Natural Language Processing

D. Thenmozhi
Associate Professor
SSNCE

Introduction

- Natural Language Processing (NLP)
  - Concerned with the development of computational models of aspects of human language processing
- Reasons
  - To develop automated tools for language processing
  - To gain a better understanding of human communication
- Requires
  - Knowledge of how human acquire, store and process language
  - Knowledge of the world and language

Origins of NLP

- Natural Language Understanding
  - Involves only the interpretation of language

- Natural Language Processing
  - Includes both understanding (interpretation) and generation (production)
  - Includes both speech and text processing (Computational Linguistics)

- Computational Linguistics
  - Concerned with the study of language using computational models of linguistic phenomena
  - Deals with knowledge representations

Computational Linguistics

- Two categories
  - Knowledge driven
  - Data driven

- Knowledge driven
  - Expressed as a set of handcrafted grammar rules
  - Disadv: knowledge bottleneck

- Data driven
  - Presume the existence of a large amount of data and employ some machine learning techniques to learn patters
  - Disadv: dependent on quantity of the data

Applications

- Information Retrieval
  - Information extraction
  - Text summarization
  - Question answering

- Information
  - Speech
  - Image
  - Text

Language and Knowledge

- Language
  - Medium of expression in which knowledge is deciphered
  - Outer form of content

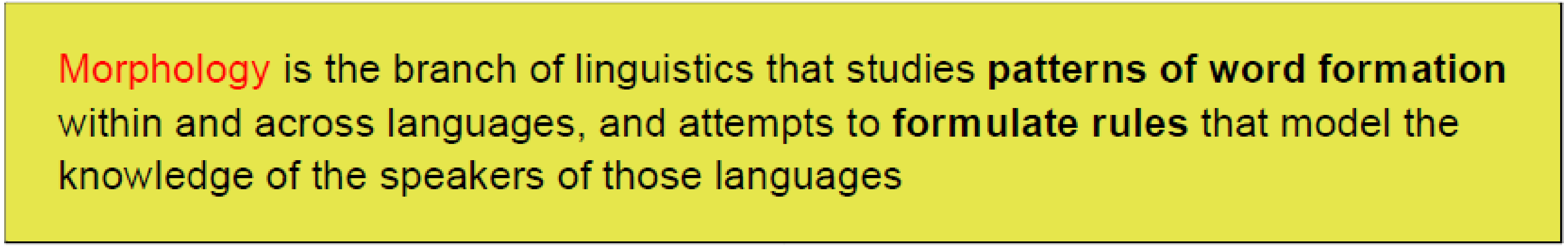
- Knowledge
  - Representation of content
  - Different levels

Knowledge Needed

- Morphology – knowledge of the meaningful *components of words*
- Syntax – knowledge of the *structural relationships between words*
- Semantics – knowledge of *meaning*
- Pragmatics – knowledge of the *relationship of meaning to the goals and intentions of the speaker*
- Discourse – knowledge about *linguistic units larger than a single utterance*
- Phonetics and Phonology – the study of sounds in language

Morphology

- Producing and recognizing the variations of individual words
- The way the word *breaks down into component parts that carry meaning like* singular or plural
- Example: dish, dishes, dishwasher
  - recognizing that *dishes is plural*
  - milk is to milkman -> infer dish is to dishwasher



Morphology is the branch of linguistics that studies **patterns of word formation** within and across languages, and attempts to **formulate rules** that model the knowledge of the speakers of those languages

Syntax

- The sequence of words does not make any sense
  - Ex: I'm I do, sorry that afraid Dave I can't

- Word Order: The knowledge needed to order and group words
  - John hit Bill
  - Bill was hit by John
  - Bill, John hit

- Constituent Structure: *Enraged Cow Injures Farmer With Ax*
  - [Enraged Cow] [Injures] [Farmer With Ax]
  - [Enraged Cow] [Injures] [Farmer] [With Ax]

Syntax is the branch of linguistics that studies the **principles and rules for constructing sentences in natural languages**

Semantics

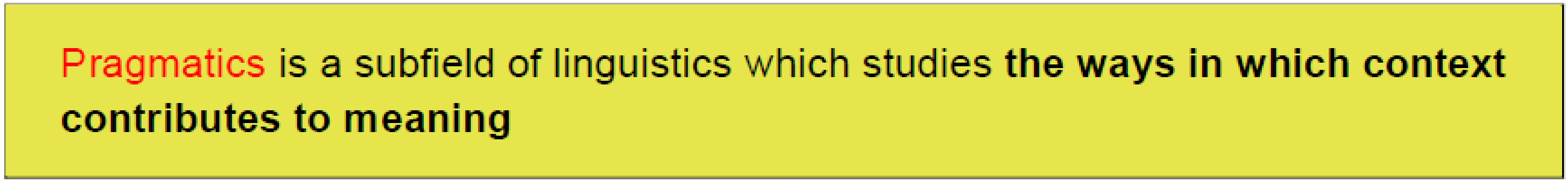
- The study of meaning
  - Ex: How much Chinese silk was exported to Western Europe by the end of the 18th century?

- Lexical semantics – the meaning of all the words
  - Export or silk
  - Europe, century, end

- Compositional semantics:
  - What exactly constitutes *Western Europe as opposed to Eastern or* Southern Europe
  - What does *end mean when combined with the 18th century*

Pragmatic

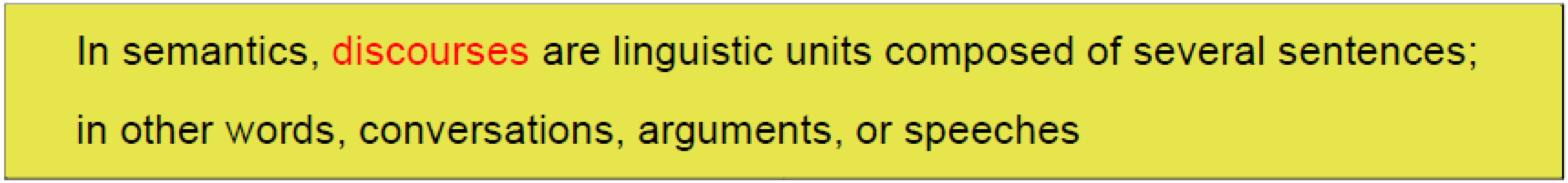
- The kind of actions that speakers intend by their use of sentences is pragmatic or dialogue knowledge
  - Request: Brad, open the door
  - Statement: Brad, the door is open
  - Information question: Brad, is the door open?



Pragmatics is a subfield of linguistics which studies **the ways in which context contributes to meaning**

Discourse

- Makes use of knowledge about how words like that or pronouns like it or she refer to previous parts of the discourse
  - How many states were there in the United States *that year?*
  - Examine the earlier sentence that mentioned about the year
  - For QA, examine the previous questions that were asked



In semantics, discourses are linguistic units composed of several sentences; in other words, conversations, arguments, or speeches

Pragmatic – the influence of context

- Scene 1: Egmore Railway station, Chennai
  - John: Parry's Corner?
  - Passerby: Ground floor, 3rd counter
- Scene 2: Ticket counter, Egmore Railway station
  - John: Parry's Corner?
  - Clerk: Rs.4.00
- Scene 3: Information Booth, Egmore Railway station
  - John: Parry's Corner?
  - Clerk: 4.25 PM, Platform 2

Pragmatic – the influence of context

- Scene 4: On the Train
  - John: Parry's Corner?
  - Passenger: Change at Park Railway Station

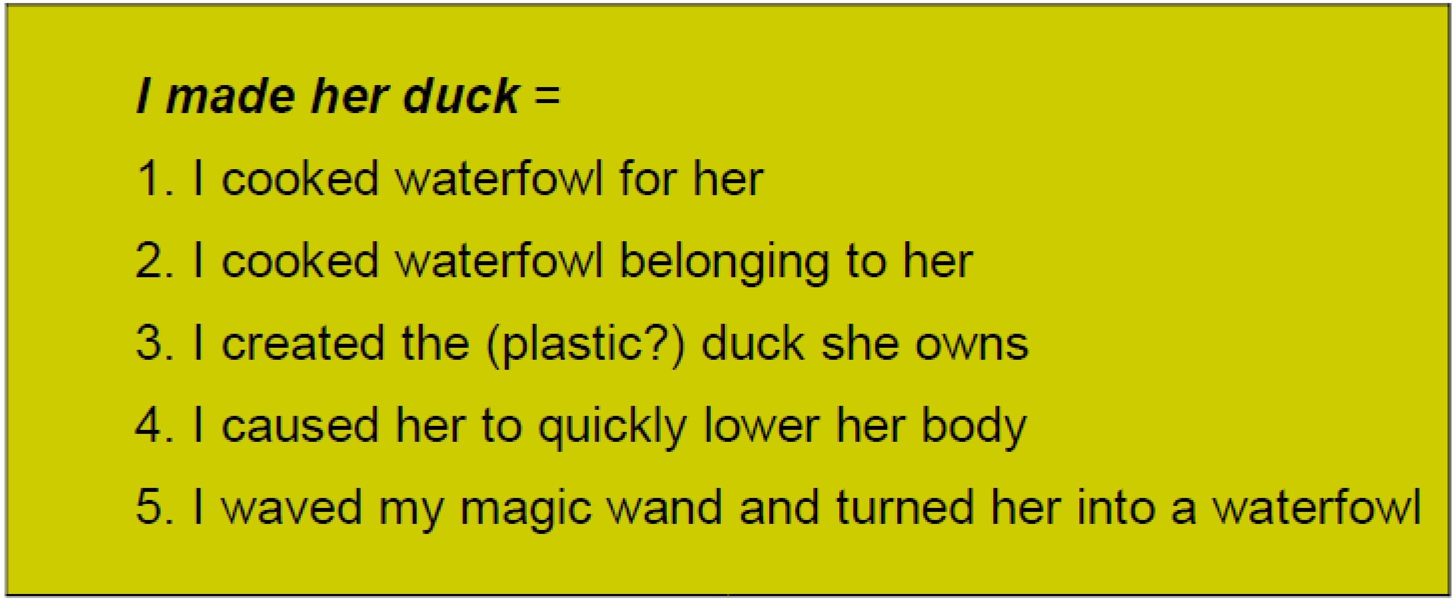
- Scene 5: On the next train, vicinity of Beach Station
  - John: Parry's Corner?
  - Passenger: Opposite to Beach Railway Station

The Challenges of NLP

- Language computing requires precise representation of content which is difficult due to ambiguity and vagueness of natural language
- Inability to capture all the required knowledge
- It is not possible to write procedures that imitate language processing as done by human
- Difficulty in identifying the semantics
  - 9/11
  - While
  - The old man finally kicked the bucket

Ambiguity

- Almost in every level ambiguity is introduced, and one of the main tasks in NLP is to resolve such ambiguities
- Input is said ambiguous in multiple, alternative linguistic structures can be built for it



I made her duck =

1. I cooked waterfowl for her
2. I cooked waterfowl belonging to her
3. I created the (plastic?) duck she owns
4. I caused her to quickly lower her body
5. I waved my magic wand and turned her into a waterfowl

Ambiguity

- *duck and her are morphologically or syntactically ambiguous in part-of-speech*
- duck --> a verb or noun, her --> dative pronoun or possessive pronoun
- *make is semantically ambiguous, i.e., make --> create or cook*
- *make is syntactically ambiguous*
  - transitive – taking single direct object (2)
  - ditransitive – taking two objects (5)
  - direct object and a verb – object (her) got caused to perform the verbal
  - action (duck) (4)

Ambiguity

- To decide whether duck is a verb or noun --> part-of-speech tagging
- To decide whether make means *create or cook* --> *word sense* disambiguation
- Resolution of part-of-speech and word sense disambiguation --> lexical disambiguation
- Deciding whether *her and duck are part of the same entity (1&4) or are* different entity (2) --> syntactic disambiguation