MCA 321: NLP

(Chapter#1 – Introduction)

Aspects of Natural Language Processing

Dr. Gitanjali Ganpatrao Nikam

Deptt. Of Computer Applications,

NIT Kurukshetra

(Lecture #1)

Introduction

- Definition: The idea of giving computers the ability to process human language.
- Goal: To get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication or simply doing useful processing of text or speech.

Examples

- Conversational Agent/ Dialogue system
- Machine translation
- Question answering

Q/A system

- What does divergent mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England by the end of the 18th century?
- What do scientists think about the ethics of human cloning?

Conversational Agent

- Dave: HAL, open the pod bay door.
- HAL: I'm sorry Dave, I'm afraid, I can't do that.

Knowledge in speech and language processing

- Knowledge of language
- It is the task of speech recognition and speech synthesis.
- It requires knowledge about phonetics and phonology, morphology, structural knowledge, Lexical semantics, Compositional semantics, Pragmatic or dialogue knowledge, Discourse knowledge

Ambiguity

"I made her duck."

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

Part-of-speech tagging

Word sense disambiguation

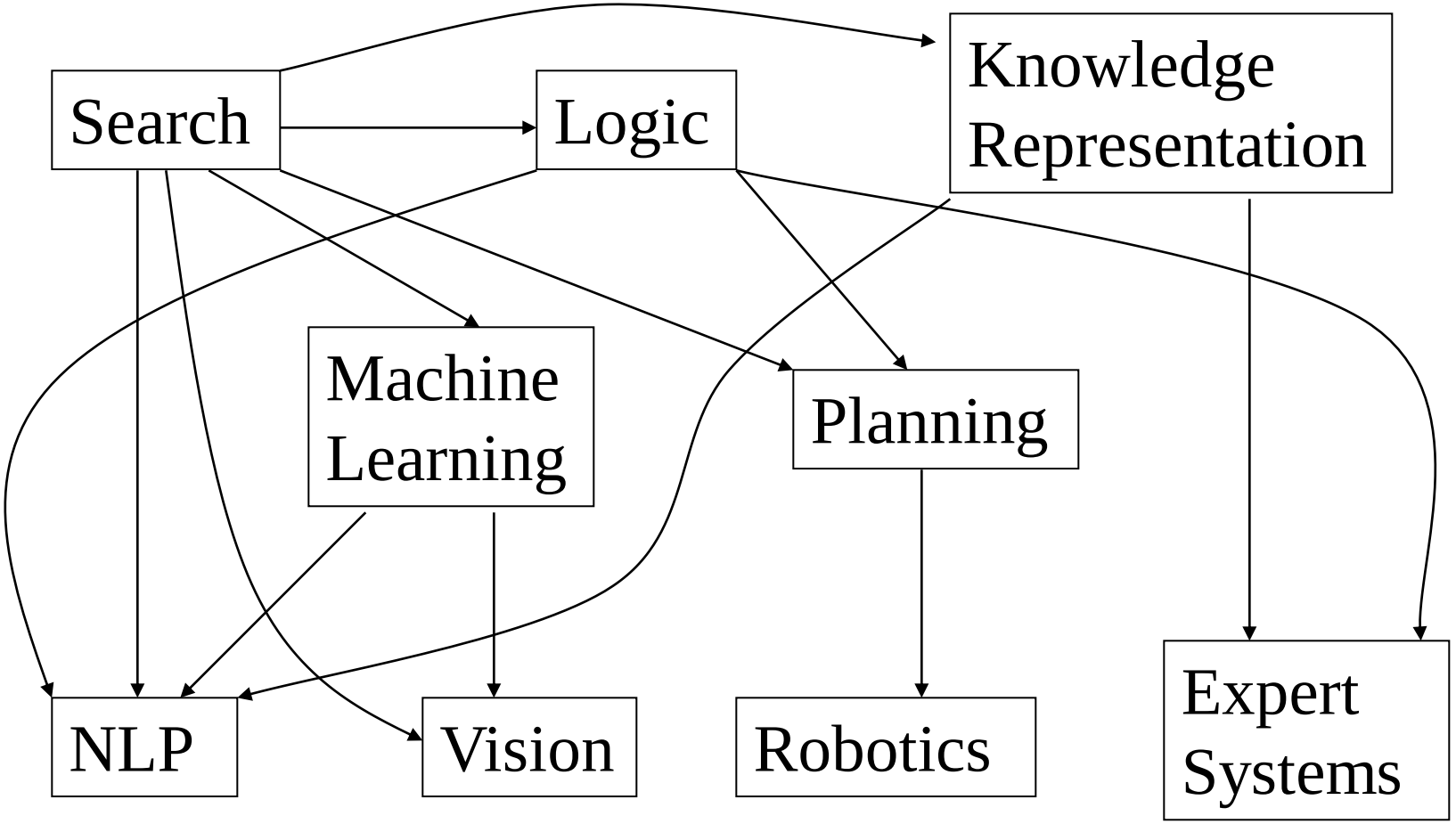
Lexical disambiguation

Syntactic disambiguation

Model and Algorithms

- State m/c
- Rule systems
- Logic
- Probabilistic models
- Vector-Space Models

Perpectivising NLP: Areas of AI and their inter-dependencies



AI is the forcing function for Computer Science

Course Content

- Sound – Biology of Speech Processing; Place and manner of Articulation; Word Boundary Detection; Argmax Based Computations; HMM an Speech Recognition
- Words and word forms – Morphology fundamentals; Morphological Diversity of Indian Languages; Morphology Paradigms; Finite State Machine Based Morphology; Automatic Morphology Learning; Shallow Parsing; Named Entities; Maximum Entropy Models; Random fields.

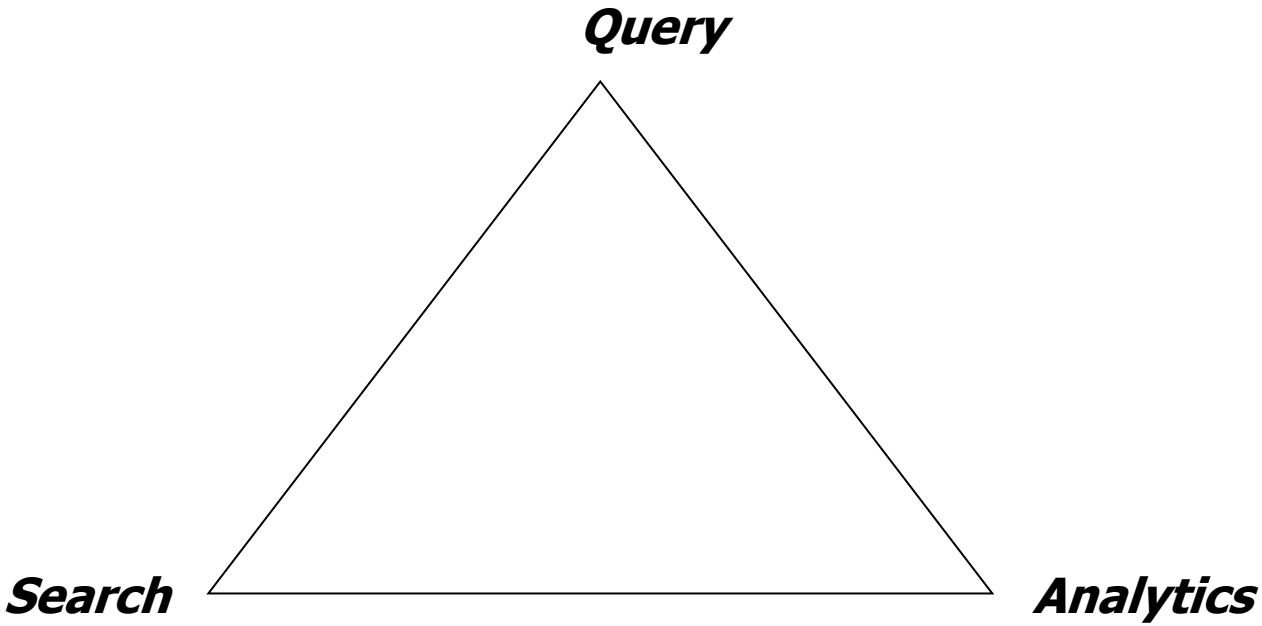
Contd..

- Structures - Theories of Parsing; Robust and Scalable Parsing on Noisy Text as in Web Documents; Hybrid of Rule Based and Probabilistic Parsing; Scope Ambiguity and Attachment Ambiguity resolution
- Meaning- Lexical knowledge networks, Wordnet Theory; Indian Language wordnets and Multilingual Dictionaries; Semantic Roles; Word Sense Disambiguation; WSD and Multilinguality; Metaphores and coreference

Contd..

- Advanced Topics/Applied NLP- Sentiment Analysis; Text entailment; Robust and scalable Machine Translation; Question Answering in Multilingual Setting; Cross Lingual Information Retrieval (CLIR)

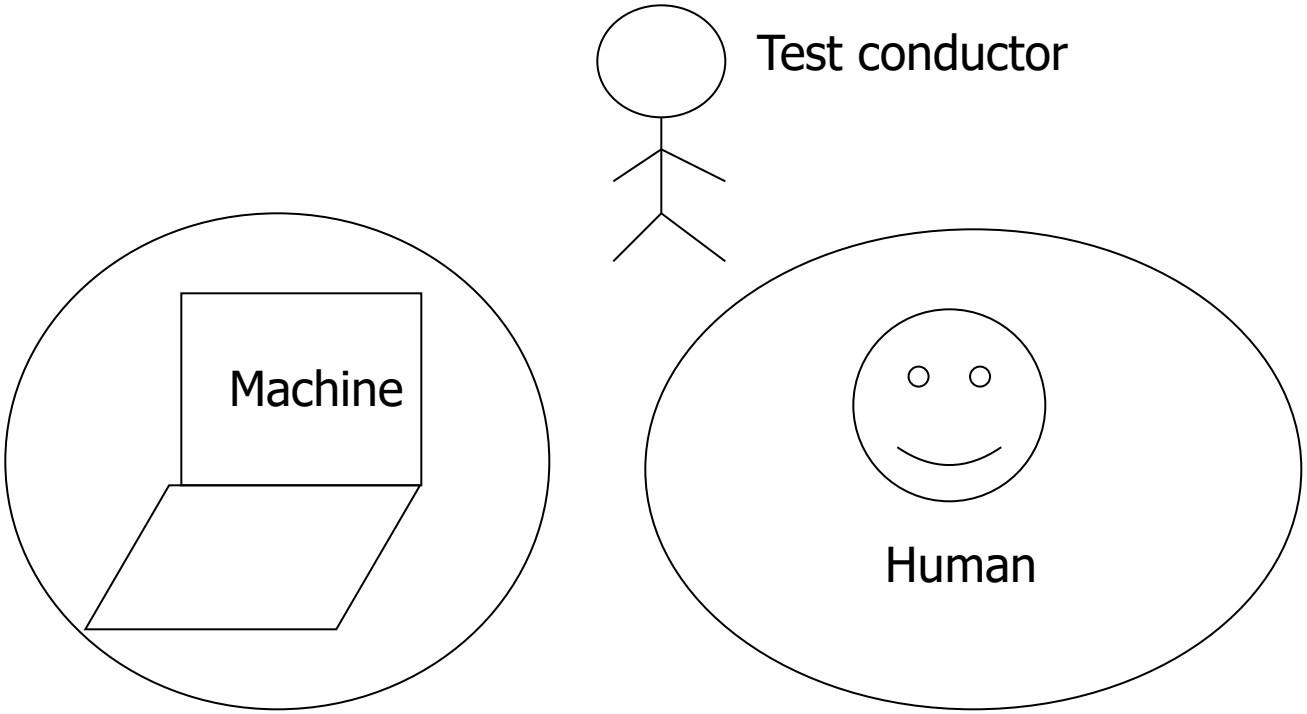
Web brings new perspectives: QSA Triangle



What is NLP

- Branch of AI
- 2 Goals
  - Science Goal: Understand the way language operates
  - Engineering Goal: Build systems that analyse and generate language; reduce the man machine gap

The famous Turing Test: Language Based Interaction

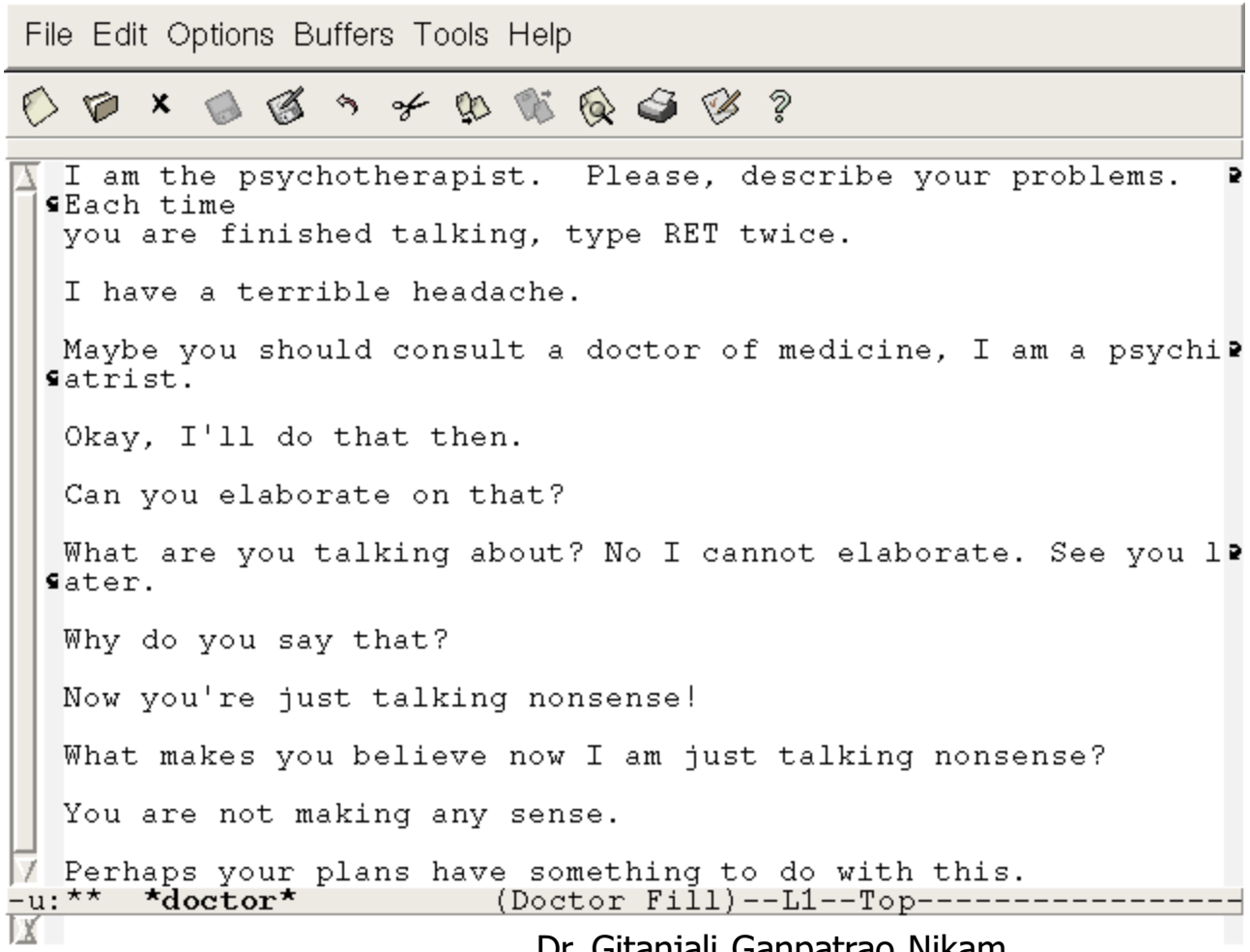


Can the test conductor find out which is the machine and which the human

Inspired *Eliza*

- http://www.manifestation.com/neurotoys/eliza.php3

Inspired *Eliza* (another sample interaction)



Ambiguity

This is what makes NLP challenging:

The Crux of the problem

Stages of language processing

- Phonetics and phonology
- Morphology
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics
- Discourse

Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatraa* vs. *maatra (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come)* or *aaj aayenge (will come today)*
    - *I got [ua]plate- I got a plate; I got up late*
  - Phrase boundary
    - *mtech1 students are especially exhorted to attend as such seminars are integral to one's post-graduate education*
  - Disfluency: *ah, um, ahem etc.*

Morphology

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys);* Gender marking (czar-czarina)
- Verbs: Tense (*stretch-stretched);* Aspect (*e.g. perfective sit-had sat*); Modality (e.g. *request khaanaa→ khaaiie)*
- First crucial first step in NLP
- Languages rich in morphology: e.g., Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing
- A task of interest to computer science: *Finite State Machines for Word Morphology*

Books etc.

- Main Text(s):
  - *Natural Language Understanding*: James Allan
  - *Speech and NLP*: Jurafsky and Martin
  - *Foundations of Statistical NLP*: Manning and Schutze
- Other References:
  - *NLP: A Paninian Perspective*: Akshar Bharti, Vineet Chaitanya and Rajeev Sangal, Prentice Hall, New Delhi.
  - *Statistical Language Learning*: E. Charniac, MIT Press
  - *"Natural language Processing and Information Retrieval",* T. Siddiqui and U. S. Tiwary, Oxford Univ. Press
  - *"Multilingual natural language processing Applications",* Daniel M. Bikel, Imed Zitouni, Pearson.
- Journals
  - Computational Linguistics, Natural Language Engineering, AI, AI Magazine, IEEE SMC

End of Lecture 1