TEXAS
The University of Texas at Austin

I320D – Topics in Human Centered Data Science

Text Mining and NLP Essentials

Week 2: Ambiguity, Multilingualism, Fundamentals layers of NLP, Overview of text corpora and datasets, Regular Expressions

Dr. Abhijit Mishra

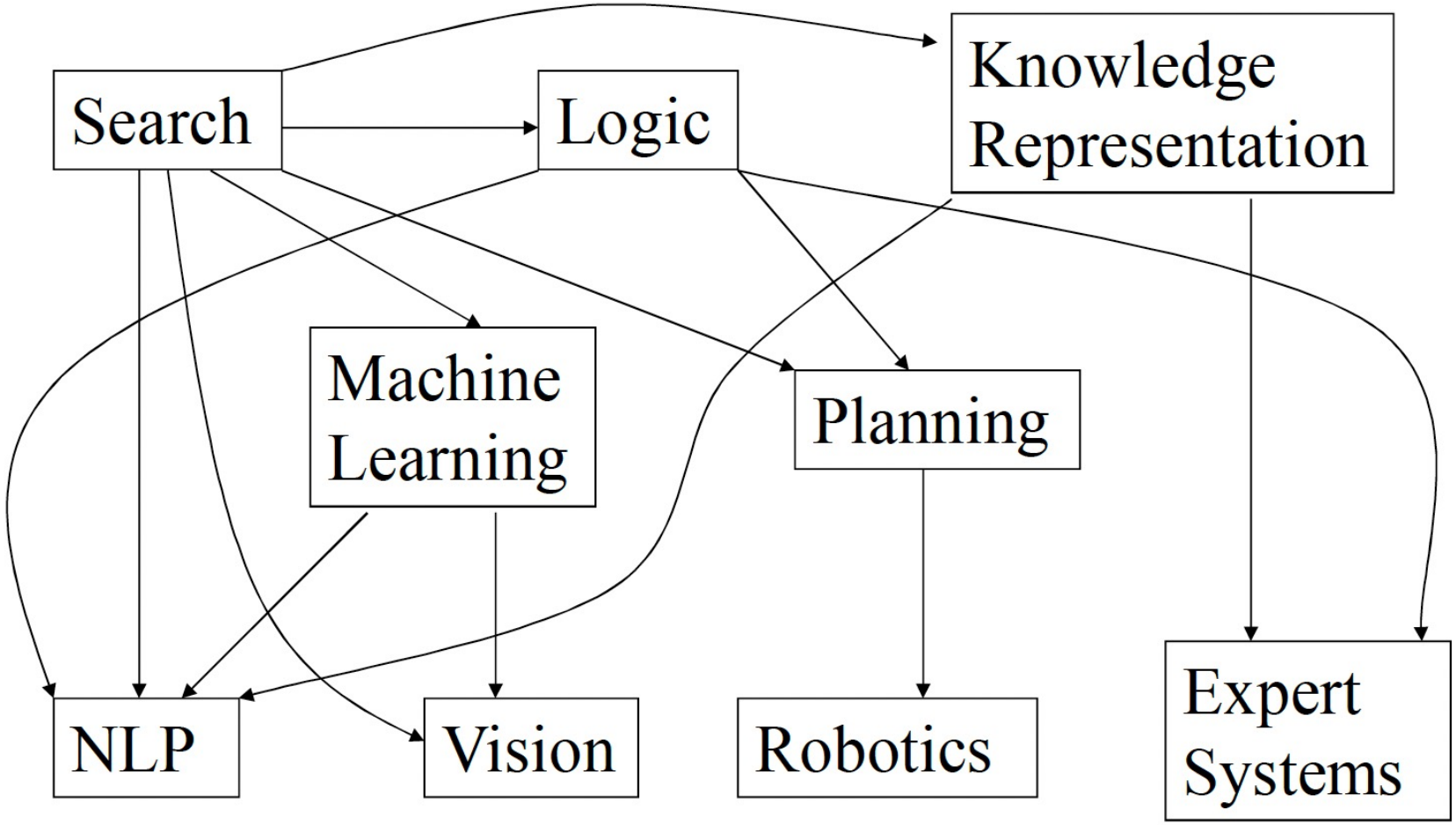
Week 1: Recap

- Lecture:
  - Syllabus Overview,
  - NLP Definition and Layered View
- Practicum:
  - Python Basics and File, String and Document Processing, Frequency Analysis and Visualization of text Data

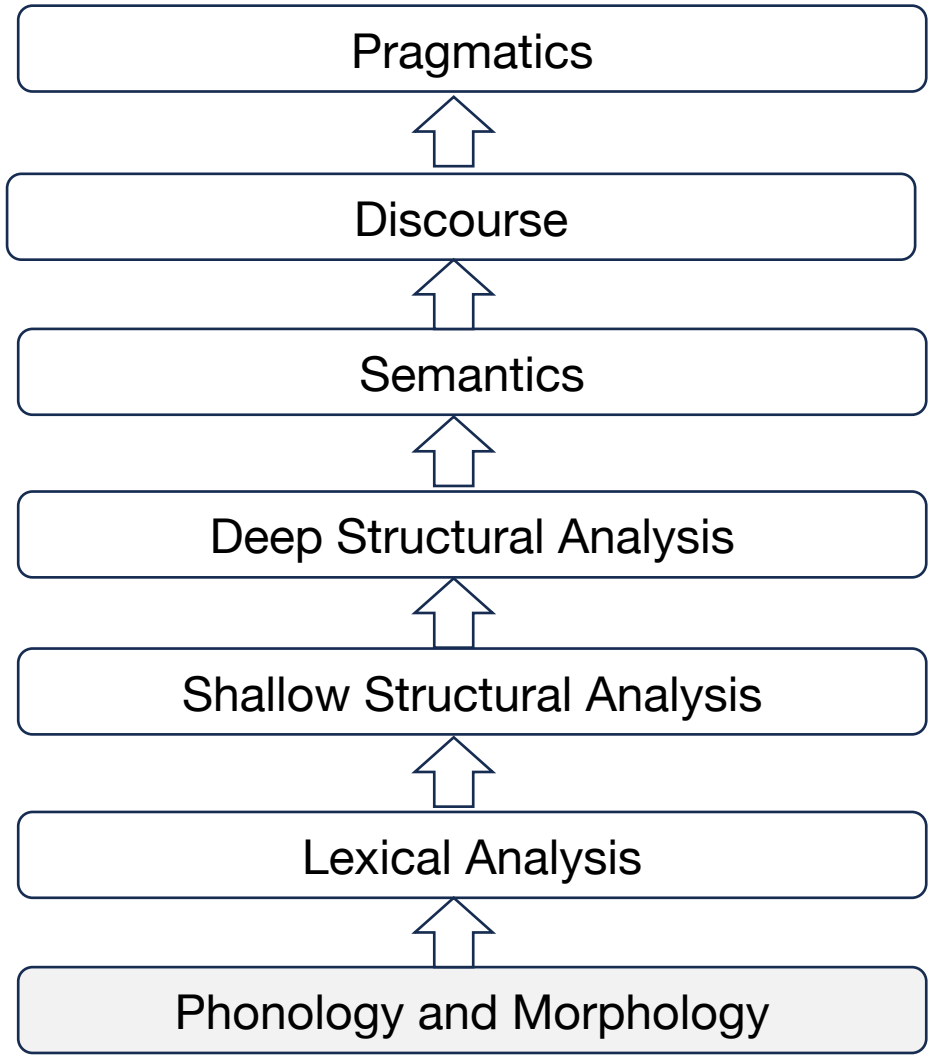
Recap: What is Natural Language Processing?

- Branch of AI
- Two Goals:
  - **Science Goal:** Understand the way language operates
  - **Engineering Goal:** Construct systems that examine and generate text (language) , bridging the divide between humans and machines.

NLP, Areas of AI : Inter-dependencies



Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context confined to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

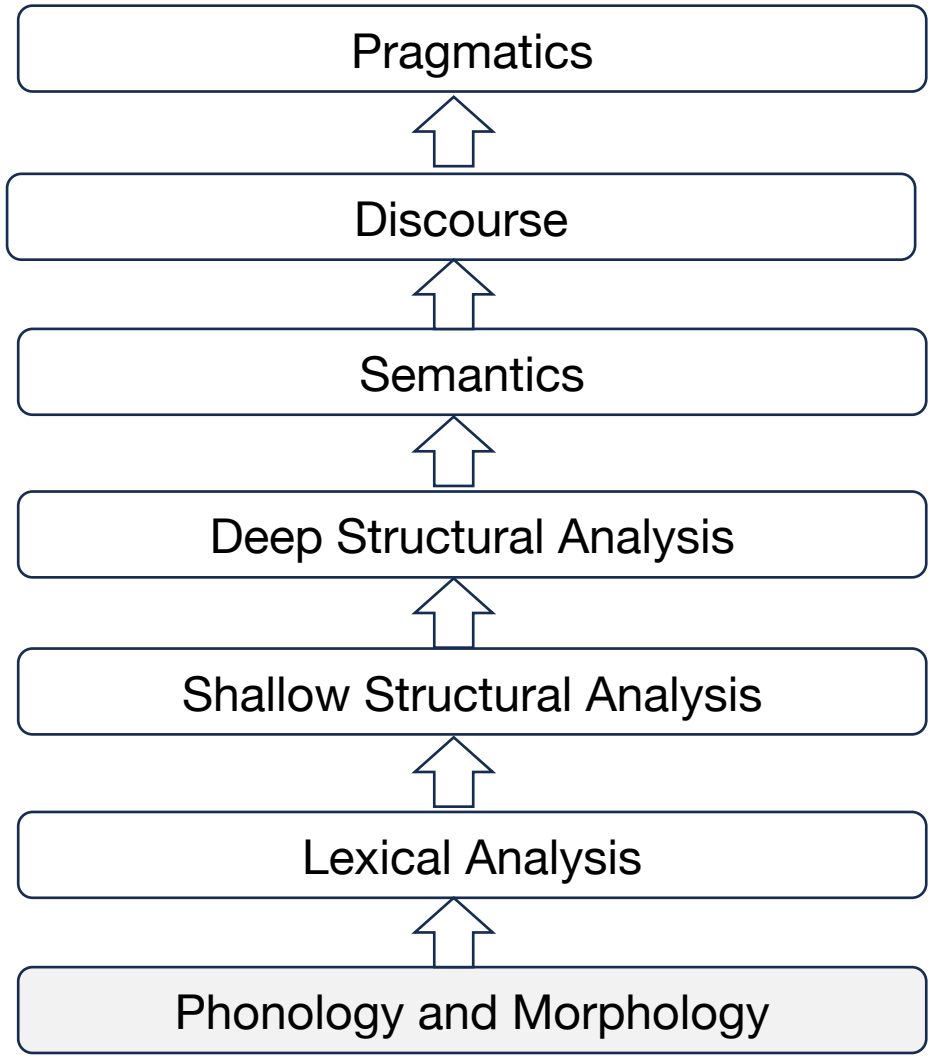
Word meaning without context

Words and surface forms

Ambiguity at the heart of NLP

- Ambiguity is what makes natural languages such as English, Mandarin different from computer languages such as JAVA, Python
- **NLP objective :** to help computers tackle ambiguity at every layer

Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context confined to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

Word meaning without context

Words and surface forms

Morphology

- Word formation rules from root words
- **Nouns**: Plural *(boy-boys); Gender marking (czar-czarina)*
- **Verbs**: Tense (dance-danced); Aspect (perfective: sit-had sat); Modality (Hindi: *khaana – khaaiie) (to eat -> please eat)*
- **Compounds** : *German*: *der Apfelbaum*: *der Apfel* (apple) *+ der Baum* (tree)

Morphology Analysis Applications

- **Direct:**
  - **Text to Speech:** accurately pronouncing words requires understanding verb conjugations, plural forms, and irregular forms
  - **Spell Correction**
    - **Example:** English spell checkers use morphology analysis to correct words with variations such as plurals (e.g., "cats" instead of "cat's") or verb forms (e.g., "running" instead of "runing")
  - **Word Auto-completion**
  - **Search:** Getting **better matches (how?)**
- **Indirect:**
  - Any application that requires higher order processing (e.g., Machine translation, Summarization, Information Extraction)

Ambiguity in Morphology Analysis

- **Ambiguity: no definite patterns (***boy-boys; woman-women)*
- ***How to break words into sub-words / prefixes-roots-suffixes? No fixed rules.***
- **Example: Turkish Word "kitaplarımızdan”**
  - **Valid Splitting 1:**"kitap-lar-ımız-dan”
    - Meaning: "from our books"
    - Components: "kitap" (book), "-lar" (plural), "-ımız" (our), "-dan" (from)
  - **Valid Splitting 2:** "kita-plar-ımız-dan"
    - Meaning: "from our continents”
    - Components: "kıta" (continent), "-lar" (plural), "-ımız" (our), "-dan" (from)

Compounding Compunds the Challenge

- German Language (Morphologically Rich)
  - “**Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz**”
  - ***Breakdown:*** Rindfleisch (beef) + Etikettierung (labeling) + Überwachung (monitoring) + Aufgaben (tasks) + Übertragung (transfer) + Gesetz (law)
  - Meaning: Beef labeling monitoring task transfer law (referring to a repealed German law)

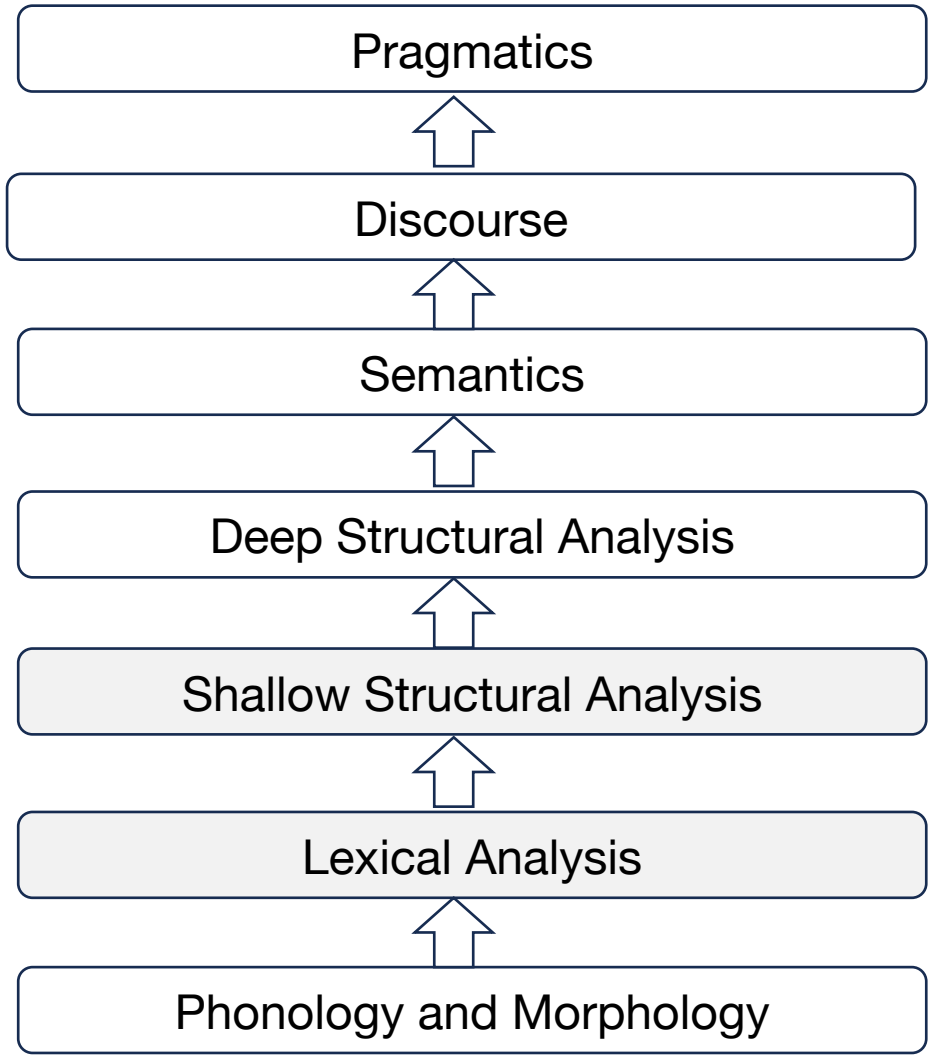
Compounding (…)

- **Hindi: "Sundarvan" (सुंदरवन):** Breakdown: Sundar (beautiful) + Van (forest), but don't breakdown **राजेश** (Rajesh) = Raja (King)+ IshA (God) = God of Kings (only refers to Vishnu, a popular Hindu deity)
- **English:**
  - "Guns and Roses" – Don't break
  - "Flowers and weapons" - Break

Ambiguity in Morphology Analysis

- First crucial step in NLP
- A task of interest to computer science: **Finite State Machines for Word Morphology**

Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context within to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

Word meaning without context

Words and surface forms

Lexical and Shallow Structural Analysis

- Dictionary and word properties
  - dog
    - *Noun (lexical property)*
    - *Takes-‘s’-in-plural (morphological property)*
    - *Animate (semantic property)*
    - *4-legged (semantic property)*
    - *Carnivore (semantic property)*
    - *Don’t spread with COVID (pragmatic property)*

Lexical Analysis Applications

- **Direct:**
  - **Text Classification** based on Topics (e.g., news classification into domains such as sports, politics)
  - **Entity Recognition (**identifying names of people, places, organizations)
    - Washington – person or place?
    - **Hindi: पूजा (Pooja) –** name of a girl **or** the act of worshiping
  - **Sentiment Analysis** given a text, identify the emotional tone expressed by analyzing words
  - **Machine Translation:** Select words in target language based on meaning given in the source language
- **Indirect:** Search Engines**,** Deep Semantic Analysis, Dialog Systems

Lexical Ambiguity

- Ambiguity in parts-of-speech
  - Dog as a noun (animal)
  - Dog as a verb (to pursue)
- Sense disambiguation
  - Dog (as animal)
  - Dog (as a very detestable person)
- Very common in day-to-day communications
  - *“Ground breaking research”*
  - *“India eradicates polio, says WHO”*

“Technological developments bring in new terms, additional meanings/nuances for existing terms”

- **Justify** as in justify the right margin (word processing context)
- **Xeroxed**: a new verb
- **Communifaking**: pretending to talk on mobile when you are actually not
- **Helicopter Parenting:** over parenting
- **Obamagain, modinomics**
- **lol, omg, imo, imho, tbh**

Ambiguity of Multiwords

- *The grandfather kicked the bucket after suffering from cancer.*
- *This job is a piece of cake*
- *Put the sweater on*
- *The 3rd white horse was the dark horse of the race*

Shallow Structural Analysis

- Involves analyzing the surface or syntactic structure of text without delving into deep meaning-based / grammatical relationships
- Includes tasks such as part-of-speech tagging and chunking

Shallow Structural Analysis

- **Two types of tasks**
  - **Part-of-Speech Tagging (POS Tagging)**
    - POS tagging involves assigning grammatical categories (such as nouns, verbs, adjectives, etc.) to each word in a sentence.
    - Sequential flow of information
    - Example:
      - “The cat chased the mouse”.
      - The / DT cat / NN chased / VBD the / DT mouse / NN
      - DT-> Determiner NN->Noun, VBD->Verb in past tense
  - **Chunking:** grouping adjacent words together into "chunks" based on their syntactic structure.
    - Noun Phrase (NP): "The cat," "the mouse"
    - Verb Phrase: “chased”

Applications and Ambiguities

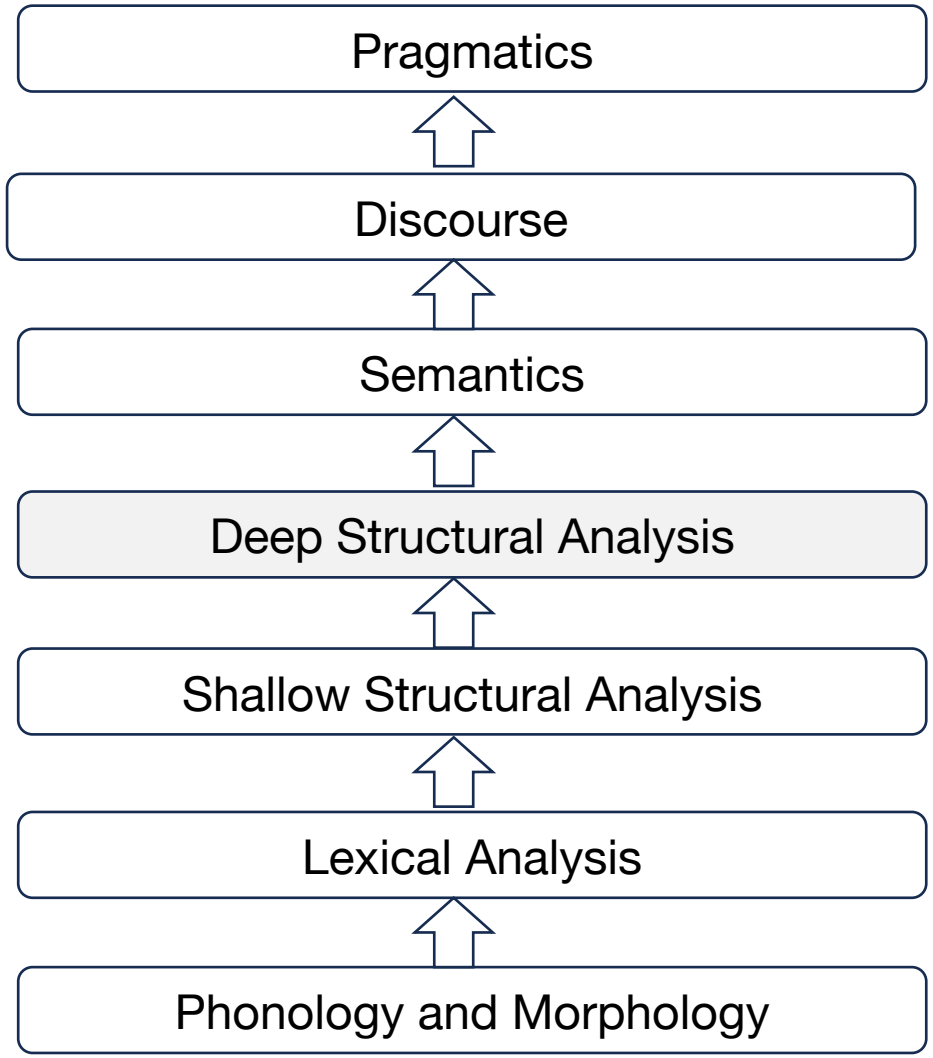
- Shallow Structural Analysis Applications
  - **Direct:** Pattern extraction, Character Analysis in Stories, Entity Extraction and linking, Knowledge Extraction
    - Example: “Barack Obama was born on August 4, 1961”
    - Knowledge: <“Barack Obama”, “birth year”, “1961”>
  - **Indirect:** Search, Translation, Question Answering

Applications and Ambiguities

- Words can have multiple grammatical interpretations, leading to ambiguity in POS tagging. For example, "bank" can be a noun (financial institution) or a verb (to tilt to one side).
- "The wind is strong as they wind their way through the forest."

"The" - Determiner, "wind" (1st occurrence) - Noun (referring to the movement of air), "is" - Verb (to be), "strong" - Adjective, "as" - Conjunction, "they" - Pronoun, "wind" (2nd occurrence) - Verb (to twist or turn), "their" - Possessive Pronoun, "way" - Noun, "through" - Preposition, "the" - Determiner, "forest" - Noun

Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context within to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

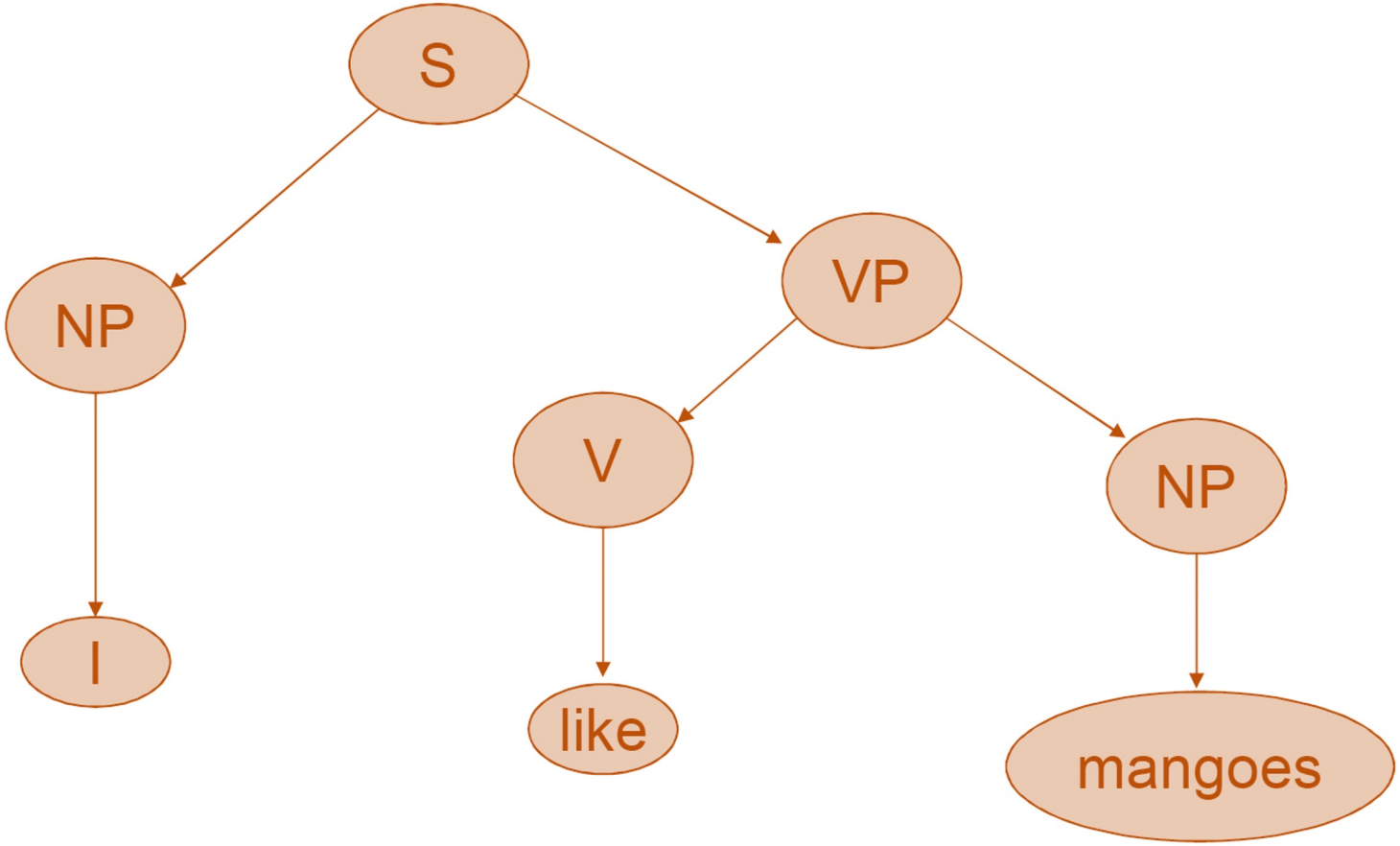
Word meaning without context

Words and surface forms

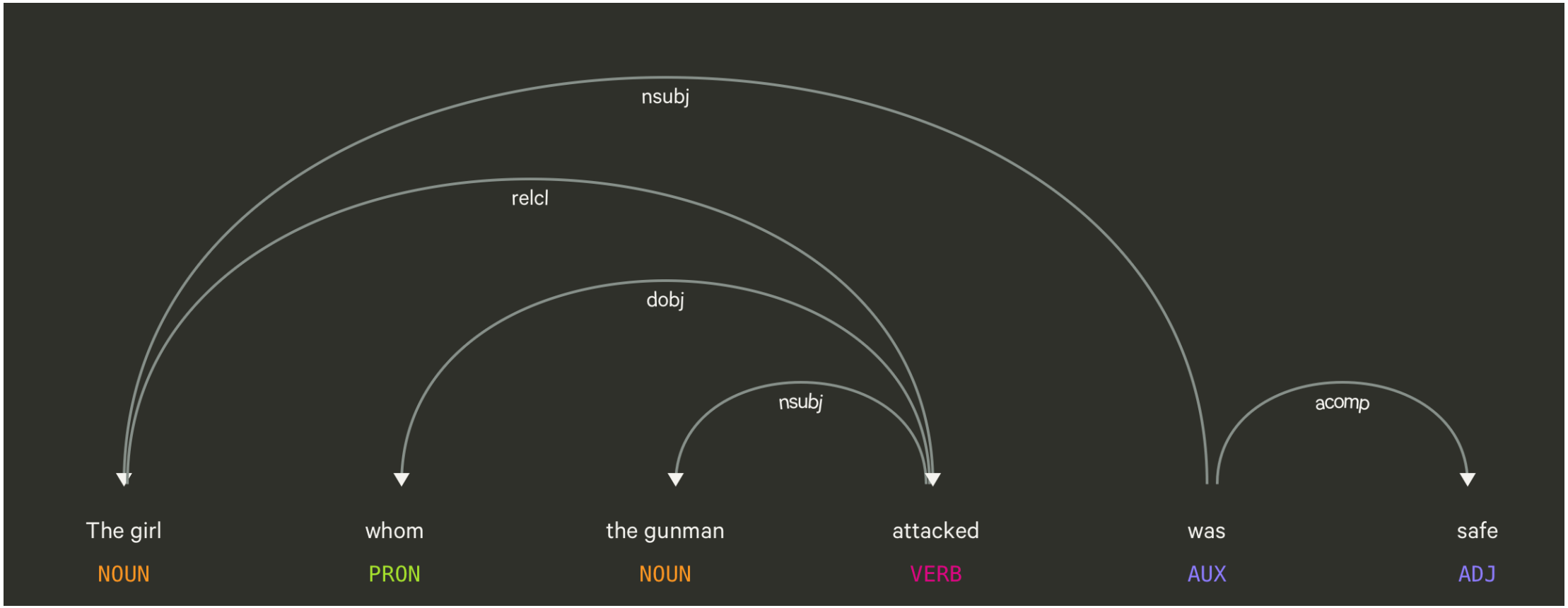
Deep Syntax, Parsing, Dependency between phrases

- Involves a more detailed and thorough examination of the grammatical structure of sentences, typically through constituency parsing and dependency parsing.
- **Constituency Parsing:** breaking down a sentence into its grammatical constituents or phrases
  - Follows **Context Free Grammar** based rules
- **Dependency Parsing:** analyzes the syntactic relationships between words in a sentence by representing them as a directed graph

Structure – Constituency Parsing



Understanding dependencies – Dependency parsing



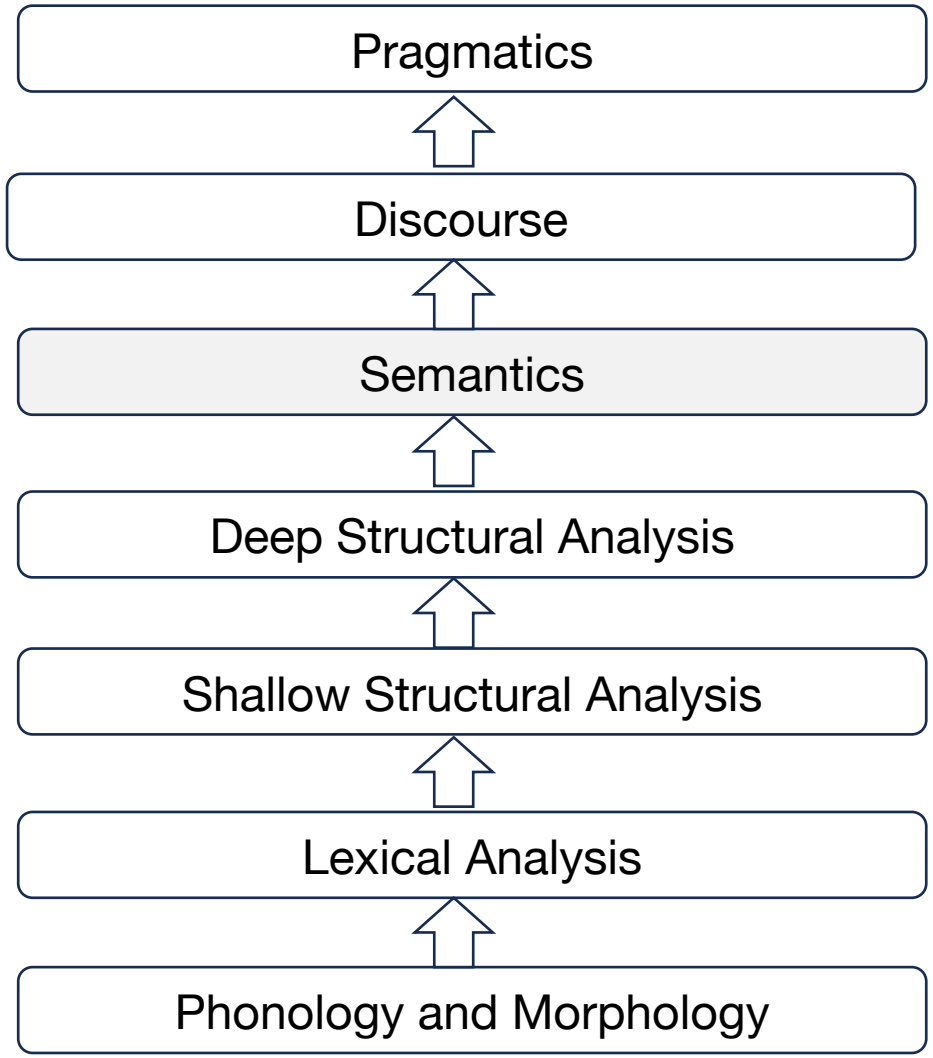
Deep Structure/Syntax Analysis Applications

- Rules / feature extraction for
  - **Machine Translation**
  - **Question Answering**
  - **Grammar Checking**
- Processing large text corpora and extracting patterns / knowledge

Ambiguity in Structure

- Scope:
  - *“The old men and women were taken to safe locations”*
  - *(old men and women) vs. ((old men) and women)*
- Preposition Phrase Attachment
  - *“I saw the boy with a telescope” (unclear who has the telescope)*
  - *“I saw the mountain with a telescope” (who has the telescope) – clear for humans, may not be for the computer*

How humans (and computers) understand text – Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context within to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

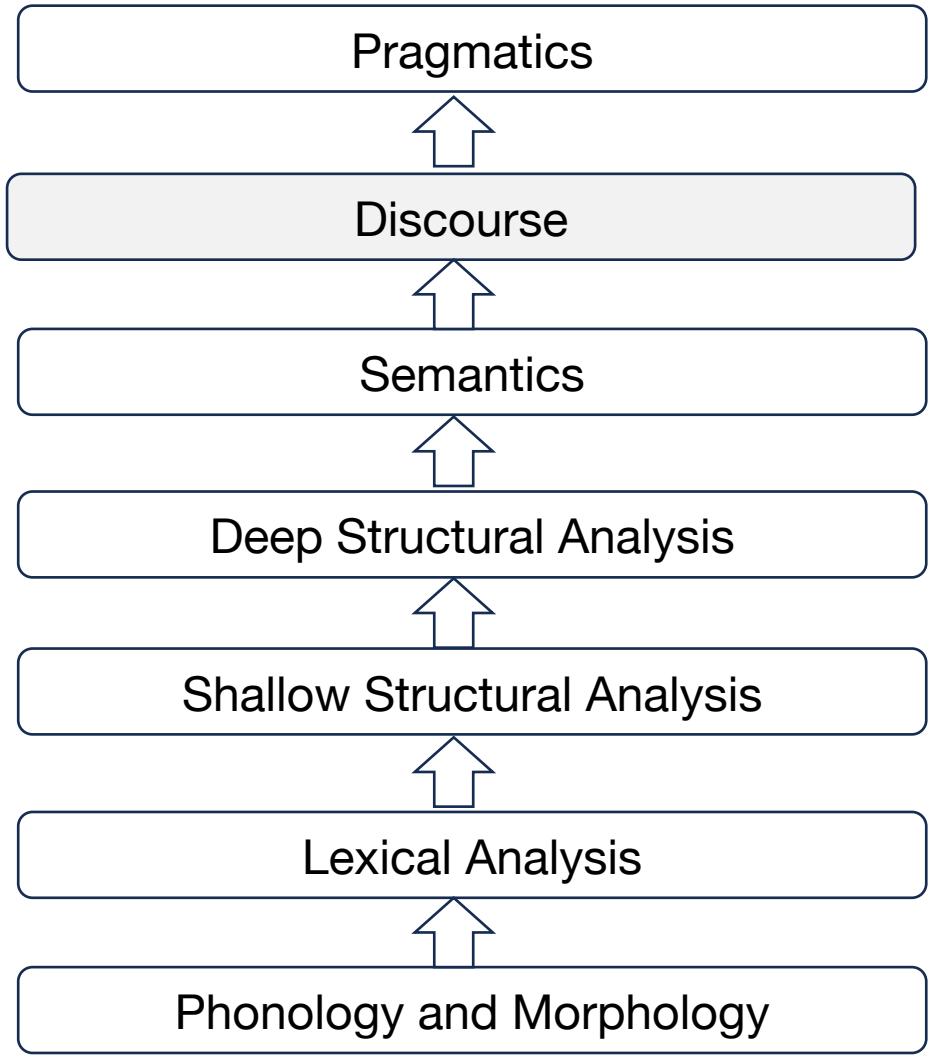
Word meaning without context

Words and surface forms

Semantic Analysis

- Representation in terms of
  - Predicate calculus/Semantic Nets/Frames/Conceptual Dependencies and Scripts
  - *“Abhijit gave a book to Bo” –*
    - *Action : Give, Agent: Abhijit, Object: Book, Recipient: Bo*
  - *Challenge: ambiguity in semantic role labeling*
    - *“Visiting aunts can be a nuisance”*
    - *“Flying planes can be dangerous”*

How humans (and computers) understand text – Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context within to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

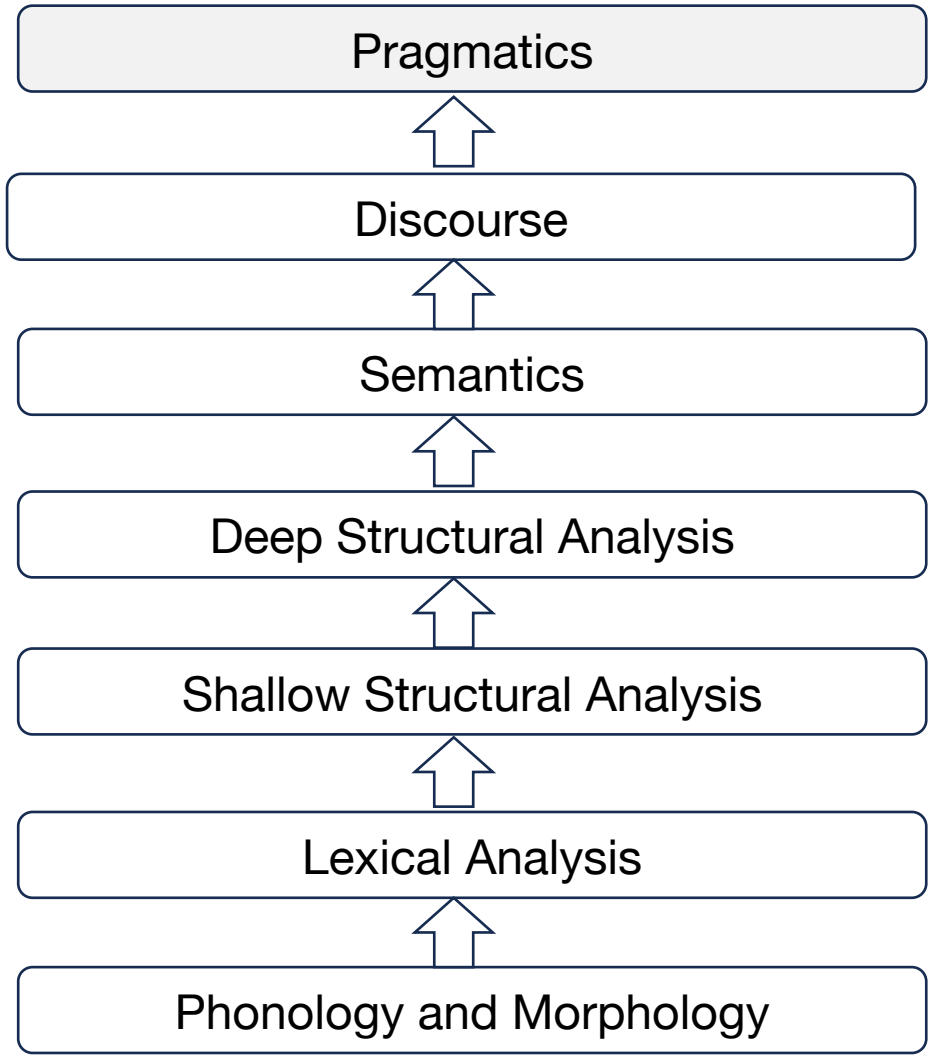
Word meaning without context

Words and surface forms

Ambiguity in Discourse

- Coreference : A challenge
- Binding of referring nouns and pronouns
  - *“The monkey ate the banana, because it was hungry”*
  - *“The monkey ate the banana, because it was ripe and sweet”*
  - *“The monkey ate the banana, because it was lunch time”*

How humans (and computers) understand text – Layered view of NLP



Context beyond presented text (world knowledge

Context from multiple sentences / documents

Context within to a sentence

Deep Syntax, Parses, Dependency between phrases

Syntax, Part-of-speech, Expressions

Word meaning without context

Words and surface forms

Pragmatism

- Very hard problem
- Model user intention
  - *Boy to girl: “Are you a Wi-Fi hotspot? Because I'm feeling a strong connection.”*
  - *Girl: “That’s soooo cheesy”*
- Requires world knowledge

Complexity of Connected Text

"*John was returning from school dejected – today was the math test*"

"He could not control the class"

"Teacher shouldn't have made him responsible"

"After all he is just a janitor"

Source : http://www.cse.iitb.ac.in/~cs626-460-2011

Textual humour

Wordplay:

- *"I'm reading a book on anti-gravity. It's impossible to put down."*
- *"I'm friends with all electricians. We have great current connections."*

Sarcasm:

"Sure, I'd love to help you move this weekend. Because what's more fun than carrying heavy furniture up and down stairs?"

Situational Disparities:

Teacher (angrily): *"did you miss the class yesterday?"*

Student: *"not much"*

John: *"I got a Jaguar car for my unemployed youngest son."*

Jack: *"That's a great exchange!"*

NLP and Multilingualism

- NLP should be non-English centric. Why?
- **Linguistic Variation:** Languages differ in structure; multilingual NLP adapts for processing diversity.
- **Global Interaction:** Multilingual NLP facilitates communication across languages for broader engagement.
- **Cognitive Empowerment:**
  - Translate every form of information and human intelligence into computer understandable form so that machines can “help everyone” alike

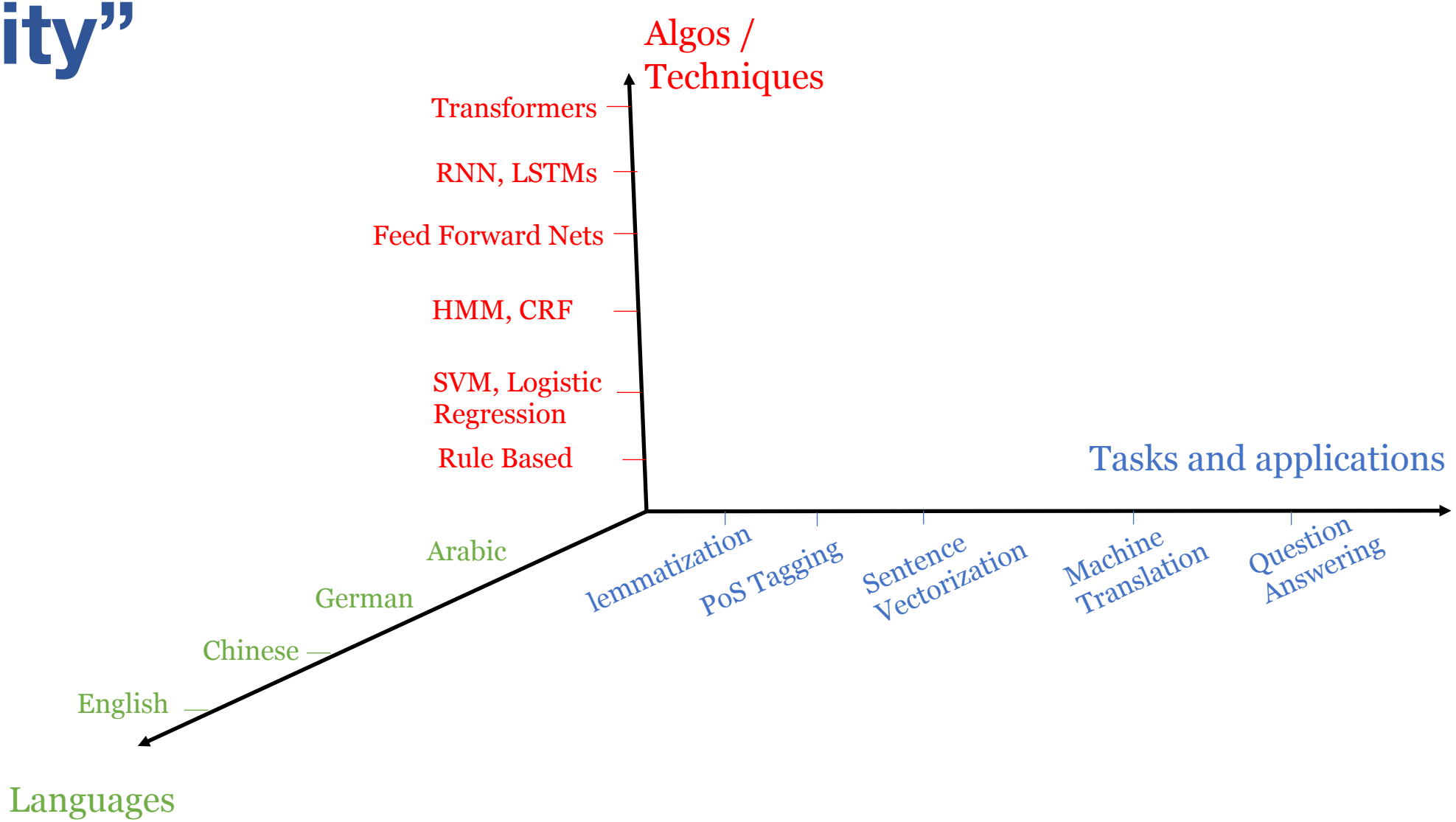
NLP Applications

- Low level:
  - Explaining words and phrases in document
  - Extract phrases, parts-of-speech, parse structure
  - Analyze words and get their root forms
  - Understand document semantics and pragmatics
- High level:
  - Translate sentences / documents
  - Analyze sentiment
  - Summarize
  - Answer Questions

NLP Algorithms

- Rule Based
- Machine Learning based or “data driven”
  - Classifiers: SVM, Logistic Regression
  - Sequence Labelers: HMM, CRFs
- Deep neural network based
  - Feed forward networks
  - Recurrent networks
  - Transformers

Putting everything together: “The NLP Trinity”



Source : http://www.cse.iitb.ac.in/~cs626-460-2011

Text Corpora for NLP processing

- A collection of text called **corpus**, is used for collecting various language data
  - **Unlabeled**: cleaned text without any annotation
  - **Labeled**: Text labeled for classification, with summaries, translations, question answering pairs
- **With annotation:** more information, but manual labor intensive

Popular Text Corpora

- **Text Classification:**
  - **IMDb Movie Reviews**
    - A dataset containing movie reviews along with sentiment labels (positive/negative), commonly used for sentiment analysis and binary classification tasks.
  - **20 Newsgroups:**
    - A collection of approximately 20,000 newsgroup documents across 20 different categories, often used for text classification and topic modeling tasks.
  - **GLUE Benchmark Datasets:** Assorted NLP classification tasks for benchmarking / testing new NLP models
    - https://gluebenchmark.com/tasks/

Popular Text Corpora – Summarization

- **CNN/Daily Mail:**
  - A dataset consisting of news articles paired with human-generated summaries. It is widely used for abstractive text summarization tasks.
- **DUC (Document Understanding Conference) datasets:**
  - Datasets from the Document Understanding Conference containing documents and manually created summaries, used for extractive and abstractive summarization evaluation.

Multilingual Corpora – Translation

- **WMT (Workshop on Machine Translation) Datasets:**
  - WMT provides datasets for machine translation tasks. The datasets cover multiple language pairs and are commonly used for training and evaluating translation models.
- **IWSLT (International Workshop on Spoken Language Translation) Datasets:**
  - IWSLT offers datasets for spoken language translation tasks, which include parallel text and audio data in multiple languages.

Corpora for Parsing / POS / Chunks

- **Penn Treebank** – from Upenn, Sentences tagged with POS and Parse trees / graphs
- **Brown Corpus**
- **CoNLL Datasets for Named Entity Recognition**

Unlabeled Corpora for Language Modeling

1. **Common Crawl:** A web archive corpus that includes data crawled from a wide range of websites.
2. **Wikipedia Dump:** Wikipedia provides periodic dumps of its entire content, including articles in various languages.
3. **OpenSubtitles:** A large collection of subtitles from movies and TV shows, available in multiple languages.

Unlabeled Corpora for Language Modeling

4. BookCorpus: A dataset containing text excerpts from books. It is used for training language models and extracting information from a diverse range of literary content.

5. Gutenberg Corpus: Project Gutenberg offers a collection of freely available literary works, including novels, essays, and poetry. It is a valuable resource for unsupervised learning and language modeling.

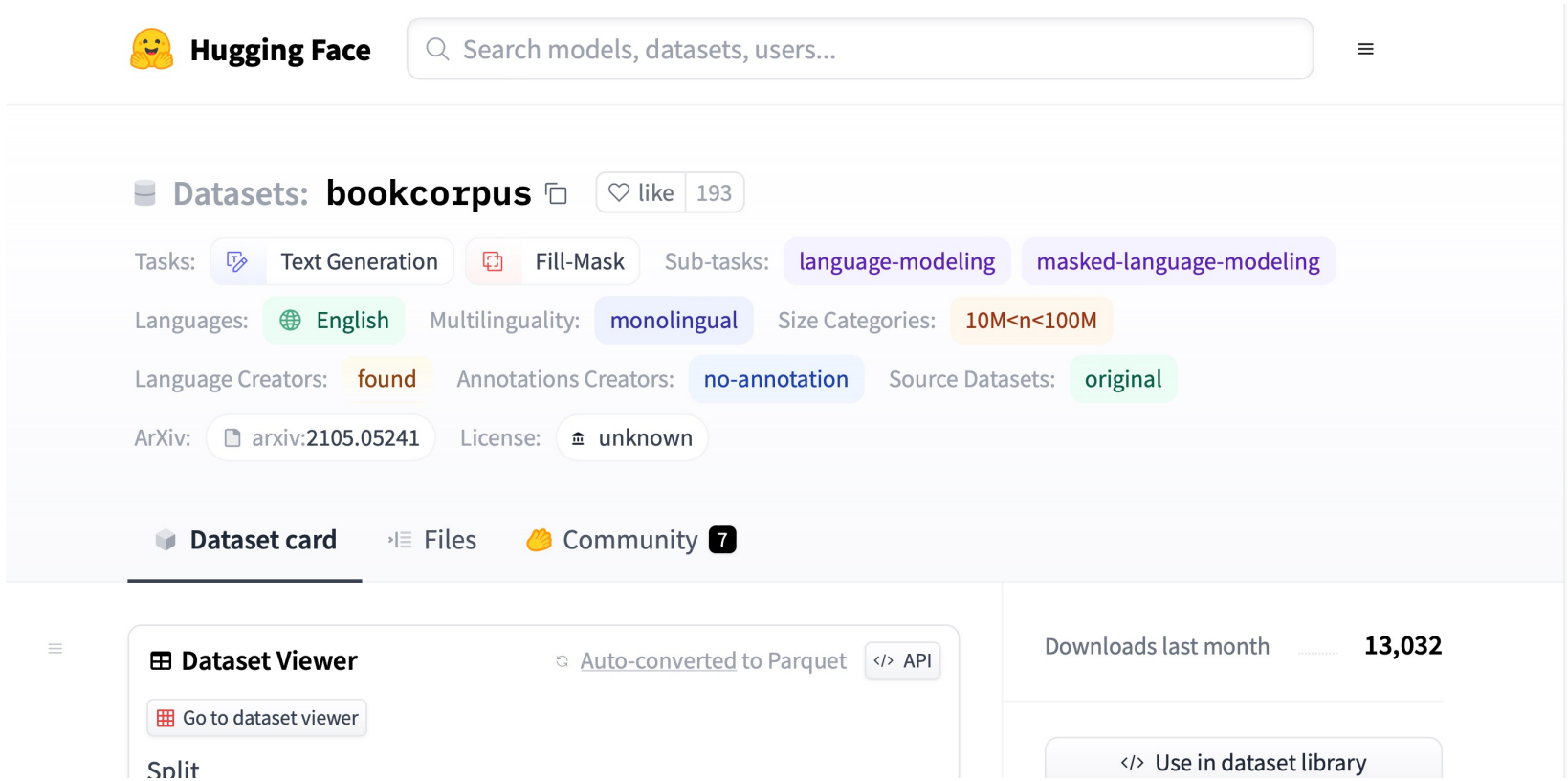
6. Reuters Corpus: The Reuters Corpus is a collection of news articles from the Reuters news agency.

Unlabeled Corpora for Language Modeling

- **English Gigaword:** A large newswire corpus containing news articles from a variety of sources. While some versions include part-of-speech tags, the raw text is often used for unsupervised learning tasks.
- **Billion Word Corpus:** A large-scale corpus consisting of a billion-word dataset from web pages. It is commonly used for training language models due to its extensive size.
- **One Billion Word Benchmark:** Similar to the Billion Word Corpus, this benchmark provides a large amount of unlabeled text for language modeling tasks.
- **Reddit Data:** Reddit provides data dumps of discussions and comments from its platform. The raw text from Reddit discussions can be used for various unsupervised learning tasks.
- **Red pajama Dataset: 1 Trillion tokens / words, open sourced for LLM development**

Major Data Sources

- Kaggle
- Huggingface Dataset



Next class

Lab: Introduction to NLTK and SpaCy libraries, Regular Expressions and Pattern Matching in Python, Loading and cleaning text data

Please fill out the pre-course survey here: https://forms.gle/FzJnF8esckd7bPtn9Links to an external site.

Next week: Ambiguity, Multilingualism, Fundamentals layers of NLP, Overview of text corpora and datasets, Regular Expressions,