19CSE453 – Natural Language Processing

Introduction

By

Ms. Kavitha C.R.

Dept. of Computer Science and Engineering

Amrita School of Computing, Bengaluru

Amrita Vishwa Vidyapeetham

Syllabus **2-0-3-3**

Unit 1

Introduction- Human languages, models, ambiguity, processing paradigms; Phases in natural language processing, applications. Text representation in computers, encoding schemes. Linguistics resources- Introduction to corpus, elements in balanced corpus, TreeBank, PropBank, WordNet, VerbNet etc. Resource management with XML, Management of linguistic data with the help of GATE, NLTK. Regular expressions, Finite State Automata, word recognition, lexicon. Morphology, acquisition models, Finite State Transducer, N-grams, smoothing, entropy, HMM, ME, SVM, CRF.

Unit 2

Part of Speech tagging- Stochastic POS tagging, HMM, Transformation based tagging (TBL), Handling of unknown words, named entities, multi word expressions. A survey on natural language grammars, lexeme, phonemes, phrases and idioms, word order, agreement, tense, aspect and mood and agreement, Context Free Grammar, spoken language syntax. Parsing- Unification, probabilistic parsing, TreeBank. Semantics- Meaning representation, semantic analysis, lexical semantics, WordNet Word Sense Disambiguation- Selectional restriction, machine learning approaches, dictionary based approaches.

Unit 3

Discourse- Reference resolution, constraints on co-reference, algorithm for pronoun resolution, text coherence, discourse structure. Applications of NLP- Spell-checking, Summarization Information Retrieval- Vector space model, term weighting, homonymy, polysemy, synonymy, improving user queries. Machine Translation–EM algorithm - Discriminative learning - Deep representation learning - Generative learning.

Text Book(s)

1. *Martin JH, Jurafsky D. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson/Prentice Hall; 2009.*

Reference(s)

1. *James A.. Natural language Understanding, Second Edition, Pearson Education;1994.*

2. Bharati A., Sangal R., Chaitanya V.. Natural language processing: a Paninian perspective, PHI; 2000.

3. Tiwary U S, Siddiqui T. Natural language processing and information retrieval. Oxford University Press, Inc.; 2008.

Course Outcomes

CO1: Understand the models, methods, and algorithms of statistical Natural Language Processing (NLP) for common NLP tasks.
CO2: Understand mathematical and statistical models for NLP.
CO3: Understand linguistic phenomena and linguistic features relevant to each NLP task.
CO4: Apply probabilistic models in code.
CO5: Apply learning models to NLP tasks such as speech recognition, machine translation, spam filtering, text classification, and spell checking

Evaluation Pattern

Internal (70)	**Continuous Assessment:50**		**Weightage**
	Quizzes	Quiz1	**5**
		Quiz2	**5**
	Lab Component	Lab Test(10 marks)	**10**
		Lab Quiz(10 marks)	**10**
	Project	1(5+15 marks)	**20**
	Mid Term	50	**20**
External (30)	**End Semester: 30**		
	End Sem	100	**30**

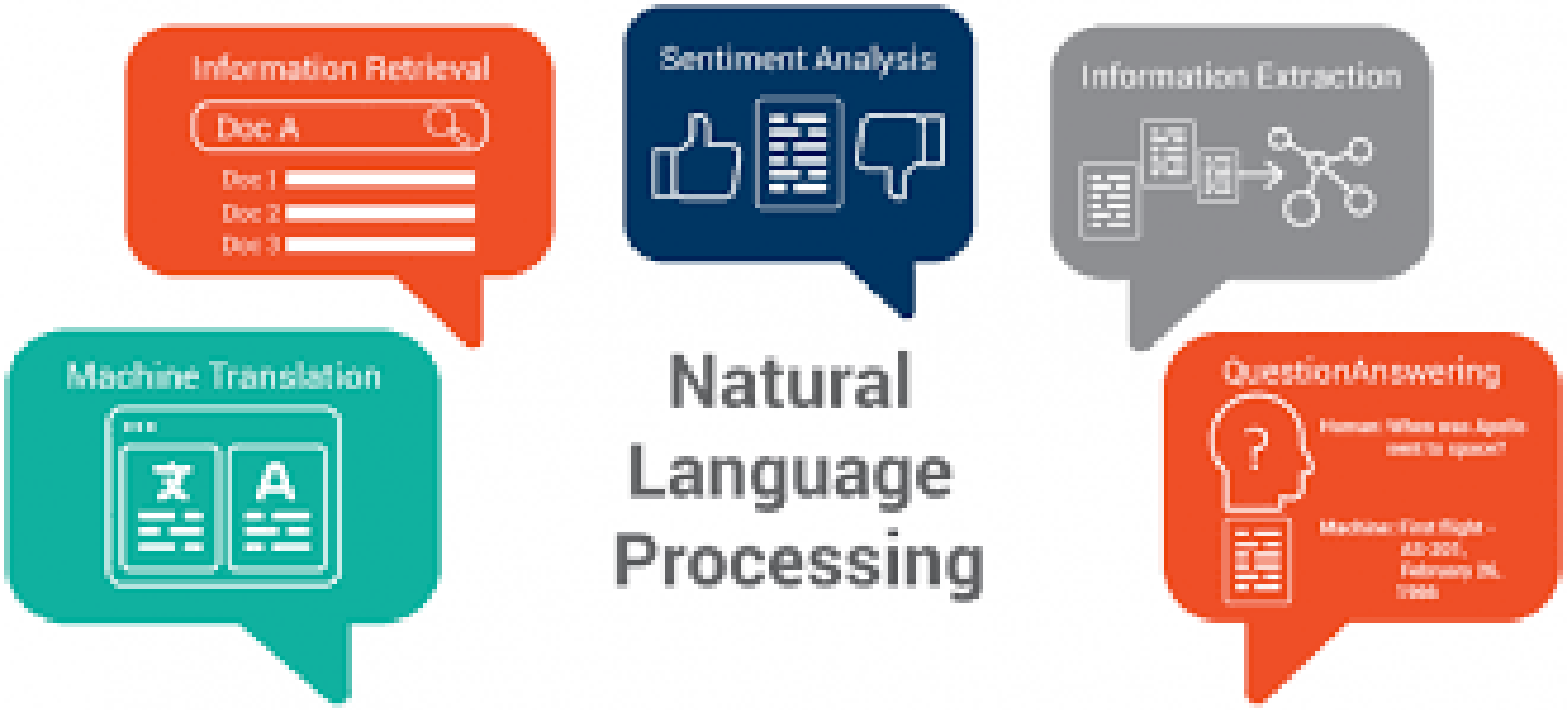
What is Natural Language?

- The **natural language** is the everyday language that is used by humans to communicate, such as English, Chinese, German etc.,
- **Programming languages** are used for developing computer programs, which enable a computer to perform some operations.
- Millions or billions of people use a certain language everyday in a slightly different way.
- A natural language can have thousands of different words, new words are created on the fly add to this complexity.
- Words can have different meanings depending on context

What is Natural Language Processing?

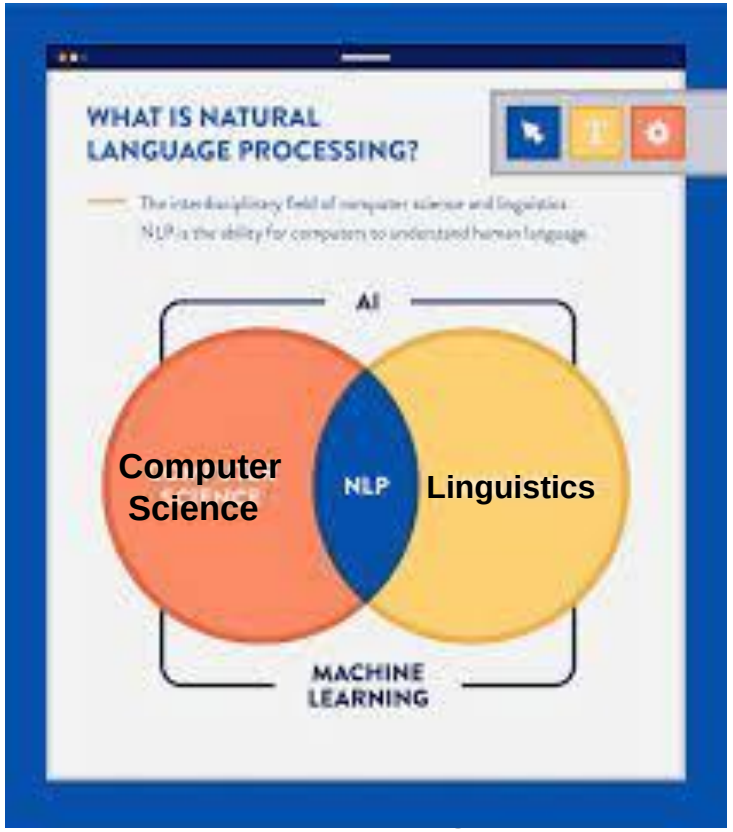
- An automated system which is capable of processing and analyzing text data
- **Natural Language Processing (NLP)** is concerned with the interactions between computers and human **(natural) languages**, in particular how to program computers to **process** and analyze large amount of **natural language** data.
- **Natural language processing** helps computers communicate with humans in their own **language**.

For example: NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment etc.



What is Natural Language Processing?

- NLP is an interdisciplinary field that uses computational methods to:
  - Investigate the properties of written human language and model the cognitive mechanisms underlying the understanding and production of written language.
  - Develop novel practical applications involving the intelligent processing of written human language by computer.



Goal of NLP

- Ultimate goal -> Teach computers to understand language to the extent what we can understand…
- When this is achieved, computer systems will be able to understand, draw inferences from, summarize, translate and generate accurate natural human text and language
- The ultimate goal of natural language processing is to develop computers that can understand natural languages

- *Fundamental Goal*

  Deep understanding of *broad* language

- *Engineering and practical Goal*

  Design, implement, and test systems that process natural languages for practical applications

Why NLP?

- Text is the largest repository of human knowledge
    - Wiki articles
    - News articles
    - Scientific research articles
    - Social media
    - Web pages
- All these channels are constantly generating large amount of text data .
- And because of the large volumes of text data as well as the highly unstructured data source, we can no longer use the common approach to understand the text and this is where NLP comes in.

Applications for better understanding of NLP in Engineering Goal

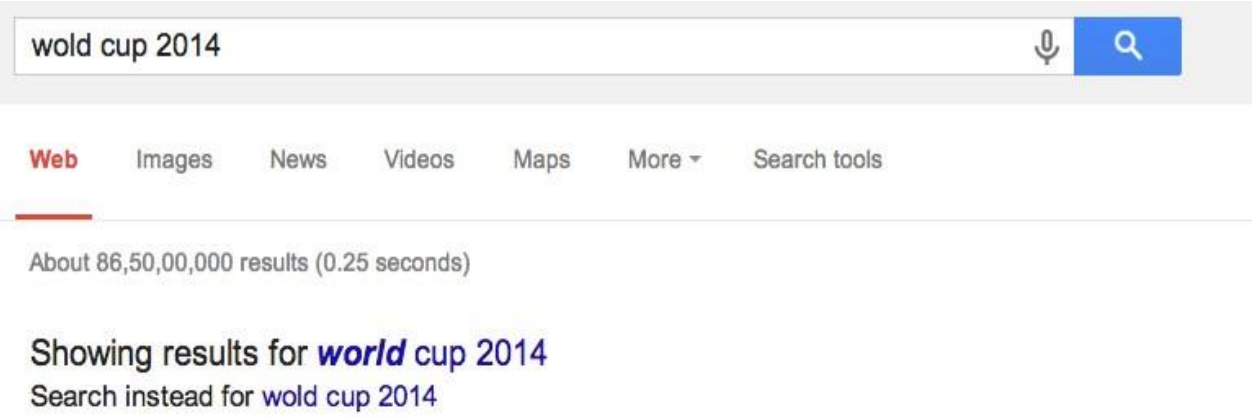
A few applications are..

Goals can be very ambitious:

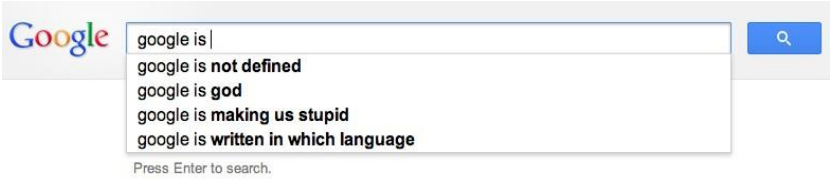
- Good quality translation-word can have multiple meanings
  – Solved by designing efficient algorithm

Goals can be practical too:

- Search Engines- Automatic query correction-Solved in NLP



- Auto Query Completion: Language modelling is applied to complete the task



- Information Extraction:

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

Other Goals

- Spam detection
- Machine Translation services on the Web
- Text Summarization

Natural Language Technology not yet perfect

But still good enough for several useful applications

Why NLP Hard/Difficult?

- It's the **nature of the human language** that makes NLP difficult. While humans can easily master a language, the **ambiguity and imprecise characteristics** of the natural languages are what make NLP difficult for machines to implement.
- Example: "I love Blackberry". In this case, Blackberry could mean both; **a phone or a fruit**. Such ambiguities are hard for computers to interpret.
- Ambiguity is the primary difference between natural and computer languages.

Why else is NLP hard?

Non standard use of English, which makes NLP Hard

Why NLP is difficult

- Human language is special for several reasons.
- It is specifically constructed to convey the **speaker/writer's meaning**.

- Understanding human language is considered a difficult task due to its complexity.

For example, there is an infinite number of different ways to arrange words in a sentence.

- Also, words can have several meanings and contextual information is necessary to correctly interpret sentences.
- Every language is more or less unique and ambiguous.

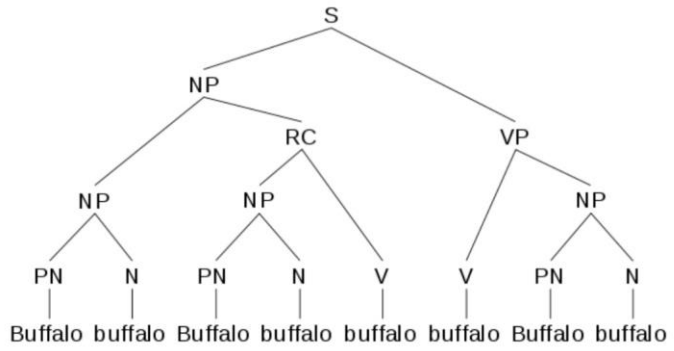
Why is NLP hard?

Lexical Ambiguity

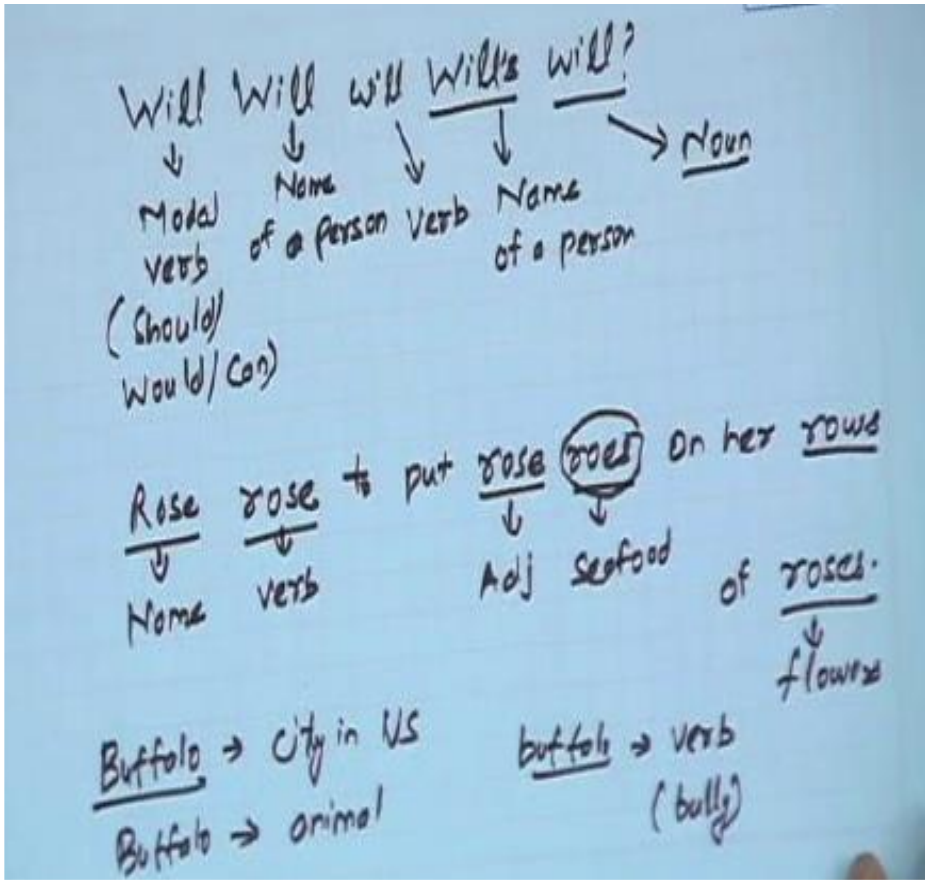
1. Will Will will Will's will?

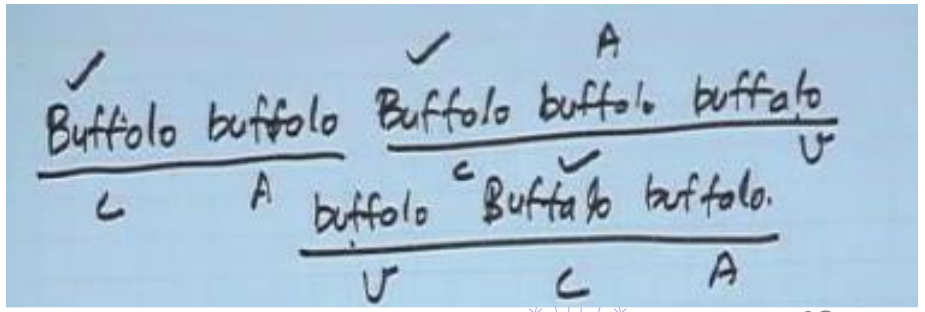
2. Rose rose to put rose roes on her rows of roses

3. Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo



Simplified parse tree S = sentence
NP = noun phrase
RC = relative clause
VP = verb phrase
PN = proper noun
N = noun
V = verb



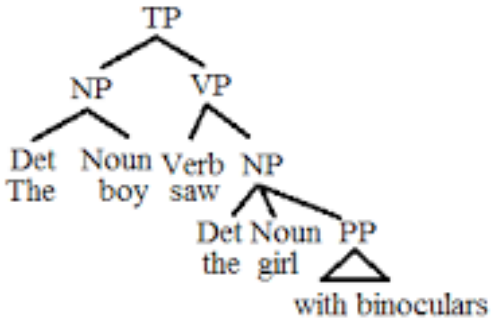


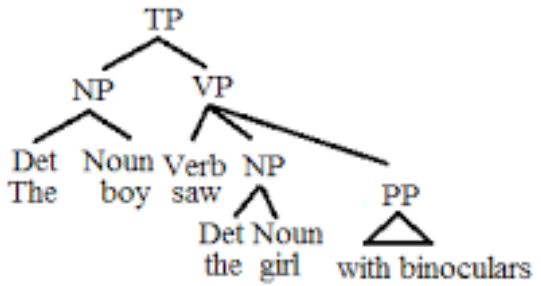
Why is NLP hard?

Language ambiguity: Structural

1.The boy saw the girl with binoculars.

2. Flying planes can be dangerous.





Language imprecision and vagueness

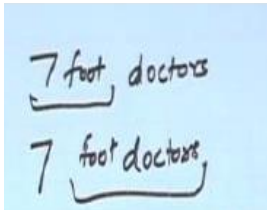
1. It is very warm here.

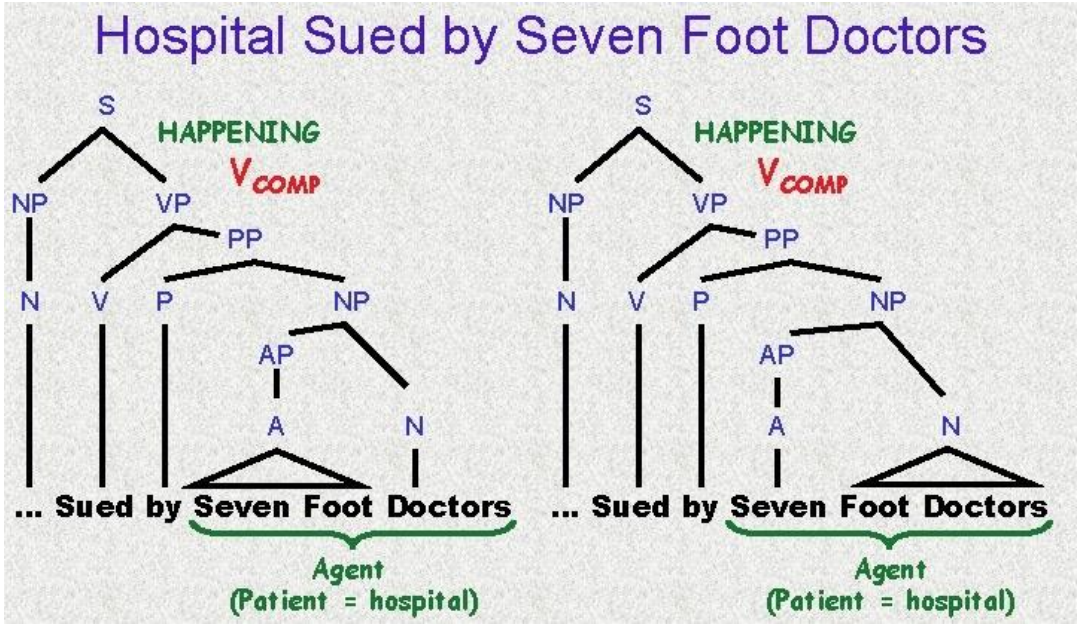
2. Q: Did your mother call your aunt last night?
A: I'm sure she must have.

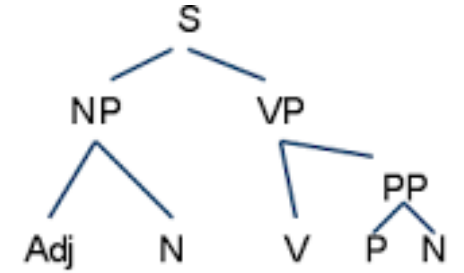
Ambiguities

News Headlines

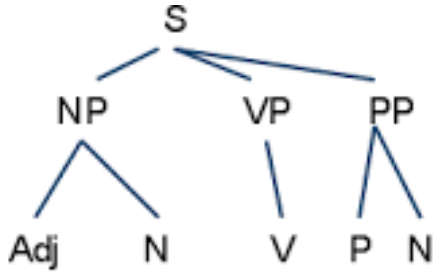
1. **Hospitals Are Sued by 7 Foot Doctors**
2. **Stolen Painting Found by Tree**







Stolen painting found by tree



Stolen painting found by tree

Ambiguity is pervasive

Find at least 5 meanings of this sentence:

I made her duck

- I cooked duck for her
- I cooked duck belonging to her
- I created the (artificial) duck, she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into a duck

Ambiguity is pervasive

Syntactic Category

‘Duck’ can be a noun or verb

‘her’ can be a possessive (‘of her’) or dative (‘for her’) pronoun

Word Meaning

‘make’ can mean ‘create’ or ‘cook’

Natural Languages vs. Computer Languages

Ambiguity is the primary difference between natural and computer languages.

Non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either

Segmentation Issues

the New York-New Haven Railroad

the [New] [York-New] [Haven] [Railroad]

the [New York]-[New Haven] [Railroad]

Idioms

- dark horse
- Ball in your court
- Burn the midnight oil

neologisms

- unfriend
- retweet
- Google/Skype/photoshop

Other issues …

- Segmentation issue:

  Multiple segmentation for same sentence..

  Finding out correct segmentation is another issue in NLP

  Eg: New York-New Heaven Railroad.

- Idioms:

We cannot construct the meaning of the page by looking at the meaning of the individual words and trying compose them together.

Eg: burn the midnight oil

not mean by: I am burning the oil at midnight

what it means is that: you are doing hard work.

What we do in NLP?

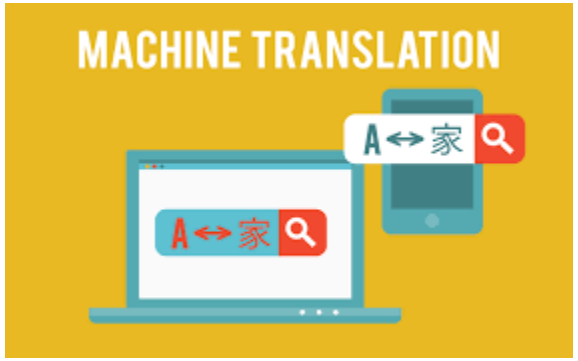
Tools Required

- Knowledge about language
- Knowledge about the world
- A way to combine knowledge resources

How is it generally done?

- Probabilistic models built from language data
  - P("maison" $\longrightarrow$ "house") is high
  - P(I saw a van) > P(eyes awe of an)
- Extracting rough text features does half the job.

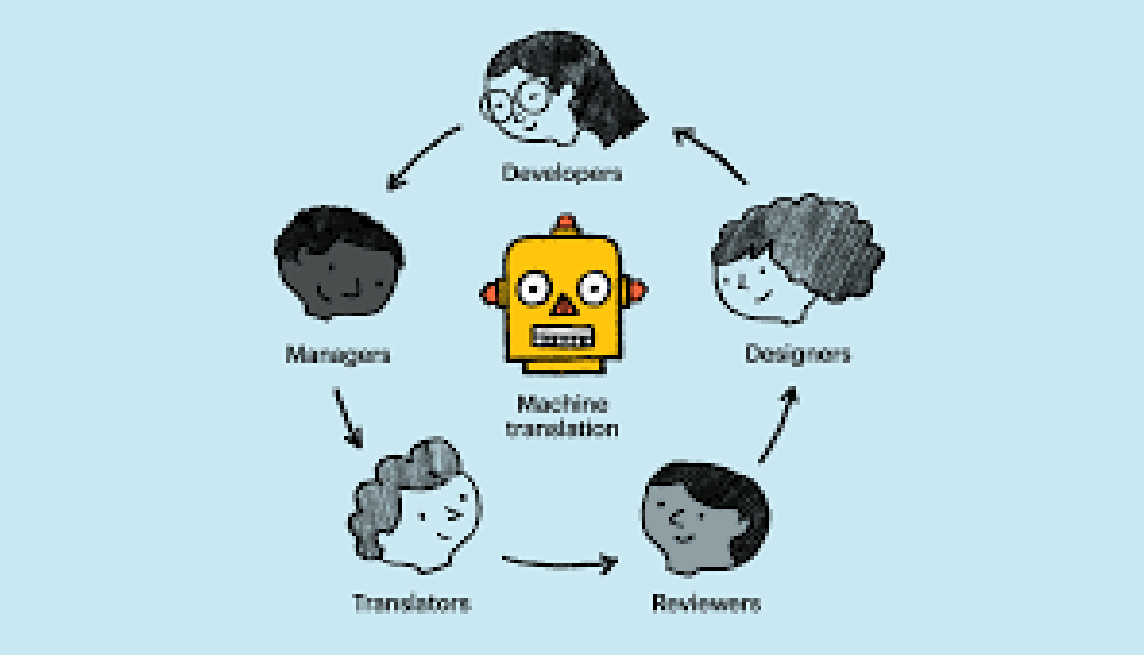
Applications of Nat. Lang. Processing



- Machine Translation
- Database Access
- Information Retrieval
  - Selecting from a set of documents the ones that are relevant to a query
- Text Categorization
  - Sorting text into fixed topic categories
- Extracting data from text
  - Converting unstructured text into structure data
- Spoken language control systems
- Spelling and grammar checkers

Major areas of research and development Applications

- Search Engines
- Machine Translation Systems
- Sentiment Analysis
- Text to Speech
- POS (Parts Of Speech) Tagging
- Automatic Speech Recognition
- Speech to Speech Translation



NLP in other domains

- Bio-medical
- Advertisement
- Politics
- Business Development
- Marketing
- Education

Python for NLP and the Natural Language Toolkit(NLTK)

- **NLTK** is Natural Language Toolkit
- NLTK is a leading platform for building Python programs to work with human language data.
- This toolkit contains packages to make machines understand human language and reply to it with an appropriate response.
- It contains text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning

- Jurafsky and Martin, Speech and Language Processing, 3rd edition
- Jacob Eisenstein, Natural Language Processing (2018)

Thank You