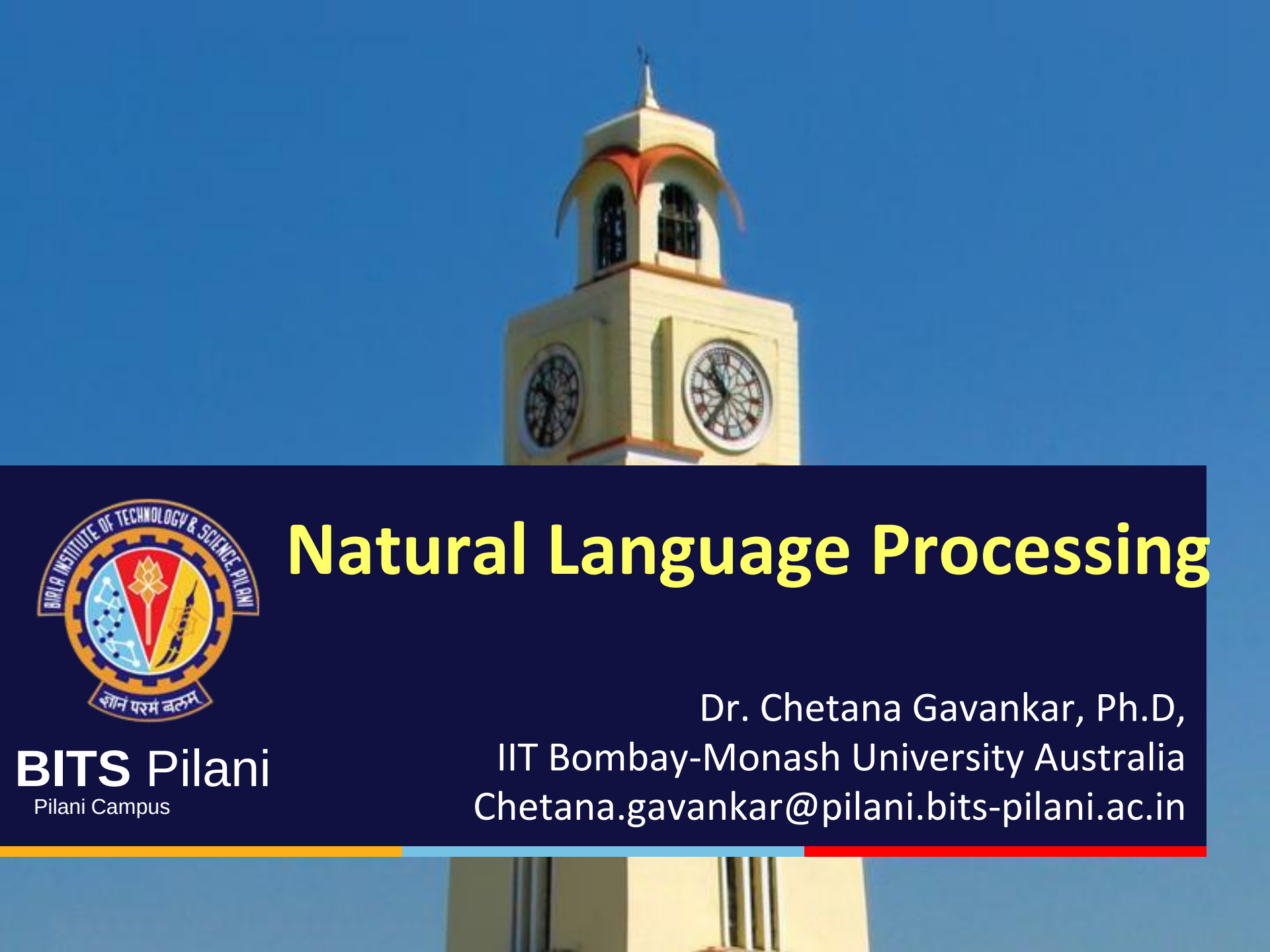


Natural Language Processing

BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in

innovate achieve lead

BITS Pilani

Pilani Campus

Session 1

These slides are prepared by the instructor, with grateful acknowledgement of James Allen and many others who made their course materials freely available online.

Session Content

- Objective of course
- What will we learn in this course?
- Text books and Reference books
- Evaluation Plan
- What is Natural Language Processing?
- Application areas of Natural Language Processing
- Introduction to Natural Language Processing

Objective of course

- To Identify and recall fundamental concepts and techniques in Natural Language Processing (NLP).
- To Explain the computational properties of natural languages and articulate the algorithms commonly utilized for processing linguistic information.
- To Apply basic mathematical models and methods in NLP applications to solve problems and perform tasks.
- To Examine research and development efforts in Natural Language Processing to discern trends, challenges, and opportunities.

What you will learn in this course

- **Vector Semantics**
- **Word Embedding**
- **Language Modelling**
  - N-gram language modeling
  - Neural Language Models
- **Introduction to LLM and Prompt Engineering**
- **Part-of-Speech Tagging**
- **Parsing**
- **Encoder-Decoder Models, Attention and Contextual Embedding's, BERT**
- **Word sense disambiguation**
- **Semantic web ontology and knowledge Graphs**
- **Retrieval Augmented Generation (RAG)**
- **NLP Application- Text Summarization**

Text books and Reference books

T1	Jurafsky and Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, McGraw Hill
T2	Manning and Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA

R1	Allen James, Natural Language Understanding
R2	Neural Machine Translation by Philipp Koehn
R3	Semantic Web Primer (Information Systems) By Antoniou, Grigoris; Van Harmelen, Frank

Evaluation Plan

Name	Weight
Quiz (best 2 out of 3)	10%
Assignment 1 and 2	20%
Mid-term Exam	30%
End Semester Exam	40%

What is Natural Language Processing?

- Analyze, understand and generate human languages just like humans do.
- 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.

Why Study NLP?

- A hallmark of human intelligence.
- Text is the largest repository of human knowledge and is growing quickly.
  - emails, news articles, web pages, IM, scientific articles, insurance claims, customer complaint letters, transcripts of phone calls, technical documents, government documents, patent portfolios, court decisions, ……

The Natural Language Processing (NLP) Market size to grow from USD 15.7 billion in 2022 to USD 49.4 billion by 2027, at a Compound Annual Growth Rate (CAGR) of 25.7% during the forecast period.

Why are language technologies needed?

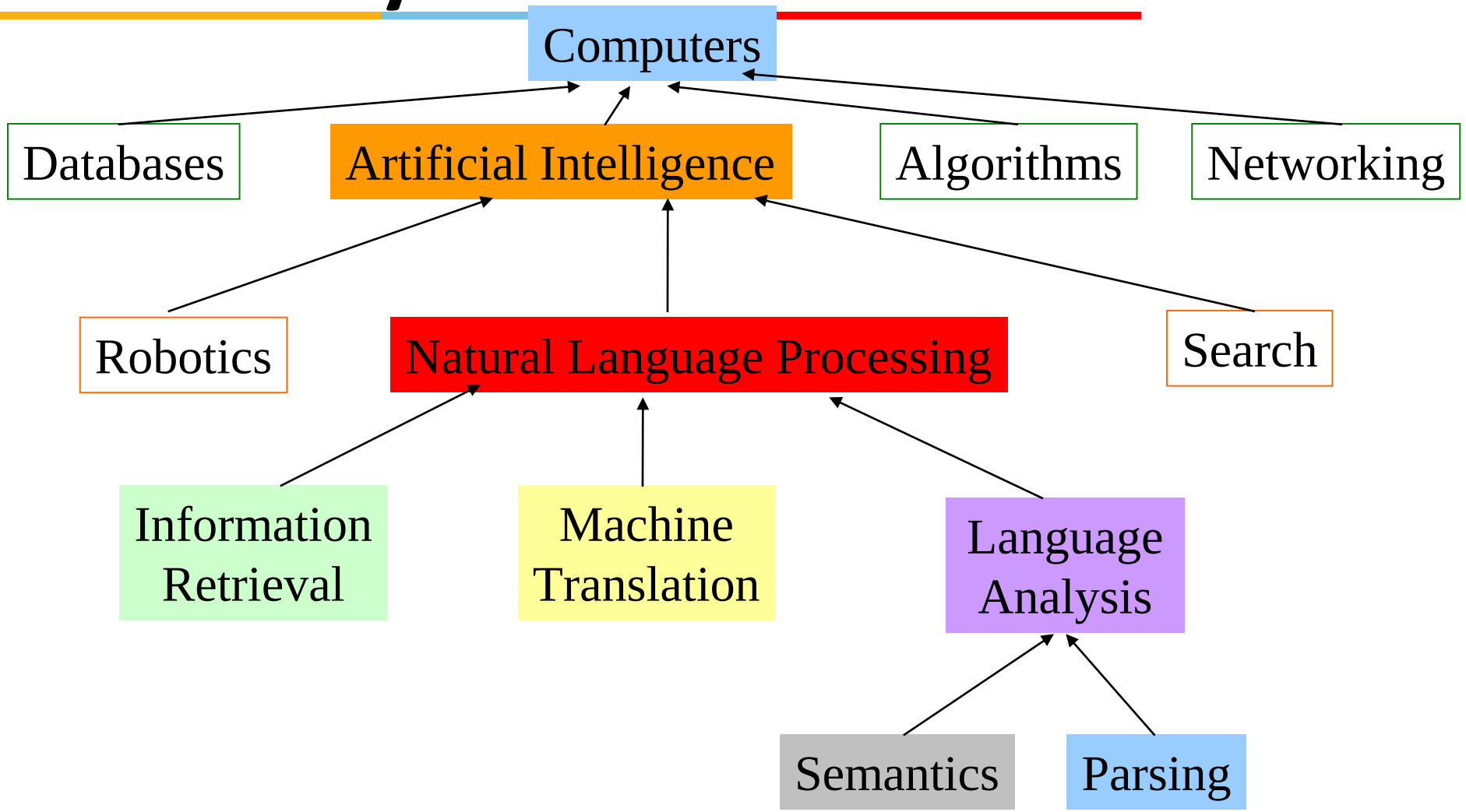
- Many companies make a lot of money if they could use computer programmes that understood text or speech.
  - answering the phone, and replying to a question
  - understanding the text on a Web page to decide who it might be of interest to
  - translating a daily newspaper from Japanese to English
  - understanding text in journals / books and building an expert systems based on that understanding

Dreams or reality??

- Will my computer talk to me like another human ??
- Will the search engine get me exactly what I am looking for??
- Can my PC read the whole newspaper and tell me the important news only..??
- Can my palmtop translate what that Japanese lady is telling me.. ??
- Ahhh.. Can my PC do my NLP assignments ??
- Do you know how our brain processes language ??

Dream come true?

Where does it fit in the CS taxonomy?



Brief history of NLP

- 1966: Eliza
- 1988: Latent Semantic Analysis patent
- January 2011: IBM Watson beats Jeopardy! champions
- October 2011: Apple Siri launches in beta
- April 2014: Microsoft Cortana demoed
- November 2014: Amazon Alexa
- May 2016: Google Assistant

2020 – Conversational Agents

Introduction to Natural Language Processing

innovate achieve lead

- The Study of Language.
- Applications of Natural Language Understanding.
- Evaluating Language Understanding Systems.
- The Different Levels of Language Analysis.
- Representations and Understanding.
- The Organization of Natural Language Processing Systems.

Dave: Open the pod bay doors, HAL.
HAL: I am sorry, Dave. I am afraid I can't do that.
Dave: What's the problem.
HAL: I think you know what the problem is just as well as I do.
Dave: I don't know what you're talking about.
HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

General speech and language understanding and generation capabilities

Politeness:	emotional intelligence
Self-awareness:	a model of self, including goals and plans
Belief ascription:	modeling others; reasoning about their goals and plans

Hal: I can tell from the tone of your voice, Dave, that you're upset.
Why don't you take a stress pill and get some rest.

[Dave has just drawn another sketch of Dr. Hunter].
HAL: Can you hold it a bit closer?
[Dave does so].
HAL: That's Dr. Hunter, isn't it?
Dave: Yes.

Recognition of emotion from speech
Vision capability including visual recognition of emotions and faces
Also: situational ambiguity

To attain the levels of performance we attribute to HAL, we need to be able to define, model, acquire and manipulate

- Knowledge of the world and of agents in it,
- Text meaning,
- Intention

NLP Applications

- Question answering
  - Who is the first Taiwanese president?
- Text Categorization/Routing
  - e.g., customer e-mails.
- Text Mining
  - Find everything that can be done with NLP
- Machine (Assisted) Translation
- Language Teaching/Learning
  - Usage checking
- Spelling correction
  - Is that just dictionary lookup?

Application areas

- Text-to-Speech & Speech recognition
- Healthcare
- Natural Language Dialogue Interfaces to Databases
- Information Retrieval_
- Information Extraction (http://nlp.stanford.edu:8080/ner/process)
- Document Classification
- Document Image Analysis
- Automatic Summarization (https://quillbot.com/summarize)
- Text Proof-reading – Spelling & Grammar_
- Machine Translation_
- **Fake News and Cyberbullying Detection**
- **Monitoring Social Media Using NLP**
- Plagiarism detection
- Look-ahead typing / Word prediction_
- Question Answering System (http://start.csail.mit.edu/index.php)
- Sentiment Analysis (https://komprehend.io/sentiment-analysis)

Open Source NLP Tools

innovate achieve lead

NLTK (Natural Language Toolkit):

Focus: Education, Research, tasks (tokenization, stemming, tagging, parsing).

Best for: Learning NLP concepts, experimentation.

spaCy:

Focus: Production, Performance, Industrial-strength NLP.

Best for: Building applications (fast NER, POS tagging, parsing), pre-trained pipelines.

Hugging Face Transformers:

Focus: State-of-the-art models (BERT, GPT, etc.), Transfer Learning.

Best for: Using and fine-tuning modern deep learning models for various tasks.

Gensim:

Focus: Topic Modeling (LDA, LSI), Document Similarity, Word Embeddings (Word2Vec).

Best for: Unsupervised analysis of large text collections.

LangChain/ Langraph / LlamaIndex:

Focus: Building LLM/Agentic applications, RAG pipelines. (Utilize other NLP tools).

Best for: Creating complex applications combining LLMs with data sources and tools.

Commercial NLP Tools

innovate achieve lead

Google Cloud AI Language: Offers services like sentiment analysis, entity recognition, syntax analysis, and text classification via pre-trained models

Amazon Comprehend: A fully managed AWS service providing APIs for tasks like keyphrase extraction, sentiment analysis, entity recognition, language detection, and topic modeling

Microsoft Azure AI Language: Part of Azure AI services, it includes features for sentiment analysis, key phrase extraction, named entity recognition (NER), language detection, and conversational language understanding

IBM Watson Natural Language Understanding: Provides capabilities to analyze text for extracting entities, keywords, sentiment, relations from unstructured data

OpenAI API: Gives access to powerful LLM (like GPT-4) for a wide range of NLP tasks including text generation, summarization, translation, classification

NLTK Demo

NLTK installation

- The Natural Language Toolkit (NLTK) is a platform used for building programs for text analysis.

Open Anaconda terminal, run

pip install nltk.

Anaconda and Jupiter are best and popular data science tools

In Jupyter, the console commands can be executed by the '!' sign before the command within the cell.

! **pip install nltk**

NLTK book

NLTK discussion forum

https://www.nltk.org/install.html

Pre-processing text data

Cleaning (or pre-processing) the data typically consists of a number of steps:

- **Remove punctuation**
- **Tokenization**
- **Remove stop words**
- Lemmatize/Stem

Why is NLP Big Deal

- L = Words + rules + exceptions..
- Ambiguity at all levels..
- We speak different languages..
- And language is a cultural entity..
- So they are not equivalent..
- Highly systematic but also complex..
- Keeps changing.. New words, New rules and New exceptions..
- Source : Electronic texts / Printed texts / Acoustic Speech Signal.. they are noisy..
- Language looks obvious to us.. But it is a Big Deal ☺!

Why is NLP difficult?

Where is Black Panther playing in Mountain View?

Black Panther is playing at the Century 16 Theater.

When is it playing there?

It's playing at 2pm, 5pm, and 8pm.

OK. I'd like 1 adult and 2 children for the first show.
How much would that cost?

Need domain knowledge, discourse knowledge, world knowledge

Types of Ambiguities

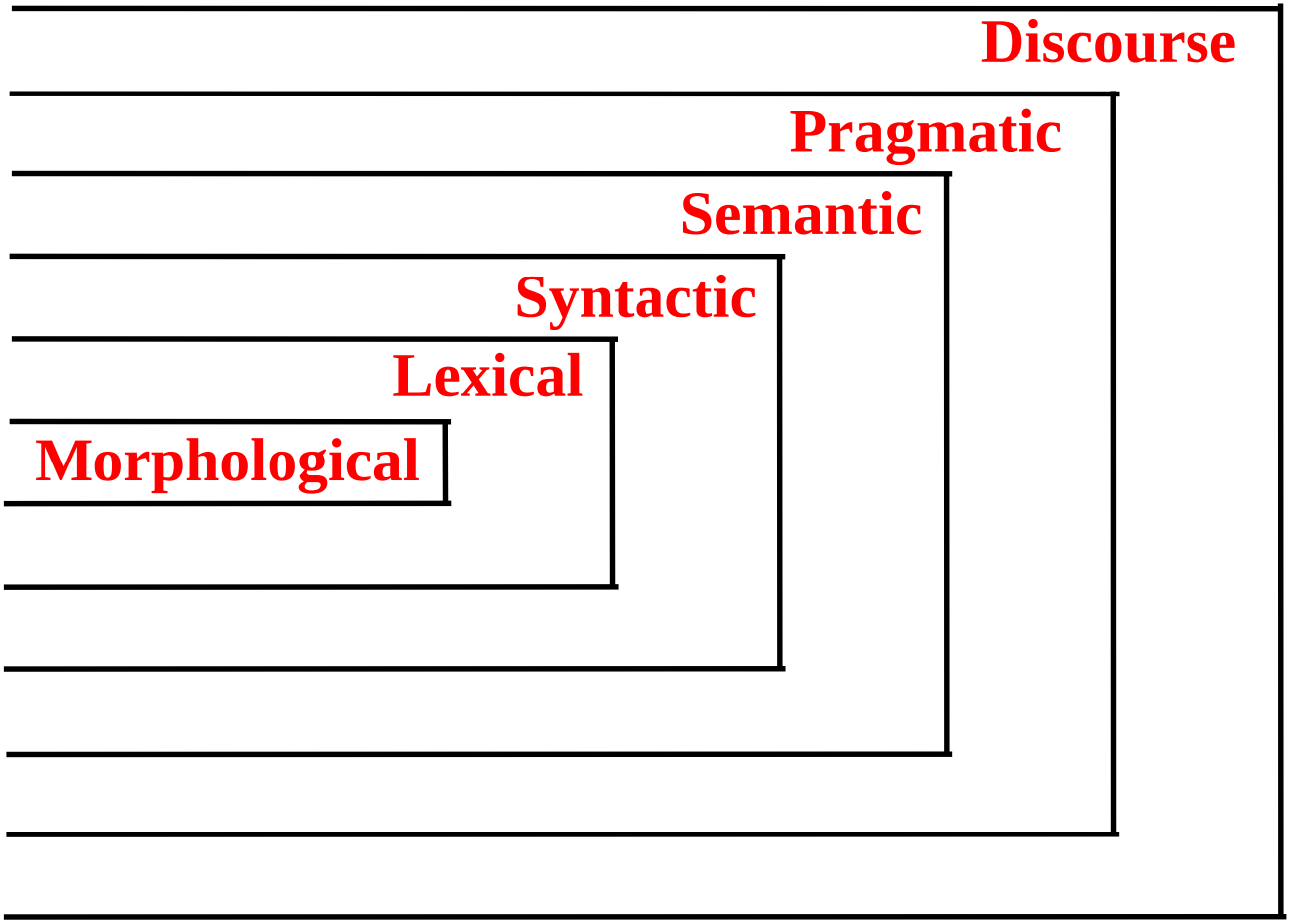
I. Structural Ambiguities
- Namrata thinks she understands me.
- She thinks Namrata understands me.
- Visiting relatives can be nuisance. (two meanings)

II. Grammatical Ambiguities
- I (feminine or masculine) go.
- Can- Noun = container, Can – Modal(auxiliary verb), Can-verb = to can means to pack etc

III. Lexical Ambiguities:

Polysemy Ex: "understand" (I get it)
- Homonymy Ex: Bank= river, financial bank

Different Levels of Language Analysis



Levels of language understanding

Morphological Knowledge : **Concerns how words are constructed from basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language (Ex: "*friendly*" is derived from the meaning of noun "*friend*" and suffix "*-ly*", which transforms noun into adjective)**

Lexical Knowledge :Concerns with listing of words and categorizing them Ex:friendful or beautyship is incorrect lexically. But friendship and beautiful is correct

Syntactic Knowledge :Concerns how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subpart of other phrases Ex: "Large have green ideas nose" is lexically correct but syntactically incorrect.

Levels of language understanding

Semantic Knowledge :Concerns what words mean and how these meanings – combine in sentences to form sentence meanings. This is the study of context-independent meaning. Ex: "Green ideas have large noses" is syntactically correct but semantically incorrect.

Pragmatic Knowledge :Concerns how sentences are used in different situations how use affects the interpretation of sentence

"She cuts banana with a pen" is semantically correct but pragmatically incorrect as it has no useful meaning.

Discourse Knowledge :Concerns how the immediately preceding sentences affect the interpretation of next sentence

Ex: Chetana completed PhD from IIT Bombay. She is a Professor at BITS Pilani.

Context Free Grammar

$$
\begin{aligned}
(1)\ S &\rightarrow NP\ VP \\
(2)\ NP &\rightarrow ART\ ADJ\ N \\
(3)\ NP &\rightarrow ART\ N \\
(4)\ NP &\rightarrow ADJ\ N \\
(5)\ VP &\rightarrow AUX\ VP \\
(6)\ VP &\rightarrow V\ NP
\end{aligned}
$$

Representations and Understanding

Allen 1995: Natural Language Understanding - Introduction

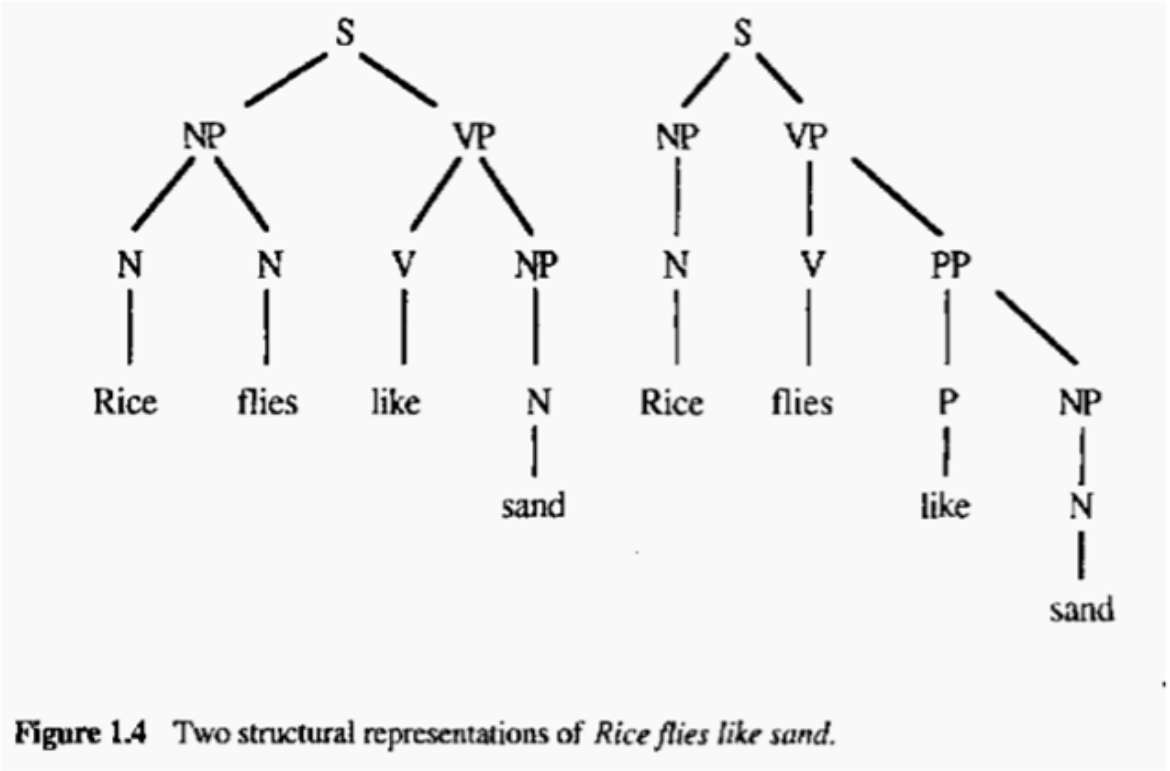


Figure 1.4 Two structural representations of *Rice flies like sand.*

Figure 1.4 Two structural representations of "*Rice flies like sand*".

CFG Rules

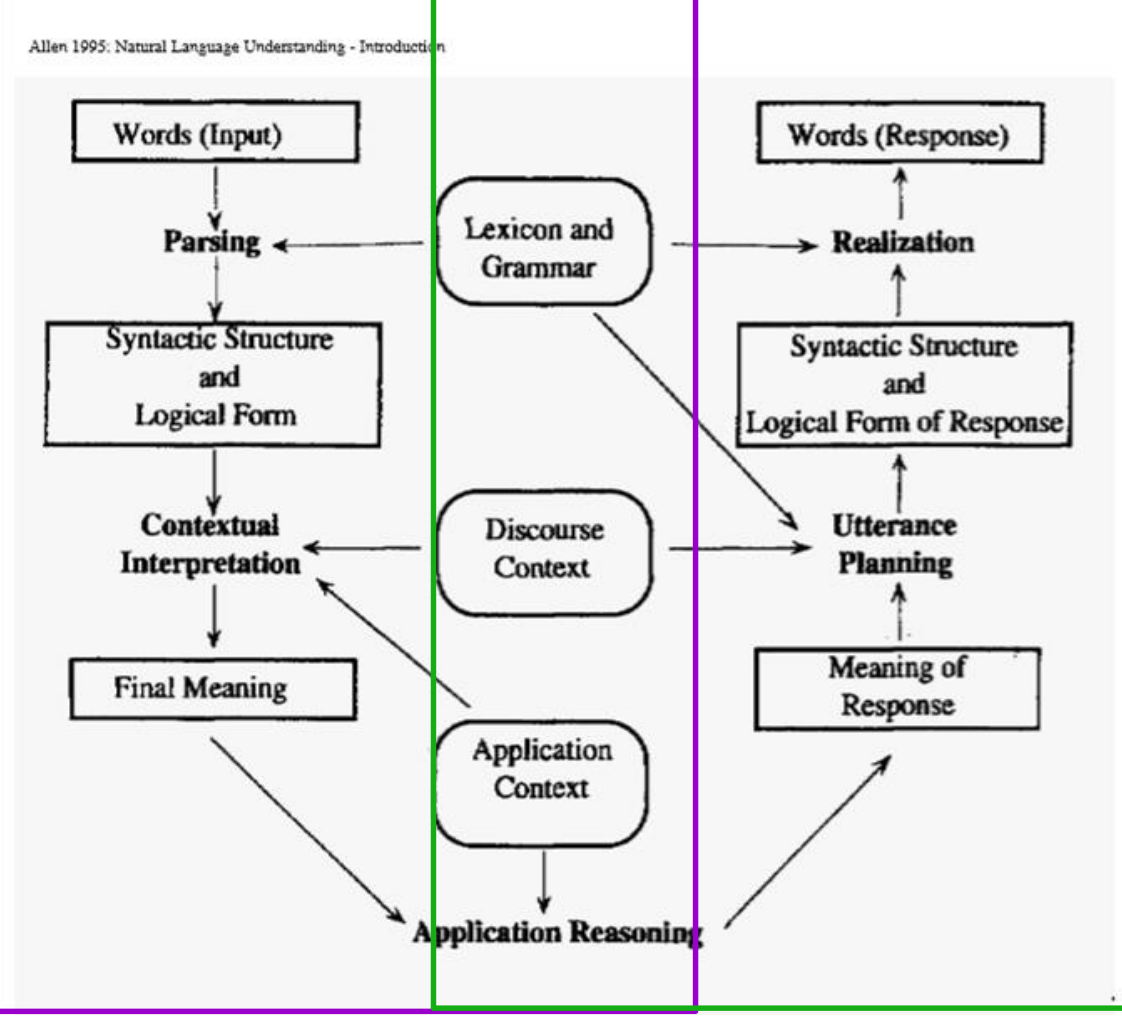
S -> NP VP
NP -> N N
NP -> N
VP -> V NP
VP -> V PP
PP -> P NP

Lexicon

Rice : N
Flies : N, V
Like : V, P
Sand : N

NP – Noun Phrases
VP – Verb Phrases
PP – Prepositional Phrases

The Organization of Natural Language Processing Systems



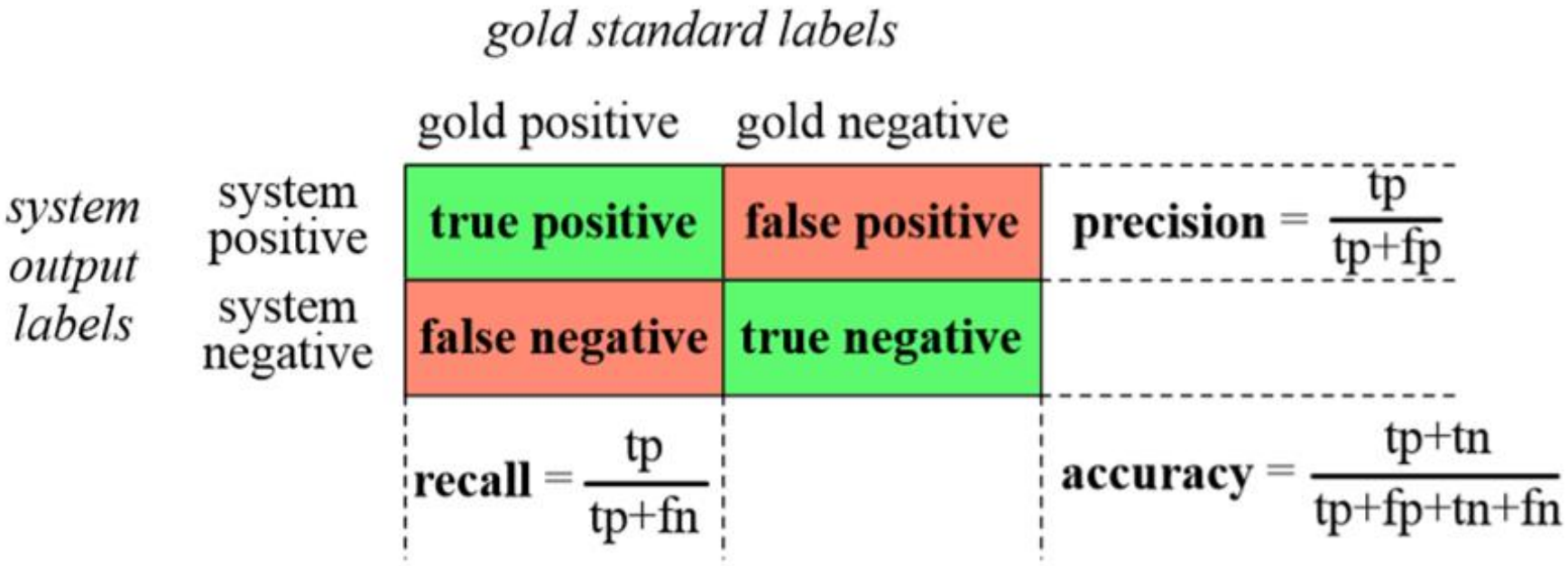
Natural Language Understanding

Natural Language Generation

Evaluating Language Understanding Systems

- What metrics to use?
- How to deal with complex outputs like translations?
- Are the human judgments measuring something real? reliable?
- Is the sample of texts sufficiently representative?
- How reliable or certain are the results?

Contingency Table



Some other evaluation measures

- Manual (the best!?):
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**

- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)

Good References

https://emerj.com/partner-content/nlp-current-applications-and-future-possibilities/

https://venturebeat.com/2019/04/05/why-nlp-will-be-big-in-2019/
https://www.nltk.org/book/

https://www.coursera.org/learn/python-text-mining/home/week/1
https://openai.com/api/
https://analyticssteps.com/blogs/top-nlp-tools

https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html

https://www.cstr.ed.ac.uk/emasters/course/natural_lang.html
https://web.stanford.edu/class/cs224u/2016/materials/cs224u-2016-intro.pdf

https://www.mygreatlearning.com/blog/trending-natural-language-processing-applications/

Dr. Chetana is a Professor and Program Lead in the CSIS department at Work Integrated Learning Programmes Division, BITS Pilani. She has more than 27 years of teaching and industry experience. She did her PhD in Computer Science and Engineering from a joint programme of IIT Bombay and Monash University, Australia. She has been working extensively on different state of art research projects and has been awarded the “Best Industry Aligned Research” at the CSI TechNext India 2019 - Awards to Academia. She has published various papers and is also a reviewer at national and international level peer reviewed conferences and journals. Her areas of expertise include Machine Learning, Natural Language Processing, LLM, Gen AI, Semantic Web, Deep Learning, Text Mining, Big Data Analytics, Information Retrieval and Software Engineering.

Thank you!!