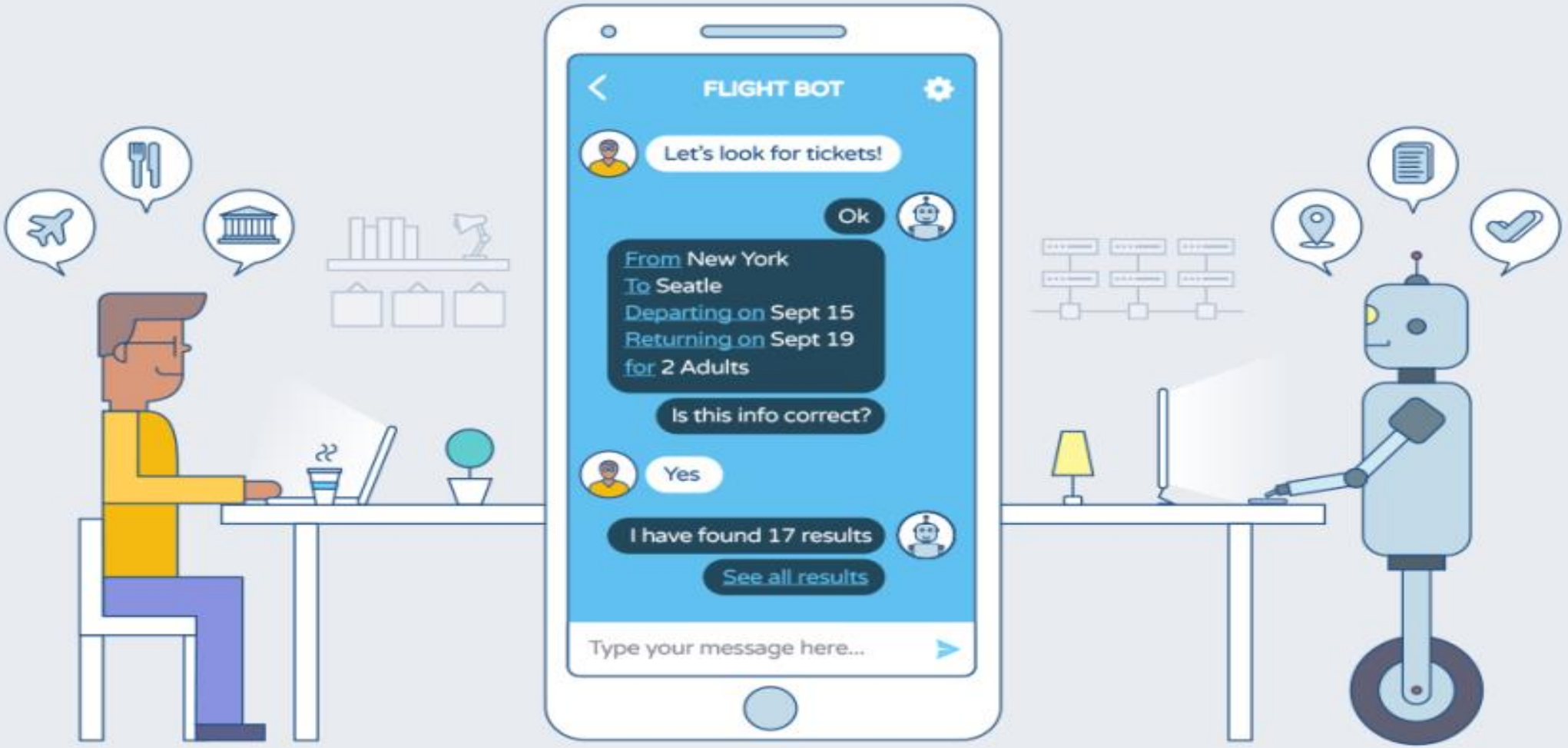
DSE 3155

Natural Language Processing

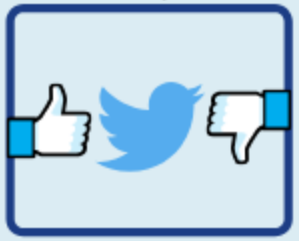

Hey robot, could you please fetch me an apple from the high shelf in the kitchen?
YES. I WILL GO TO THE KITCHEN AND TRY FIND AN APPLE IN THE UPPER SHELF.
fetch
I HAVE REACHED THE KITCHEN, BUT I'M SORRY. THE APPLE IS TOO HIGH FOR ME TO REACH.
WOULD YOU LIKE TO GET IT YOURSELF? OR I CAN BRING YOU A BANANA INSTEAD.



An Application of NLP

Information Retrieval

Sentiment Analysis



Information Extraction

Natural Language Processing (NLP)

Machine Translation

Question Answering

NLP

- Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages
- In particular, concerned with programming computers to fruitfully process large natural language corpora
- Also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)
- A natural language is just like a human language which we use for communication like English, Hindi,..etc.
- Language processing problem can be divided as:
  - – Processing written text, semantic and syntactic knowledge of the language.
  - – Processing spoken language

NLP for Machines

- Analyze, understand and generate human languages just like humans do
- Applying computational techniques to language domain
- To explain linguistic theories, to use the theories to build systems that can be of social use
- Started off as a branch of Artificial Intelligence
- Borrows from Linguistics, Psycholinguistics, Cognitive Science & Statistics
- Make computers learn our language rather than we learn theirs

Challenges in NLP

- Text is the largest repository of human knowledge and is growing quickly.
- It is not an easy task to teach a person or a computer, a natural language.
- The problem is Syntax (rules governing the way in which the words are arranged) & understanding context.
- Computers traditionally require humans to “speak” to them in a programming language that is clear, unambiguous and highly structured.
- Human speech is not always precise.

Components of NLP

Natural Language Understanding

- Taking some spoken/typed sentence and working out what it means
- It helps the machine to understand and analyze human natural language

Natural Language Generation

- Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)
- It is a translator that converts the computerized data into natural language representation.

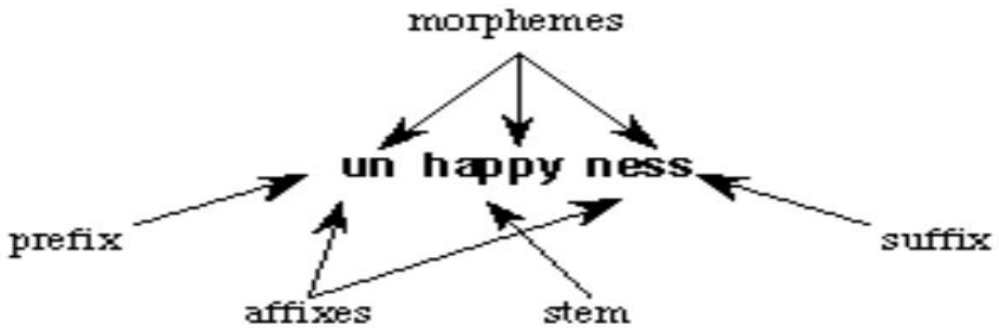
Figure :Components of NLP

Linguistics and Language

- Linguistics is the science of language
- Its study includes:
  - Pronunciation of different speakers which refers to **phonetics**
  - Sounds which refers to **phonology**
  - Study of meaningful components of word formation refers to **morphology**
  - Structural relationship between the words refers to **syntax**
  - Meaning refers to **semantics**
  - How language is used to accomplish task refers to **pragmatics.**
  - Linguistic units larger than a single utterance refers to **Discourse.**

Morphological and Lexical Analysis

- Morphology is the study of the structure and formation of words.
- One of the widespread task here is lemmatizing or stemming which is used in many web search engines. Ex: dog-dogs, run-ran, bus-buses etc.
- Its most important unit is the Morpheme, which is defined as the "minimal unit of meaning”.
- Consider a word like: "unhappiness". This has three parts: un means "not", while ness means "being in a state or condition" and happy is a free morpheme because it can appear on its own.



Morphological and Lexical Analysis

- Lexical Analysis involves identifying and analyzing the structure of words.
- Lexicon of a language means the collection of words and phrases in a language.
- Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words

Syntactic Analysis

- Syntax concerns the proper ordering of words and its affect on meaning
- This involves analysis of the words in a sentence to depict the grammatical structure of the sentence.
- The words are transformed into structure that shows how the words are related to each other.
- In other words, Syntactic Analysis exploit the results of Morphological analysis to build a structural description of the sentence.

Syntactic Analysis

sentence -> noun_phrase, verb_phrase
noun_phrase -> noun
noun_phrase -> determiner, noun
verb_phrase -> verb, noun_phrase
verb_phrase -> verb
noun -> [mary]
noun -> [apple]
verb -> [ate]
determiner -> [the]

Syntactic Analysis

- Main problems on this level are:
  - part of speech tagging (POS tagging)
  - chunking or detecting syntactic categories (verb, noun phrases)
  - sentence assembling (constructing syntax tree)
- Parsing is the process of assigning structural descriptions to sequences of words in a natural language.

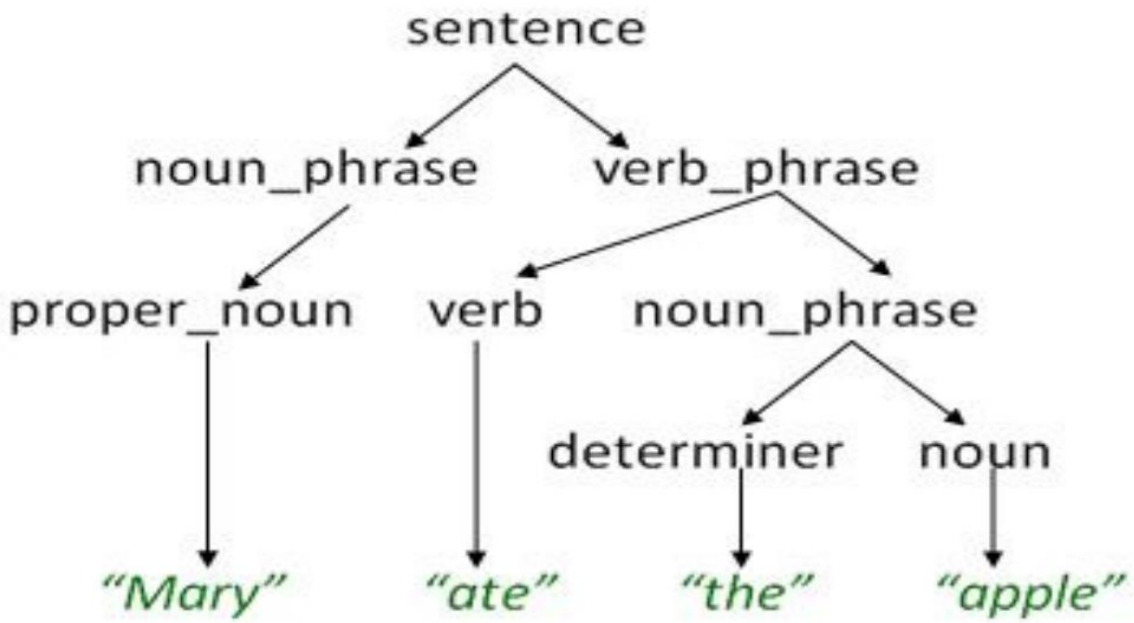
Top-down Parsing

- In this kind of parsing, the parser starts constructing the parse tree from the start symbol and then tries to transform the start symbol to the input.
- The most common form of top-down parsing uses recursive procedure to process the input. The main disadvantage of recursive descent parsing is backtracking.

Bottom-up Parsing

- In this kind of parsing, the parser starts with the input symbol and tries to construct the parser tree up to the start symbol.

Syntactic Analysis



Semantic Analysis

- Semantics concerns the (literal) meaning of words, phrases, and sentences.
- This abstracts the dictionary meaning or the exact meaning from context.
- The structures which are created by the syntactic analyzer are assigned meaning
- E.g.. “colorless blue car”. This would be rejected by the analyzer as colorless blue do not make any sense together.

Basically one needs to complete morphological and syntactical analysis before trying to solve any semantic problem.

Discourse Integration

- The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it.
- In addition, it also brings about the meaning of immediately succeeding sentence

Example:

1. Bill had a red balloon. John wanted it.

   “ It” refers to red balloon and it should be identified like that only .

Pragmatic Analysis

- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
- It means abstracting or deriving the purposeful use of the language in situations.
- Importantly those aspects of language which require world knowledge
- The main focus is on what was said is reinterpreted on what it actually means
- E.g. “Close the window?” should have been interpreted as a request rather than an order

  “Do you know what time it is?” should be interpreted as a request to be told the time

Ambiguity

- Resolving ambiguity is the task performed in all the steps of NLP
- Language is ambiguous: one word, one phrase, or one sentence can mean different things depending on the context
- Ambiguity at multiple levels
  - Word senses: bank (finance or river ?)
  - Part of speech: chair (noun or verb ?)
  - Syntactic structure: I can see a man with a telescope
  - Multiple: I made her duck

Ambiguity

Example: I made her duck

Five different meanings this sentence could have:

i. I cooked waterfowl for her
ii. I cooked waterfowl belonging to her
iii. I created the (plaster?) duck she owns
iv. I caused her to quickly lower her head or body
v. I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity

Ambiguities arises due to :

- Words duck and her are morphologically or syntactically ambiguous in their parts-of-speech (duck can be verb or noun and her can be pronoun or possessive pronoun)
- Word make is syntactically ambiguous:

word make can be transitive (ii) , or it can be ditransitive (v) , or it can take direct object or verb

Resolving Ambiguities

Lexical disambiguation: Resolution of part-of-speech and word sense ambiguities

- Deciding whether duck is a verb or a noun can be solved by part-of-speech tagging.
- Deciding whether make means create or cook is solved by word sense disambiguation

Syntactic disambiguation: represents sentences that can be parsed in multiple syntactical forms

- Deciding whether her and duck are part of same entity or are different entity.

Models and Algorithms

- **Well known Models:**

  State Machines, Formal rule systems, logic, Probability theory, machine learning tools

- **Well known Algorithms:**

State space search and Dynamic Algorithms

- State Machines: are formal models that consists of states, transition among states and an input representation.

  - Variations of basic model are deterministic and non-deterministic finite state automata, finite-state transducers (can write to output devices), weighted automata, Markov models(also called as Markov models), hidden Markov models (Probabilistic component)

Models and Algorithms

- **Formal rule systems:** Regular grammars, regular relations, context-free grammars, feature-augmented grammars.
- State machines and Formal rule systems are used when dealing with the knowledge of phonology, morphology and syntax.
- Algorithms associated with state machines and formal rule systems involve a search through a space of states representing hypothesis about an input. Example: Depth first search, breadth first search

Logic:

Examples : First order logic (Predicate calculus), semantic networks, conceptual dependency

Models and Algorithms

- **Probability theory :**

  -All the methods (state machines, formal rule systems, logic) can be augmented with probabilities.

  - Used to solve many kinds of Ambiguity problems.

- **Machine Learning Models :**

  - Automatically learn the various representations (automata, rule systems, search heuristics, classifiers)

  - Used as a powerful modeling technique where good casual models or algorithms are not available

END