatic Summarization



Some of the APIs : Aylien Text Analysis, MeaningCloud Summarization, ML Analyzer, Summarize Text, Text Summary.

Application

11. Information Retrieval

- Lot of unstructured data. Identify entities of interest and their relationship.

Example:

College of Engineering Pune, selected Dr. Jibi Abraham, as Dean Academics from June 2019, responsible for all academic based activities. She was institute MIS-Incharge before this. She succeeds Dr. M.S. Sutaone, who is now Deputy Director of the institute.

Person Name	Institute /Company Name	Post	State	Year
Dr. Jibi Abraham	COEP	Dean Academics	Start	2019
Dr. Jibi Abraham	COEP	MIS-Incharge	End	2019
Dr. M.S.Sutaone	COEP	Deputy Director	Start	2019

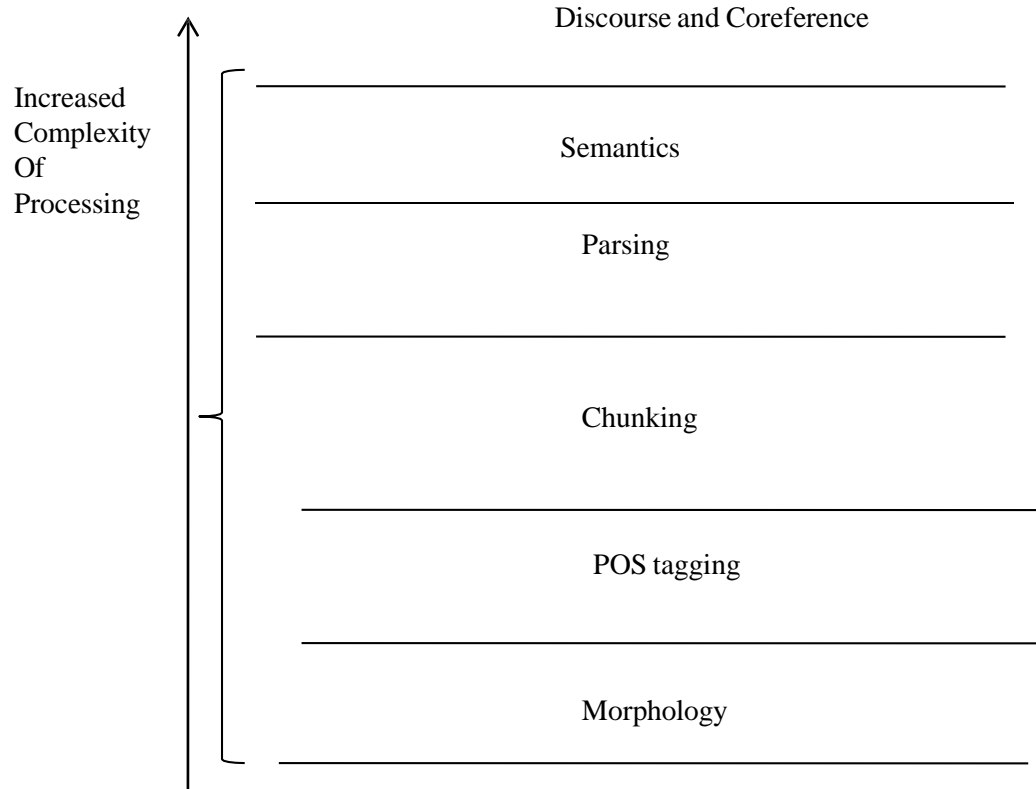
NLP: Projected Growth

- Growing in an exponential manner
- Expected to touch the market of more than **\$25 billion in 2022**
 - With compound growth rate of 16% annually
- Reasons behind this growth
 - Rising of the Chatbots
 - Urge of discovering the customer insights
 - Transfer of technology of messaging from manual to automated
 - Translation of contents, and
 - many other tasks which are required to be automated and involve language/Speech at some point
 - Etc.

Major Industries: Amazon, Google, Microsoft, Facebook, IBM etc.

Layers in NLP

Layers of Language Processing



Different Levels of Language Analysis

1. Phonetic and phonological knowledge

- Words are related to sound

2. Morphological Knowledge

- Word are constructed from morphemes

3. Syntactic Knowledge

- Words put together to form correct sentence
- Structural role played by each word

4. Semantic Knowledge

- What word means
- Context independent meaning- meaning the sentence has regardless of the context in which it is used

Different Levels of Language Analysis

5. Pragmatic Knowledge

- How sentences are used in different situations and how use affects the interpretation of the sentence

6. Discourse Knowledge

- How the immediately preceding sentences affects the interpretation of next sentence

7. World Knowledge

- Includes general language about the structure of the world that language users must have in order to maintain a conversation.