Introduction to Natural Language Processing

This lecture

- **Introduction to Language**
- What is NLP? Why it is important?
- NLP Applications
- Linguistic Knowledge
- Challenges
- NLP course- What will you learn from this course?

- Natural language vs. Artificial Language
- A vocabulary consists of a set of words
- A text is composed of a sequence of words from a vocabulary
- A language is constructed of a set of all possible texts
- What is NLP
  - **Wiki: Natural language processing**(**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.
  - Identify the structureand meaningof words, sentences, textsand conversations
  - Deep understanding of broad language

NLP

- Identify the structure and meaning of words, sentences, texts and conversations
- Deep understanding of broad language
- NLP is all around us

This lecture

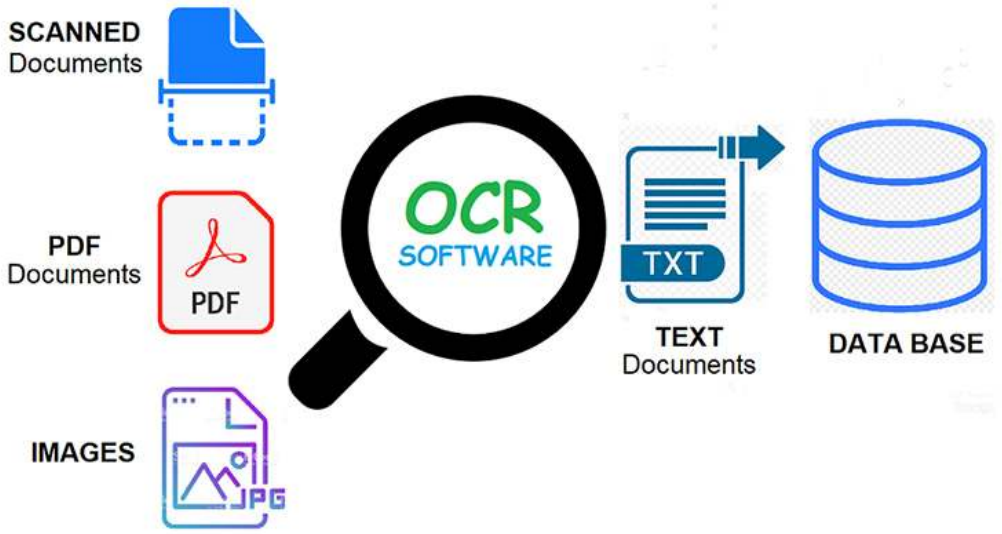
- Introduction to Language
- What is NLP? Why it is important?
- **NLP Applications**
- Linguistic Knowledge
- Challenges
- NLP course- What will you learn from this course?

NLP Applications

- Spell and Grammar Checking
- Checking spelling and grammar
- Suggesting alternatives for the errors

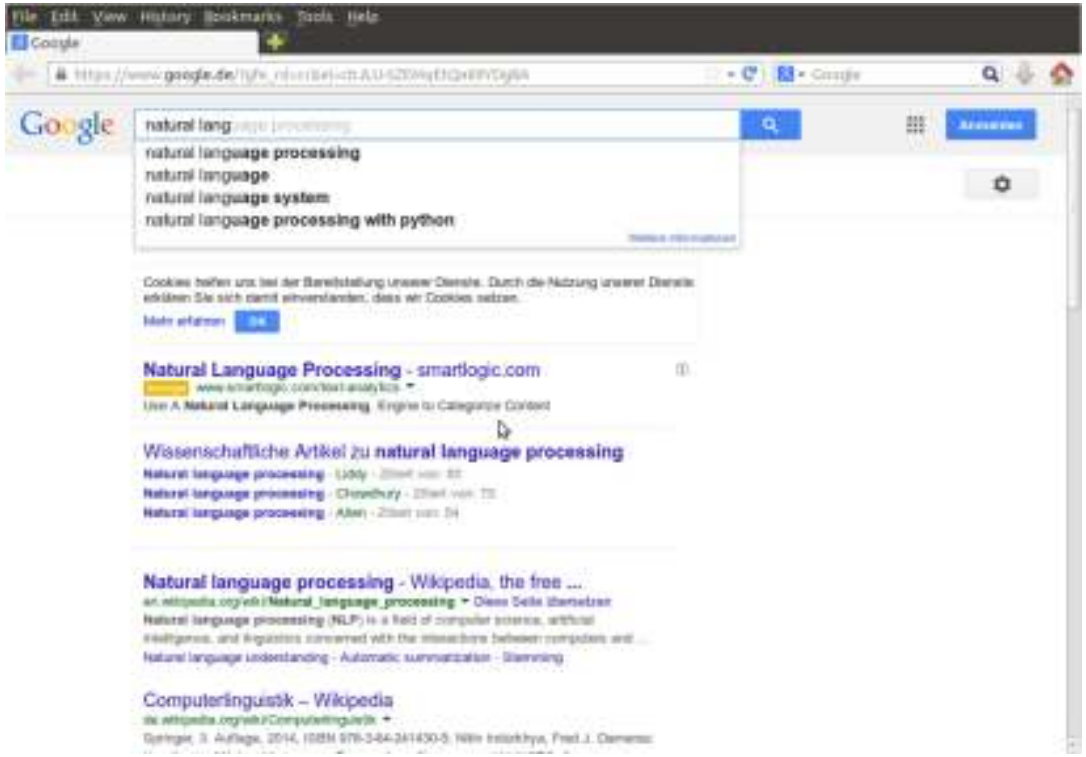
Optical Character Recognition

- Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image



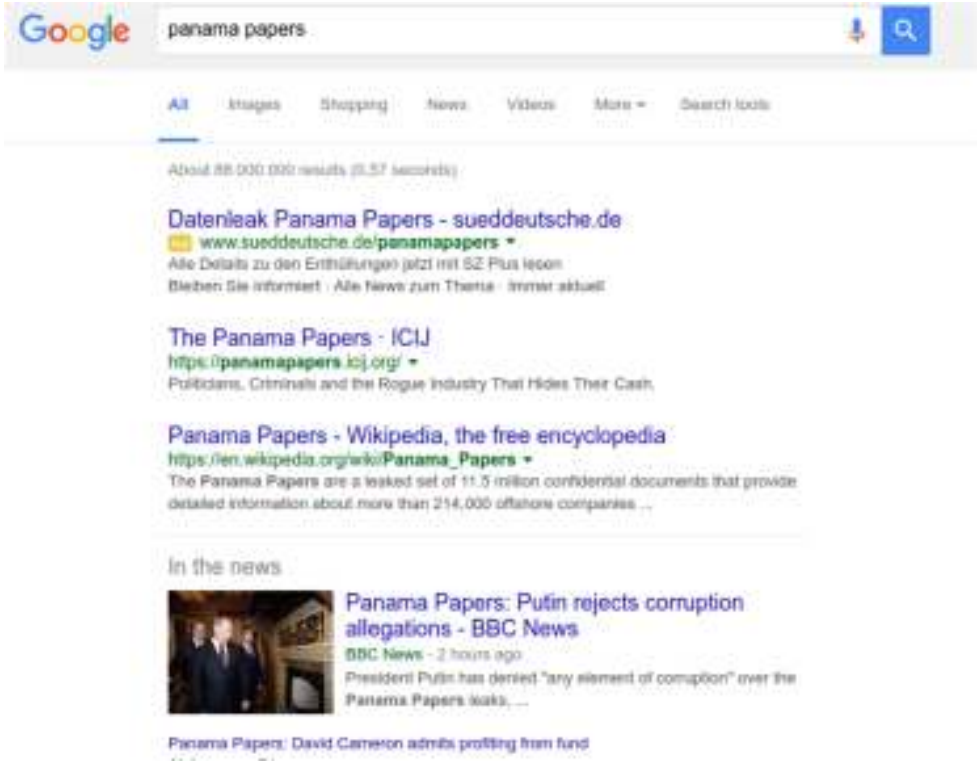
Word Prediction

- Predicting the next word that is highly probable to be typed by the user



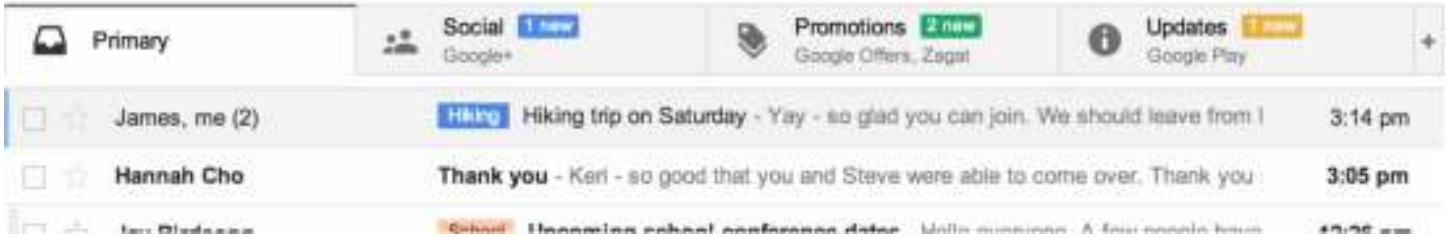
Information Retrieval

Finding relevant information to the user's query



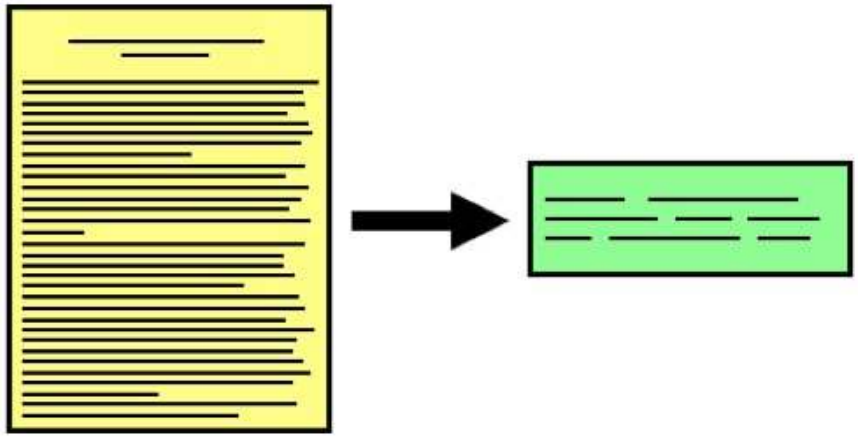
Text Categorization

- Assigning one (or more) pre-defined category to a text



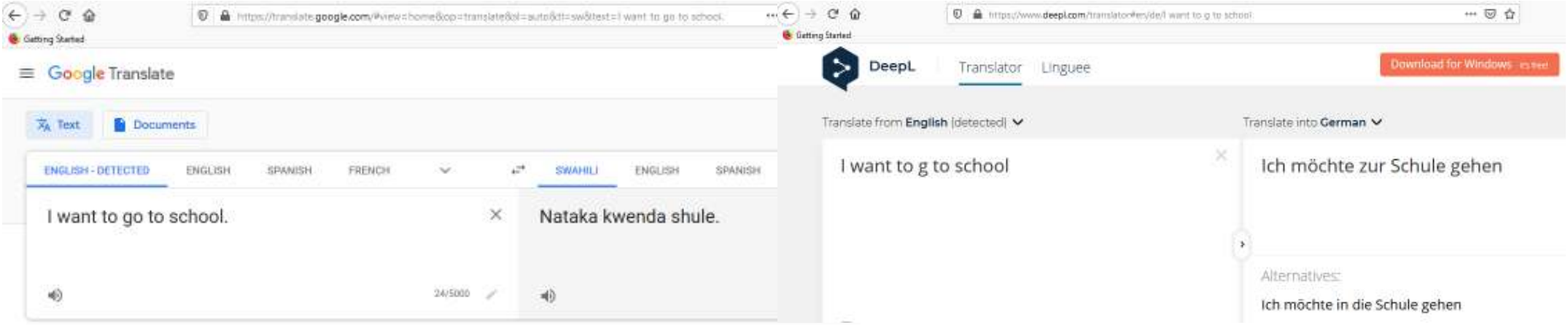
Summarization

- Automatic summarization is the process of shortening a set of data computationally, to create a subset that represents the most important or relevant information within the original content



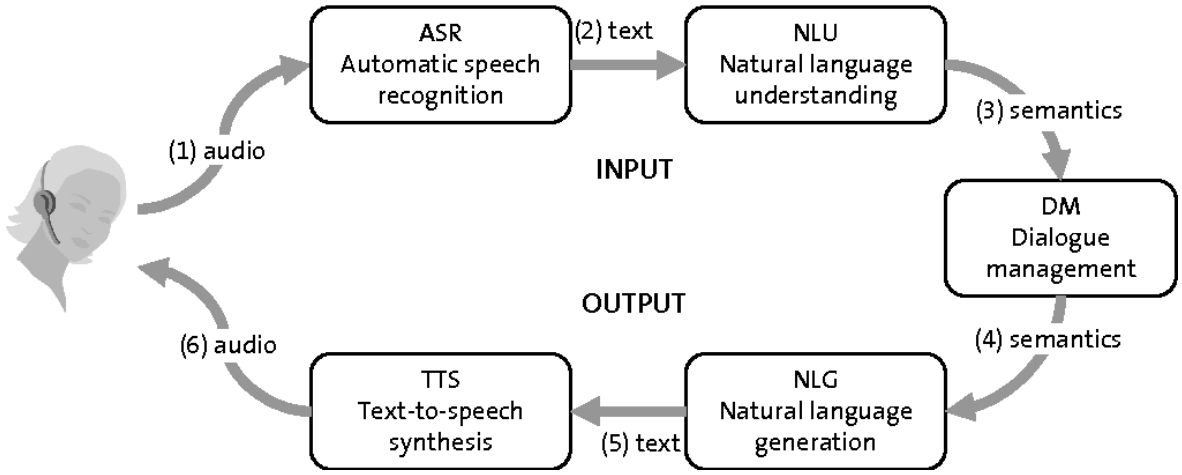
Machine translation

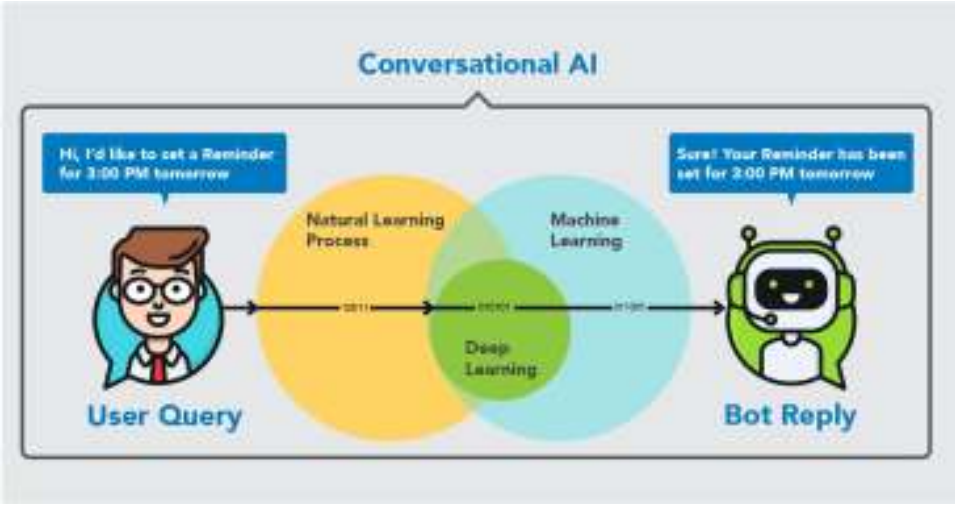
- The task of automatically converting source text in one language to text in another language.
- In a **machine translation** task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.



Dialogue systems

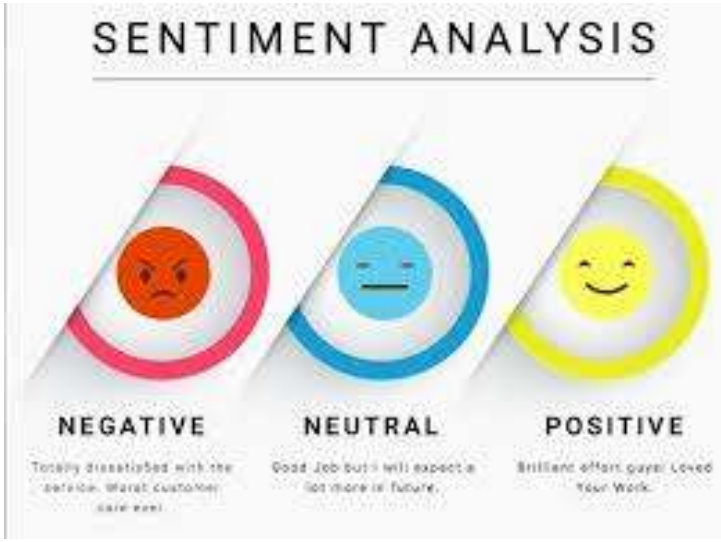
- A dialogue system, or conversational agent, is a computer system intended to converse with a human. Dialogue systems employed one or more of text, speech, graphics, haptics, gestures, and other modes for communication on both the input and output channel.





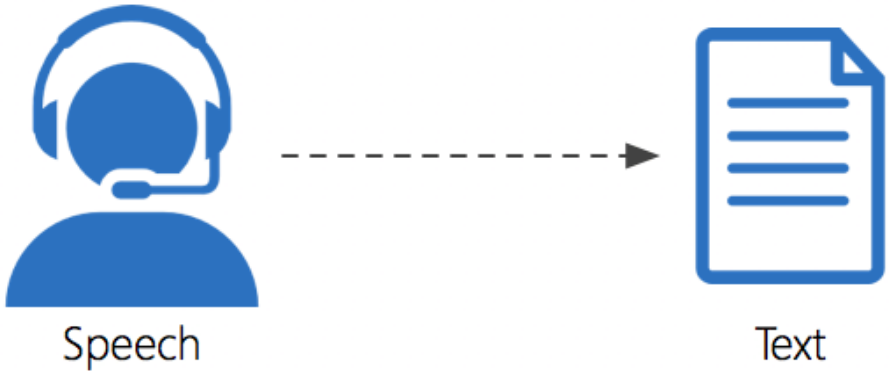
Sentiment/ opinion analysis

- the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.



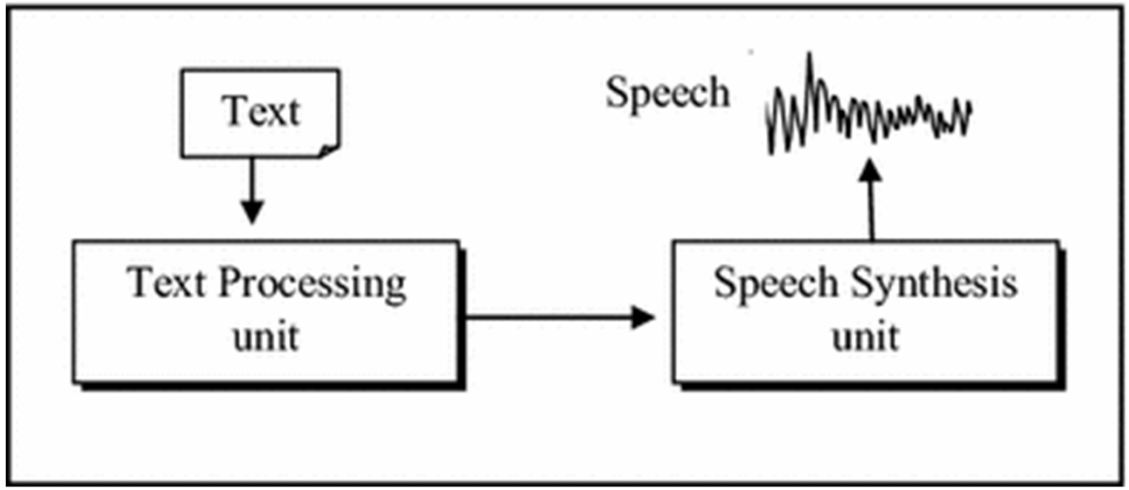
Speech recognition

- Recognizing a spoken language and transforming it into a text



Speech synthesis

- Producing a spoken language from a text



Question answering

- Question answering is a computer science discipline within the fields of information retrieval and natural language processing, which is concerned with building systems that automatically answer questions posed by humans in a natural language
- IBM Watson is an automated question answering system and won the Jeopardy contest.

Digital personal assistant

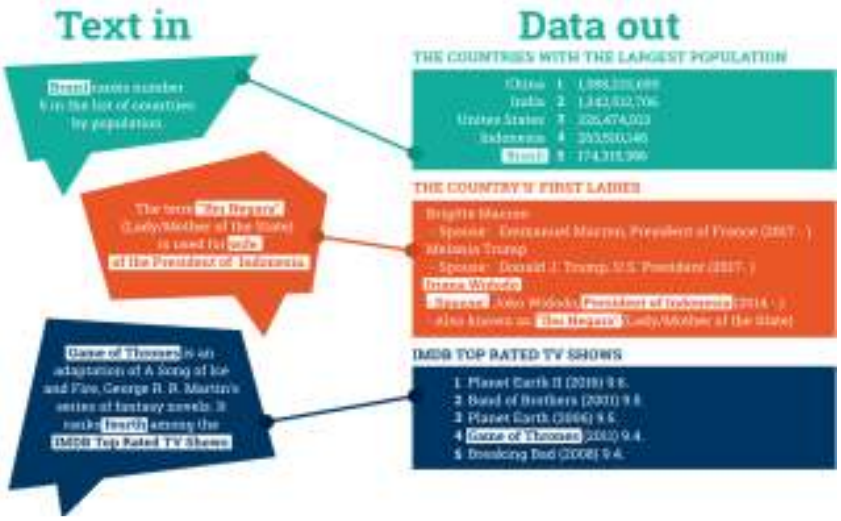
- An **intelligent virtual assistant** (**IVA**) or **intelligent personal assistant** (**IPA**) is a software agent that can perform tasks or services for an individual based on commands or questions (Natural Language Instructions). Sometimes the term "chatbot" is used to refer to virtual assistants generally or specifically accessed by online chat.

Information extraction

- Unstructured text to database entries
- **Information extraction** is the process of **extracting** specific (pre-specified) **information** from textual sources.
- One of the most trivial examples is when your email extracts only the data from the message for you to add in your Calendar.

Food Tutorials are Infinitely Better When Directed By Wes Anderson. Bruce Lee's biopic, 'Little Dragon', to be directed by Shekhar Kapur. Stallone directed his first short film Vic.

- Wes Anderson **directed** Food Tutorials
- Shekhar Kapur **directed** Little Dragon
- Stallone **directed** Vic



Language Comprehension

- **Machine Comprehension** (MC), or the ability to read and understand unstructured text and then answer questions about it remains a challenging task for computers.

> **Christopher Robin** is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh.** As **a boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

- Q: who wrote Winnie the Pooh?
- Q: where is Chris lived?

Level of difficulties

- Easy (mostly solved)
  - Spell and grammar checking
  - Some text categorization tasks
  - Some named-entity recognition tasks
- Intermediate (good progress)
  - Information retrieval
  - Sentiment analysis
  - Machine translation
  - Information extraction

Levels of difficulty

- Difficult (still hard)
  - Question answering
  - Summarization
  - Dialog systems

This lecture

- Introduction to Language
- What is NLP? Why it is important?
- NLP Applications
- **Linguistic Knowledge**
- Challenges
- NLP course- What will you learn from this course?

Linguistic knowledge

NLP and linguistics :

- Letters – a,b,c … z
- Words – combining letters to form words
- Phonetics and phonology - The study of linguistic sounds and their relations to words
- Morphology - The study of internal structures of words and how they can be modified, parsing complex words into their components e.g. (ni)(na)(kula).
- Syntax - The study of the structural relationships between words in a sentence
- Semantics - The study of the meaning of words, and how these combine to form the meanings of sentences

Discourse

- The study of linguistic units larger than a single statement

  John reads a book. He borrowed it from his friend.

Pragmatics

- Social use of language
- The study of how language is used to accomplish goals, and the influence of context on meaning
- Understanding the aspects of a language which depends on situation and world knowledge

Give me the salt!

Could you please give me the salt?

This lecture

- Introduction to Language
- What is NLP? Why it is important?
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- **Challenges**
- NLP course- What will you learn from this course?

Challenges

- Word sense ambiguity

Word sense / meaning ambiguity

Ambiguity

- Ambiguous headlines:
  - Include your children when baking cookies – as an ingredient??
  - Hospitals are Sued by 7 Foot Doctors – Doctors with 7 feet??
  - Iraqi Head Seeks Arms – Head seeking arms??

- Discourse ambiguity- Pronoun reference ambiguity
  - Madam Lilian normally brings her cat , Cindy, to class. **She** just loves to give big kisses.
    - Does she refer to Madam Lilian or her cat?
  - Alice understands that you like your mother, but **she** …
    - Does she refer to Alice or your mother?

Ambiguity

- Semantic ambiguity
  - Words with multiple meanings
    - Fall
      - The third season of the year
      - Moving down towards the ground or towards a lower position
  - Sentences expressing more than one meaning
    - The door is open.
      - Expressing a fact
      - A request to close the door
- Syntax and ambiguity
  - John killed a man with a knife

Phonetics and phonology

- Words which sound the same but have different meanings
  - Plain –plane
  - Sea – see
  - But – butt
  - Two –too
  - Ice cream – I scream
  - Your students – you’re students

Paraphrasing

- Different words/sentences express the same meaning
- Season of the year
  - Fall
  - Autumn
- Book delivery time
  - When will my book arrive?
  - When will I receive my book?

Language is not static

- Language grows and changes
- E.g. cyber lingo - lol, lmao, bff, luwanh

scale

- What is the size of a language? How many words?
- Examples:
  - Bible (King James version): ~700K words
  - Penn Tree bank ~1M from Wall street journal
  - Newswire collection:500M+
  - Wikipedia: 2.9 billion words (English)
  - Web: several billions of words

This lecture

- Introduction to Language
- What is NLP? Why it is important?
- NLP Applications
- Linguistic Knowledge
- Challenges
- **NLP course- What will you learn from this course?**

What you will learn in this course

- The NLP pipeline - key components of text understanding and
  - Core NLP techniques: tokenization, lemmatization, stemming, chunking, Sentence splitting, part of speech tagging, syntactic parsing
  - Core NLP technologies : named entity recognition, co-reference resolution, event extraction, language modelling
- Text analytics using Python, NLTK, Spacy
- Text classification and sentiment analysis
  - Sentence representation – bag of words, tf-idf
  - Building a simple text classifier
- Recent trends in NLP – words embeddings and Neural Networks