V.E.S.
Since 1962

VES INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING

Introduction to NLP

Mrs. Vidya S Zope

outline

- What is NLP?
- Why NLP?
- Where does NLP fits in Computer taxonomy?
- History of NLP
- Generic NLP system
- Levels of NLP
- Knowledge in language processing

Natural?

- Natural Language?

  Refers to the language spoken by people, e.g. English, Hindi, Marathi as opposed to artificial/programming languages, like C, C++, Java, etc.

- Natural Language Processing

  Applications that deal with these natural languages in a way or another.

What is Natural Language Processing (NLP)

- The process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.

- **The sub-domain of artificial intelligence concerned with the task of developing programs possessing some capability of 'understanding' a natural language in order to achieve some specific goal.**

- The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages.

- The field of NLP is secondarily concerned with helping us come to a better understanding of human language.

Goals of NLP

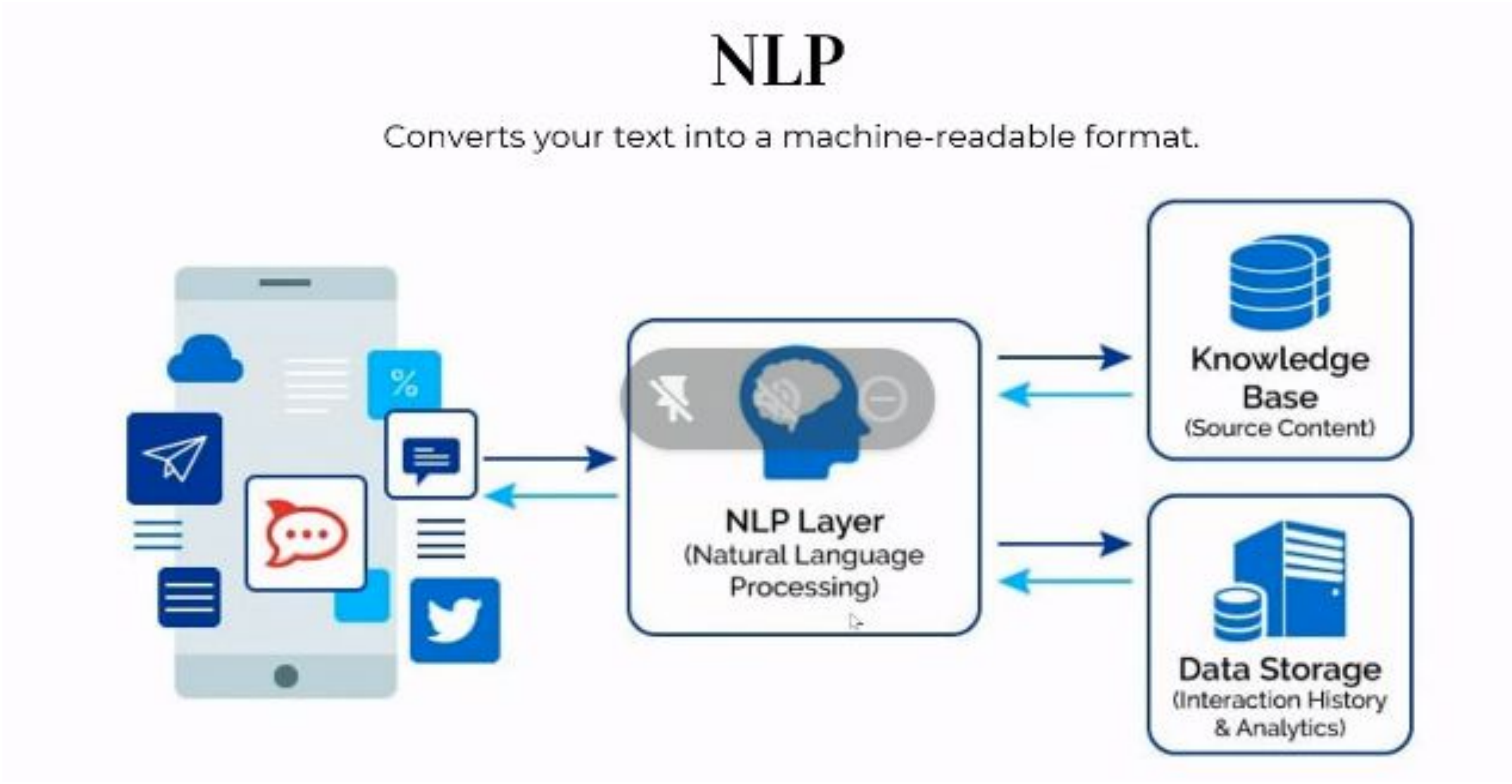
Scientific goal: Understand the way language operates.

Engineering goals: build systems that analyze and generate language, and reduce the man machine gap.

Why Natural Language Processing?

language data (Structured / Unstructured) in electronic form in different languages.

- e-commerce companies need to know sentiment of online users, sifting through 1 lakh e-opinions per week: needs NLP

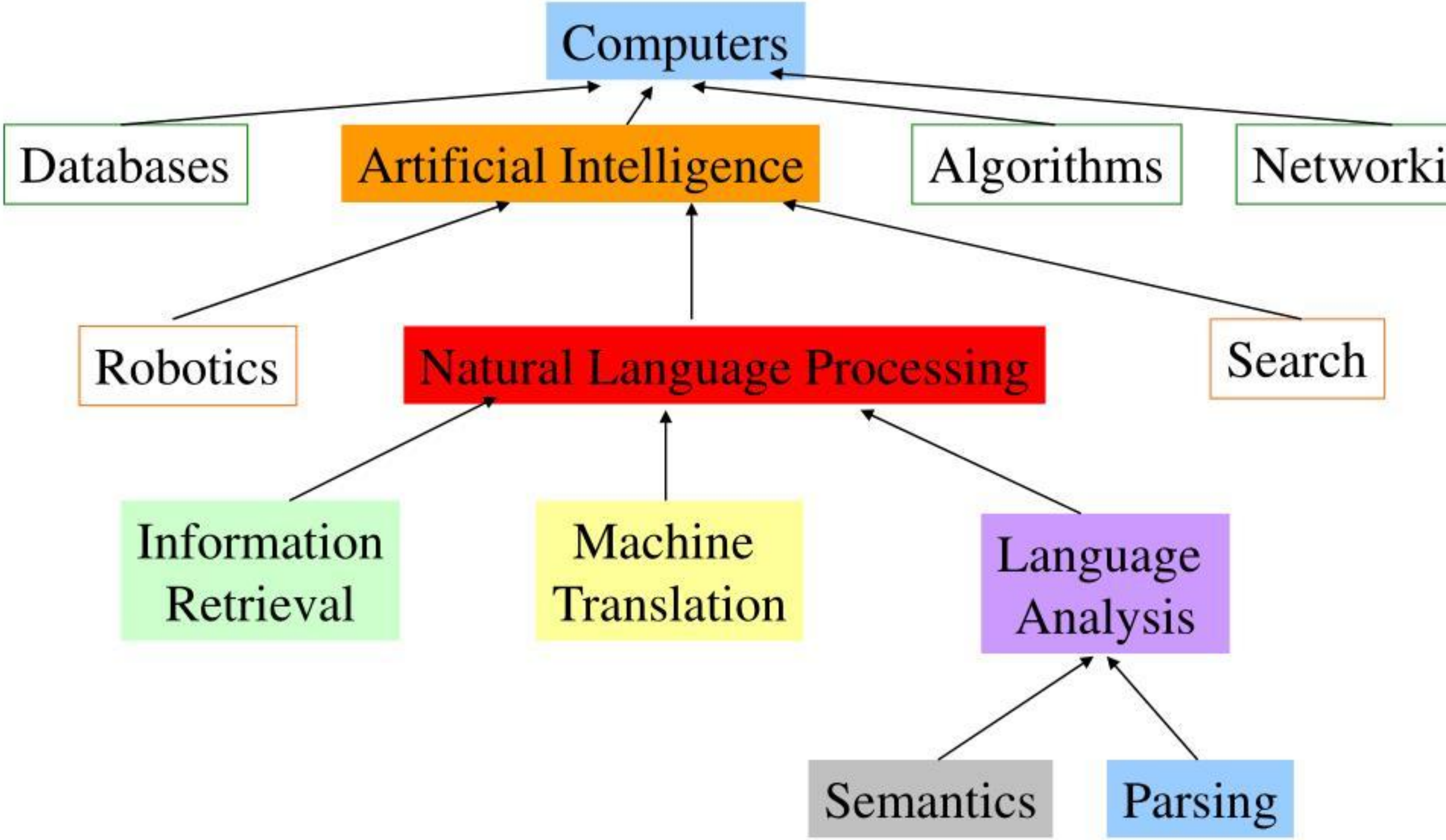


NLP

Converts your text into a machine-readable format.

NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)

Where does it fit in the CS taxonomy?



Forms of Natural Language

- The input/output of a NLP system can be:
  1. **Written text**
  2. **Speech**

- To process written text, we need: lexical, syntactic, semantic knowledge about the language, discourse information, real world knowledge.

- To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

Why Natural Language Processing?

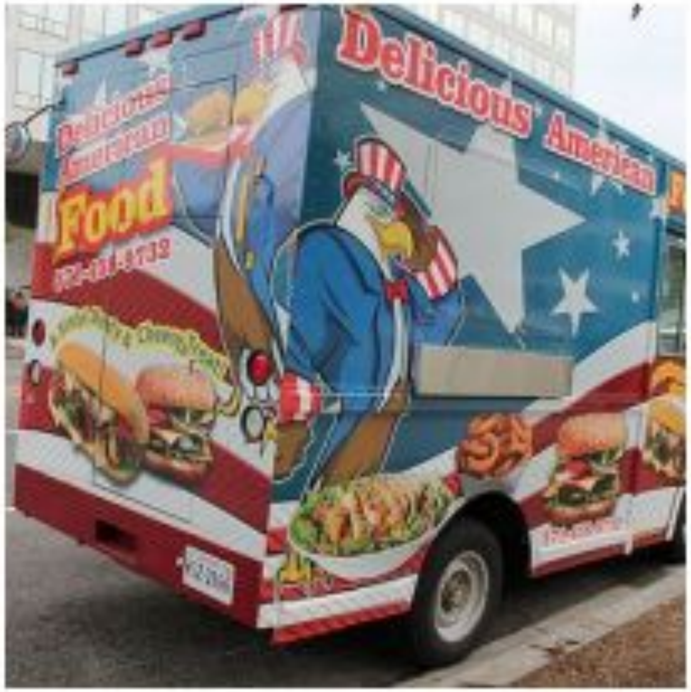
- kJfmmfj mmmvvv nnnffn333
- Uj iheale eleee mnster vensi credur
- Baboi oi cestnitze
- Coovoel2^ ekk; ldsllk lkdf vnnjfj?
- Fgmflmllk mlfm kfre xnnn!

Computers Lack Knowledge!

- Computers “see” text in English the same you have seen the previous text!
- People have no trouble understanding language
    - Common sense knowledge
    - Reasoning capacity
    - Experience
- Computers don’t have

  common sense knowledge

  reasoning capacity
- **Unless we teach them**



Q: What endangered animal is featured on the truck?

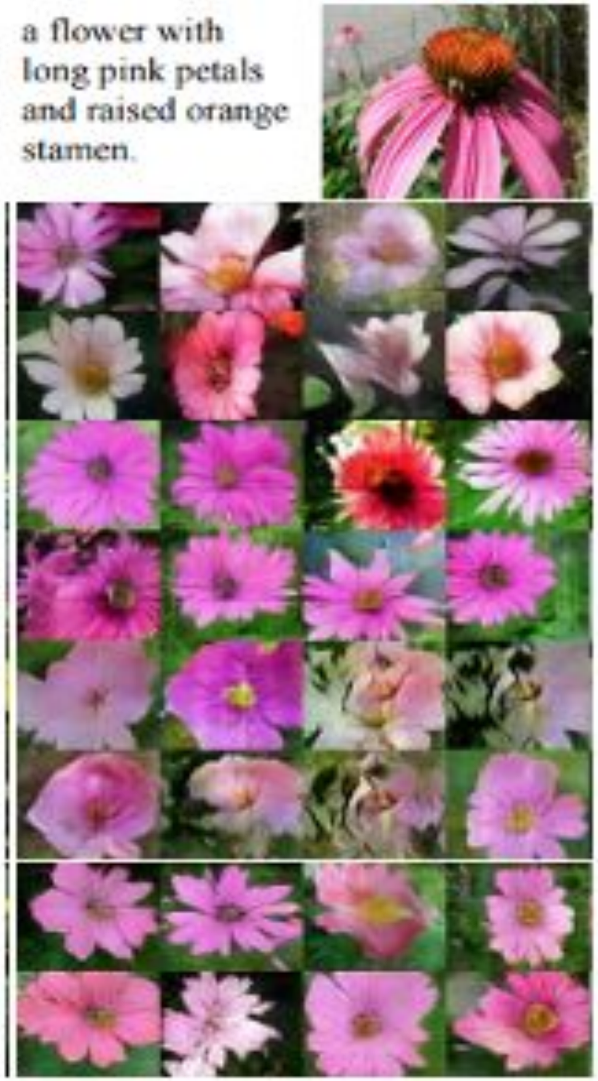
A: **A bald eagle.**

A: A sparrow.

A: A humming bird.

A: A raven.

Source: http://vision.stanford.edu/pdf/zhu2016cvpr.pdf



Results on Oxford-102

https://towardsdatascience.com/text-to-image-a3b201b003ae

Example:

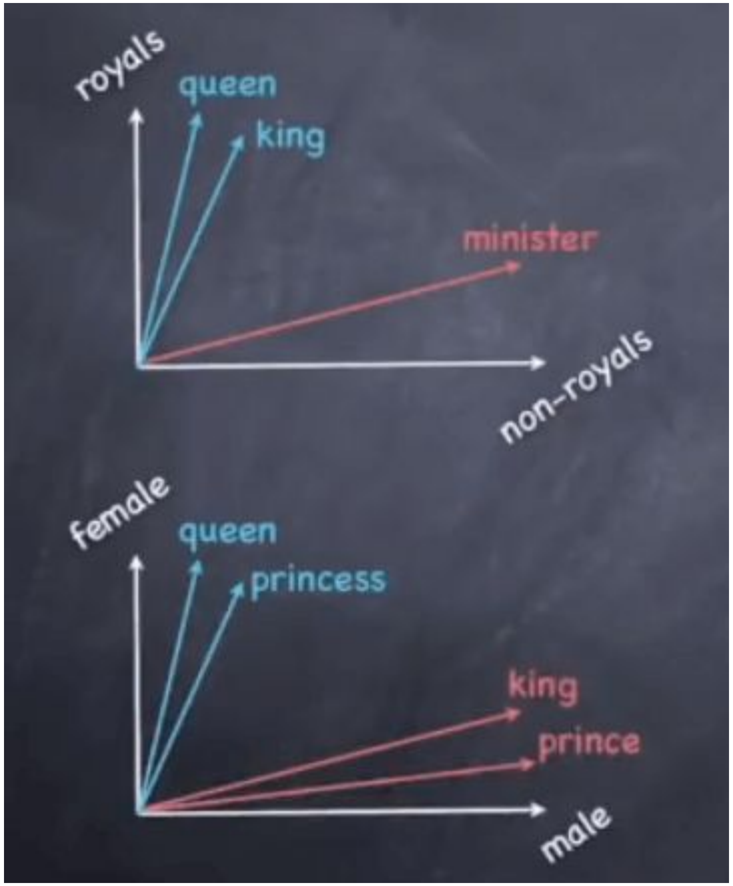
1. Man is to woman as king is to ___________?

Meaning (king) – meaning (man) + meaning (woman)=?

The answer is-

2. Is King to kings as the queen is to________?

The answer is---



NLP can be used for:

- Speech recognition (text-to-speech and speech-to-text).
- Segmenting previously captured speech into individual words, sentences, and phrases.
- Recognizing basic forms of words and acquisition of grammatical information.
- Recognizing functions of individual words in a sentence (subject, verb, object, article, etc.)
- Extracting the meaning of sentences and parts of sentences or phrases, such as adjective phrases (e.g., “too long”), prepositional phrases (e.g., “to the river”), or nominal phrases (e.g., “the long party”).
- Recognizing sentence contexts, sentence relationships, and entities.
- Linguistic text analysis, sentiment analysis, translations (including those for voice assistants), chatbots, and underlying question-and-answer systems

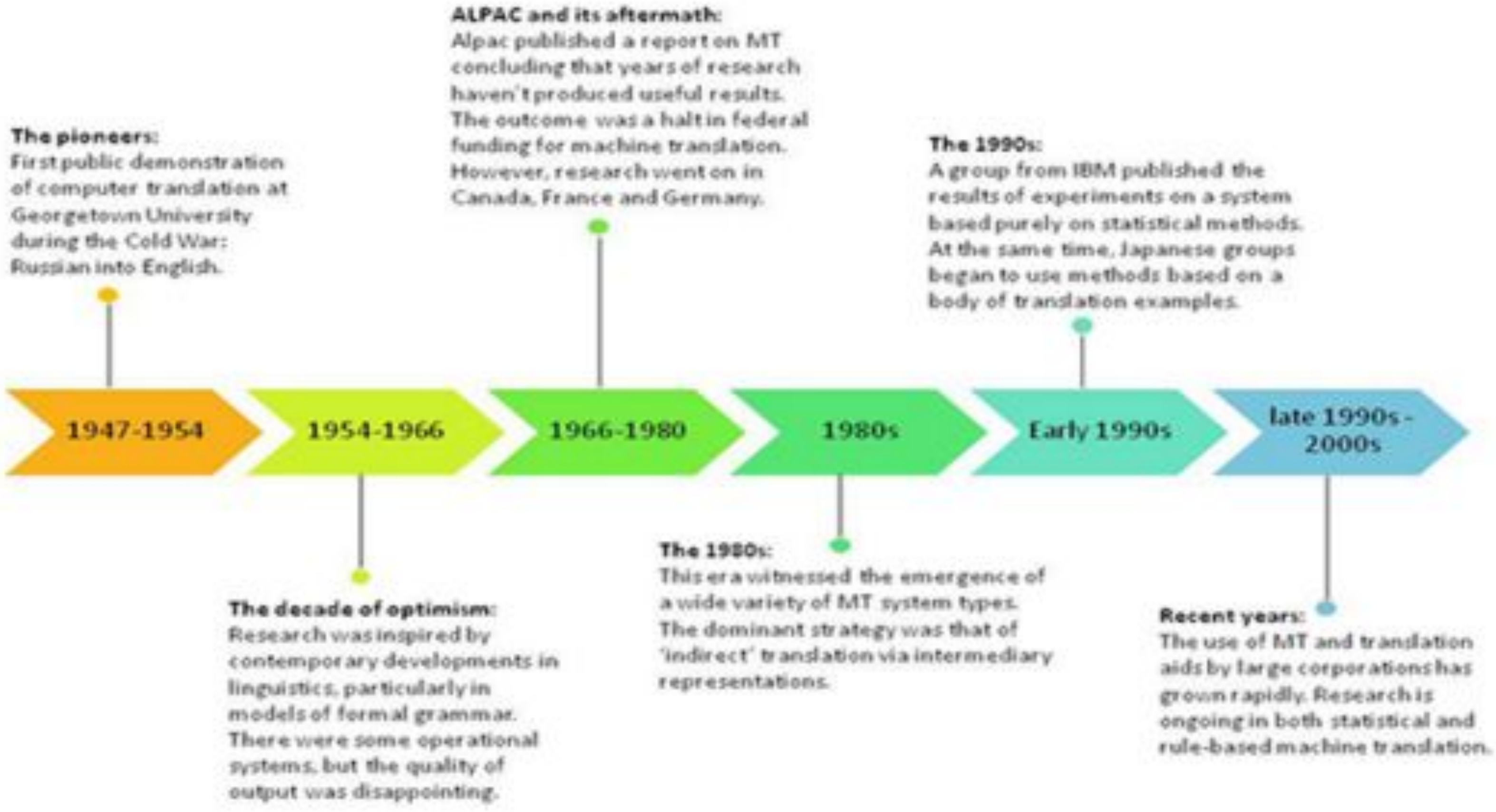
Natural language vs. Computer Language

Parameter	Natural Language	Computer Languages
Ambiguous	They are ambiguous in nature.	They are designed to unambiguous.
Redundancy	Natural languages employ lots of redundancy.	Formal languages are less redundant.
Literalness	Natural languages are made of idiom & metaphor	Formal languages mean exactly what they want to say

History of NLP

- **1950-** NLP started when Alan Turing published an article called "Machine and Intelligence."
- **1950-** Attempts to automate translation between Russian and English
- **1960-** The work of Chomsky and others on formal language theory and generative syntax
- **1990-** Probabilistic and data-driven models had become quite standard
- **2000-** A Large amount of spoken and textual data become available

MT timeline



1947-1954

The pioneers:
First public demonstration of computer translation at Georgetown University during the Cold War: Russian into English.

1954-1966

The decade of optimism:
Research was inspired by contemporary developments in linguistics, particularly in models of formal grammar. There were some operational systems, but the quality of output was disappointing.

1966-1980

ALPAC and its aftermath:
Alpac published a report on MT concluding that years of research haven't produced useful results. The outcome was a halt in federal funding for machine translation. However, research went on in Canada, France and Germany.

1980s

The 1980s:
This era witnessed the emergence of a wide variety of MT system types. The dominant strategy was that of 'indirect' translation via intermediary representations.

Early 1990s

The 1990s:
A group from IBM published the results of experiments on a system based purely on statistical methods. At the same time, Japanese groups began to use methods based on a body of translation examples.

late 1990s - 2000s

Recent years:
The use of MT and translation aids by large corporations has grown rapidly. Research is ongoing in both statistical and rule-based machine translation.

What makes NLP Challenging?

Ambiguity

- The dog is in the pen.
- The ink is in the pen.
- Stones cried.
- I need some paper.
- I wrote a paper.
- I read the paper.

Components of NLP

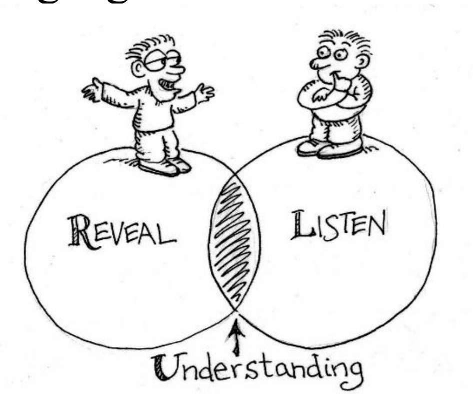
- **Natural Language Understanding**
  - Mapping the given input in the natural language into a useful representation.
  - Different level of analysis required:

    morphological analysis,

    syntactic analysis,

    semantic analysis,

    discourse analysis



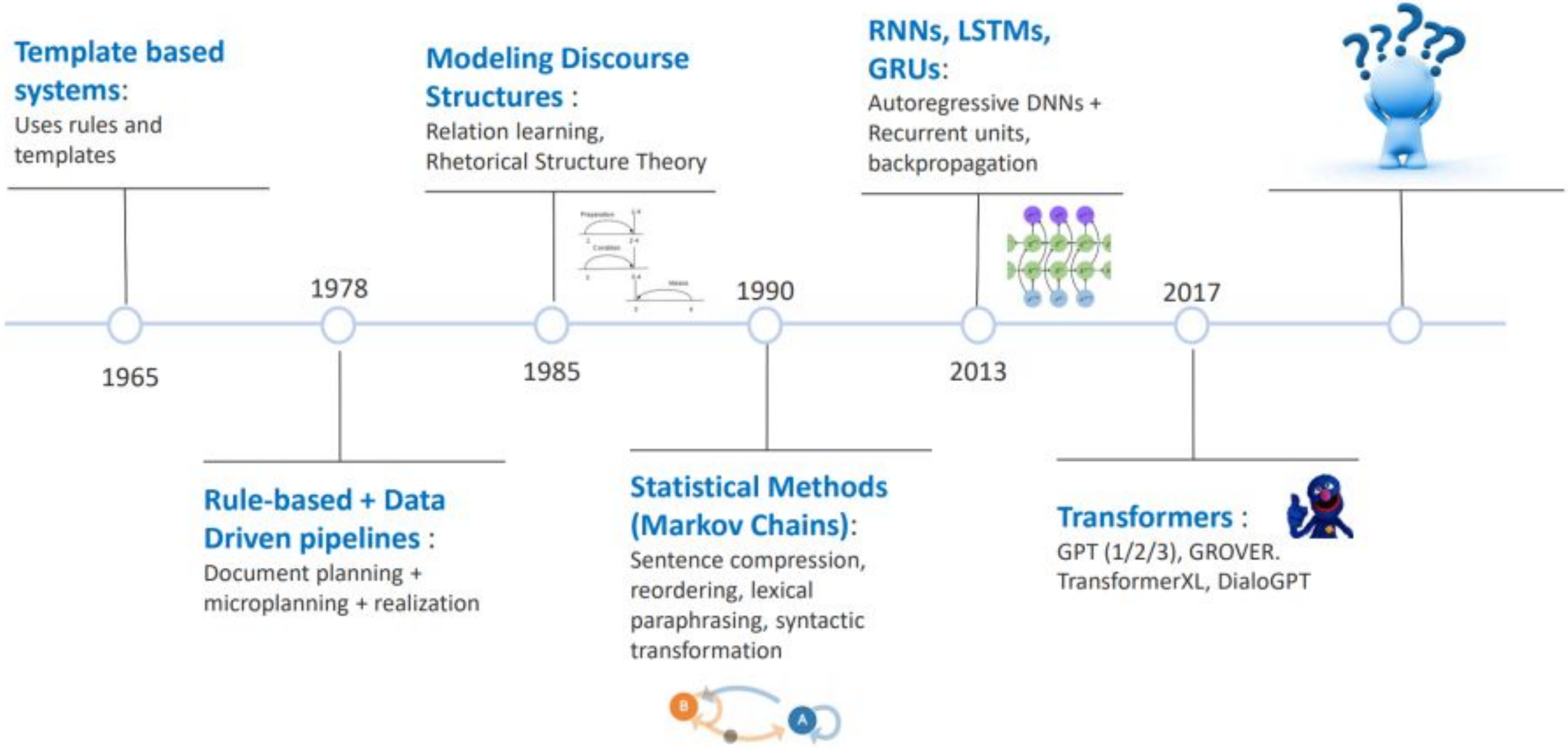
- **Natural Language Generation**
  - Producing output in the natural language from some internal representation.
  - Different level of synthesis required:

    deep planning (what to say),

    syntactic generation

- NL Understanding is much harder than NL Generation. But, still both of them are hard.

History of NLG



1965

Template based systems: Uses rules and templates

1978

Rule-based + Data Driven pipelines : Document planning + microplanning + realization

1985

Modeling Discourse Structures : Relation learning, Rhetorical Structure Theory

1990

Statistical Methods (Markov Chains): Sentence compression, reordering, lexical paraphrasing, syntactic transformation

2013

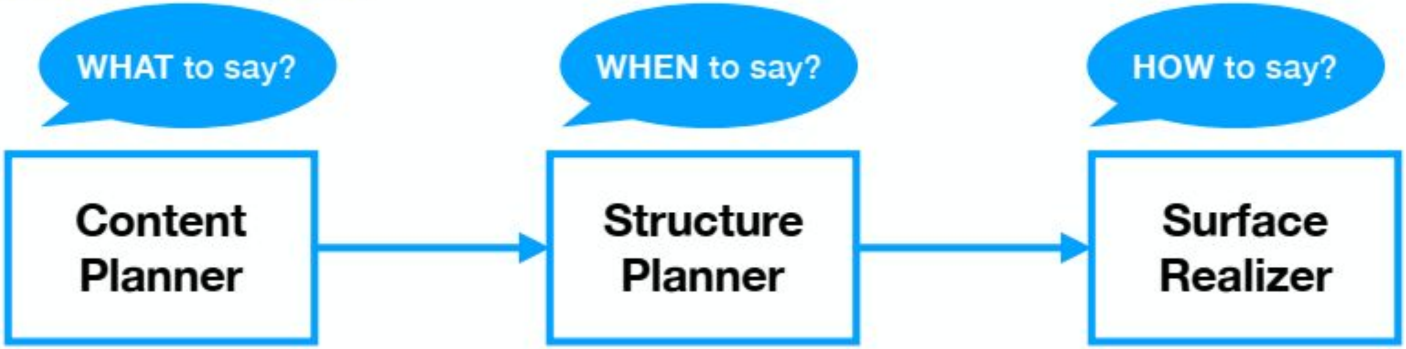
RNNs, LSTMs, GRUs: Autoregressive DNNs + Recurrent units, backpropagation

2017

Transformers : GPT (1/2/3), GROVER. TransformerXL, DialoGPT

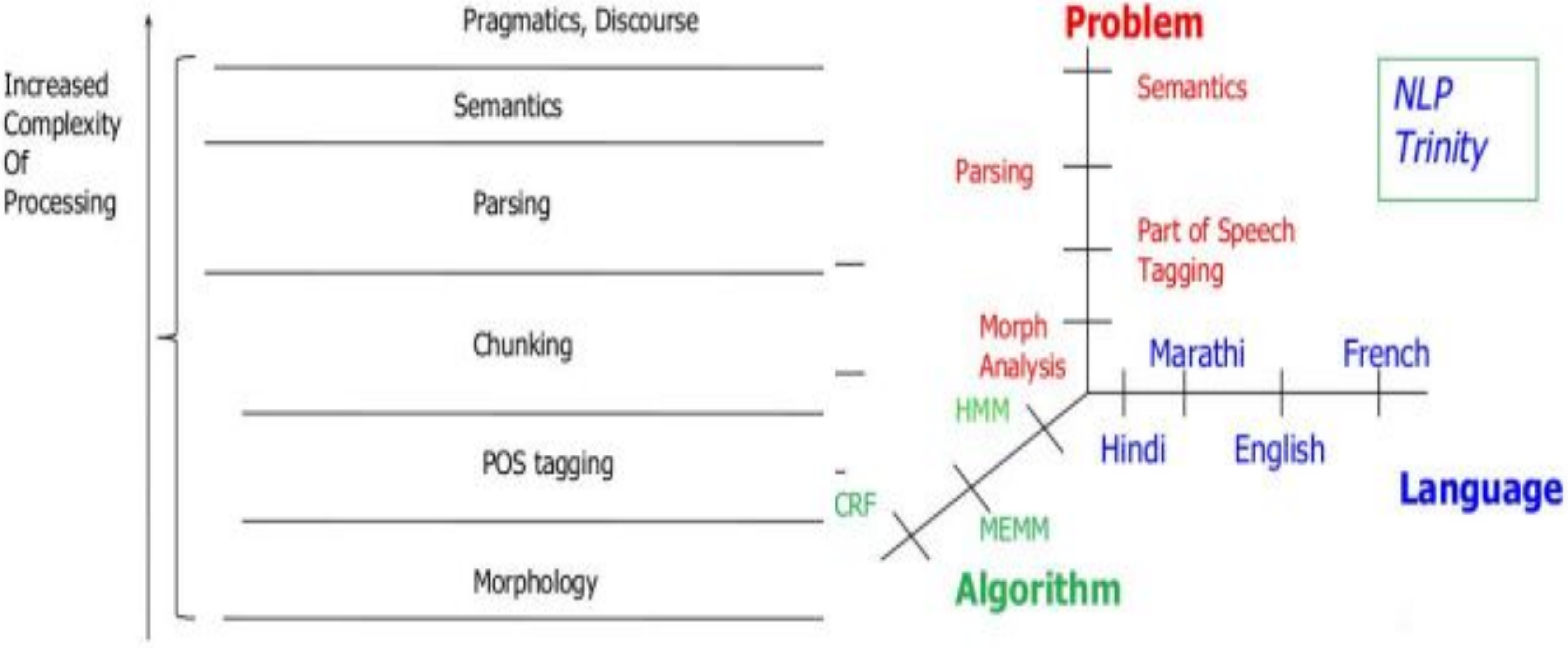
Credits: Tutorial on "The amazing world of Neural Text Generation"

Prior Methods in NLG



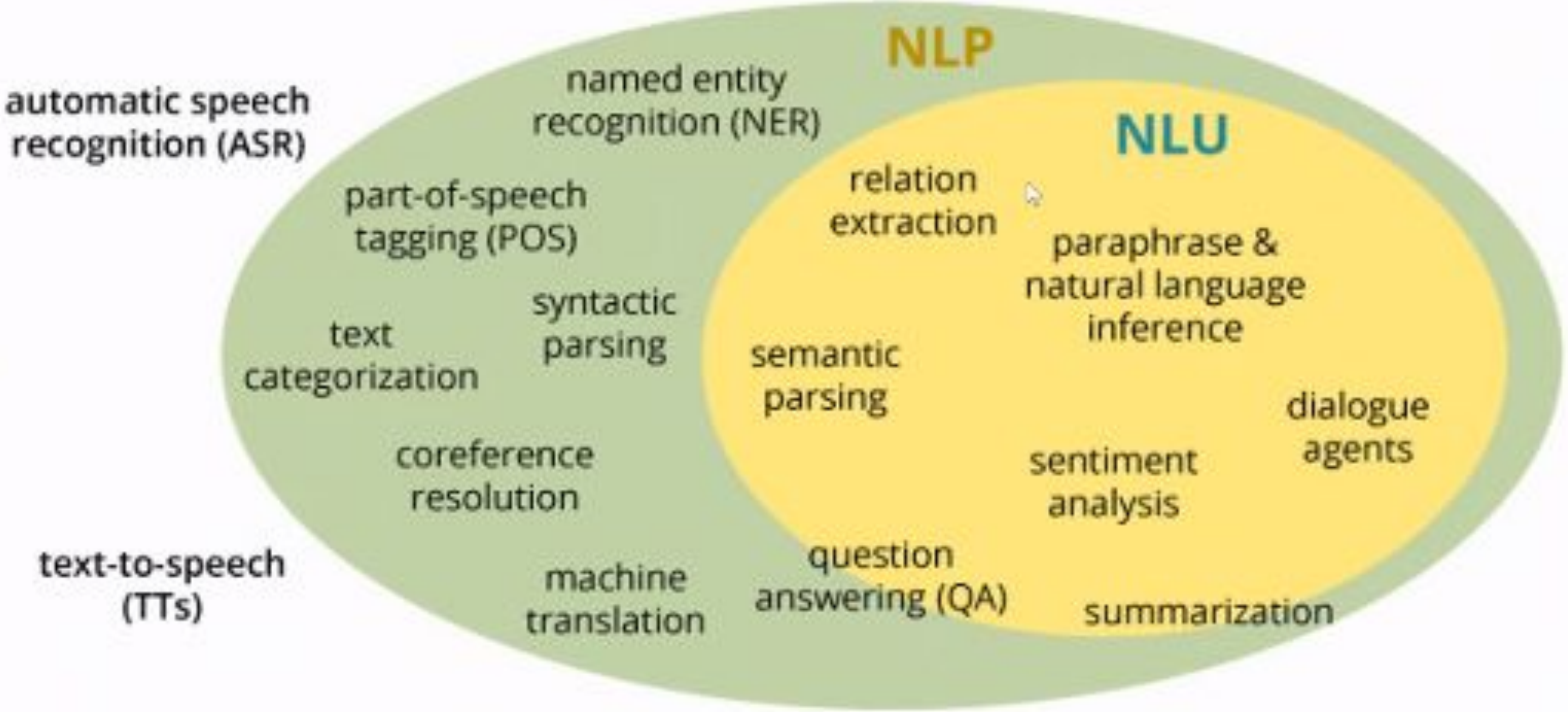
Classical	• Sentence Retrieval • MMR • Similarity Functions: • Tf-idf • Graph Algorithms	• Sentence Reordering • Sub-topic grouping • Heuristic Grammar • Templates • Rhetorical Structure Theory	• Dictionary lookup • Cut and replace • From ontologies and graph structures
Neural	• Attention • Bottom Up (masking) • Pointer Generator • Coverage Penalty • Diversity Loss	• Sequence of entities • Dynamic and Static Schema • Event Structures • Hierarchical Structure	• Attention • Latent variables (style)

NLP: Multi-layered, multi-dimensional

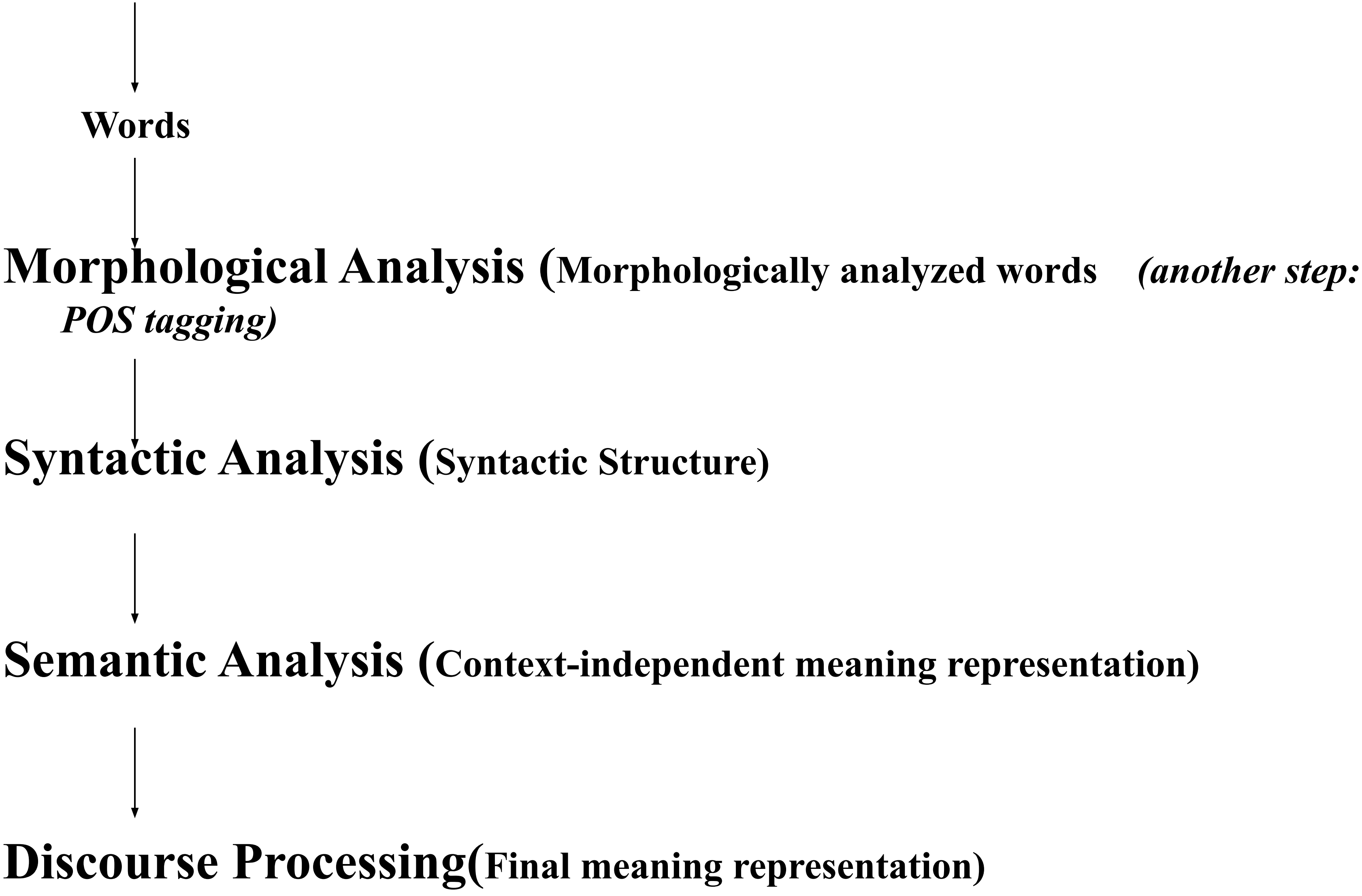


Source: Fuss Talk : Dr Pushpak Bhattacharyya

NLP v/s NLU


NLP
NLU
automatic speech recognition (ASR)
named entity recognition (NER)
part-of-speech tagging (POS)
relation extraction
paraphrase & natural language inference
syntactic parsing
text categorization
semantic parsing
dialogue agents
coreference resolution
sentiment analysis
text-to-speech (TTs)
question answering (QA)
machine translation
summarization

Natural Language Understanding



Words

Morphological Analysis (Morphologically analyzed words ***(another step: POS tagging)***

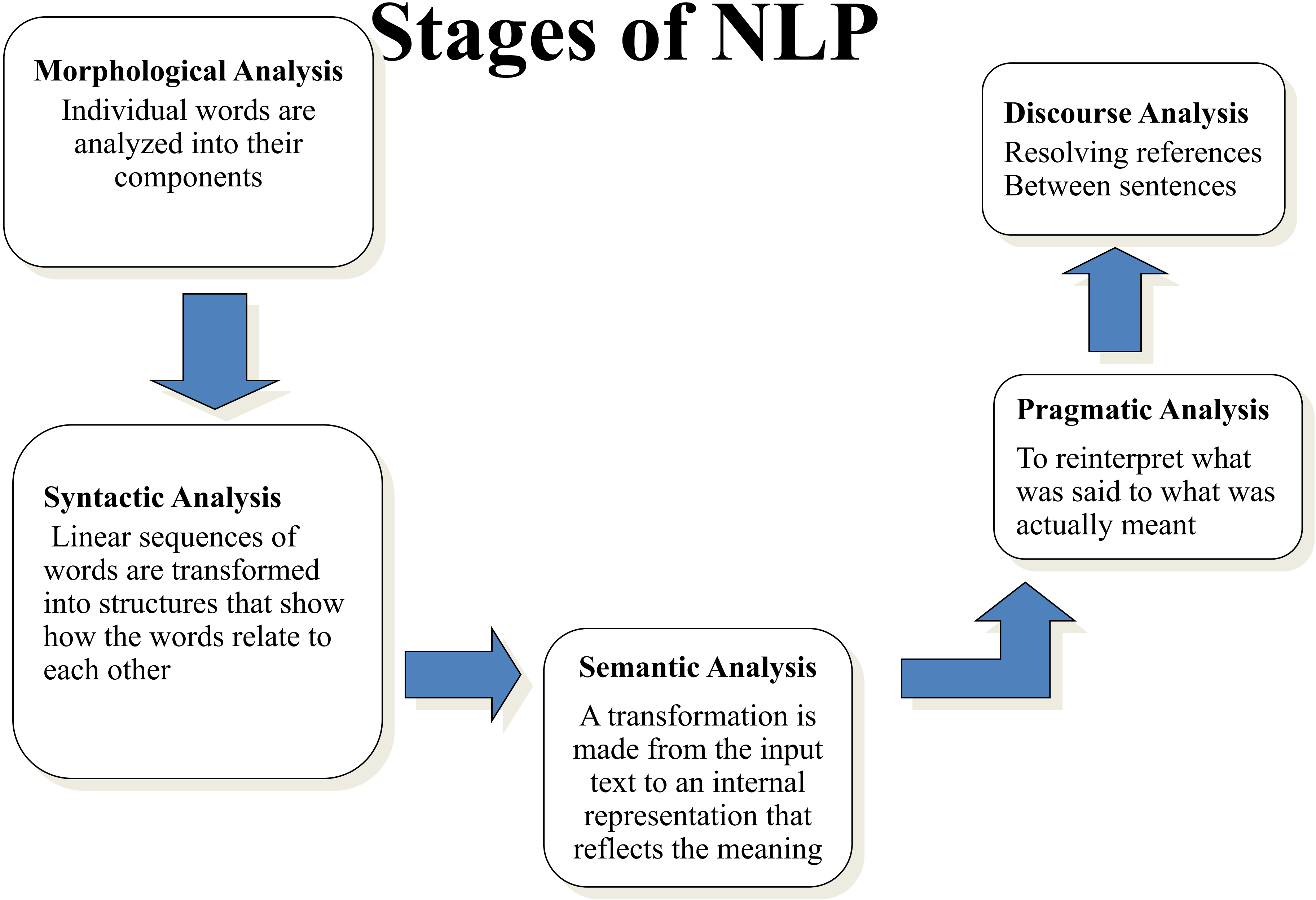
Syntactic Analysis (Syntactic Structure)

Semantic Analysis (Context-independent meaning representation)

Discourse Processing(Final meaning representation)

Stages of NLP

Stages of NLP



Morphological Analysis
Individual words are analyzed into their components

Syntactic Analysis
Linear sequences of words are transformed into structures that show how the words relate to each other

Semantic Analysis
A transformation is made from the input text to an internal representation that reflects the meaning

Pragmatic Analysis
To reinterpret what was said to what was actually meant

Discourse Analysis
Resolving references Between sentences

The stages Involved in NLP

- **Morphology:** Concerns the way words are built up from smaller meaning bearing units.
- **Syntax:** concerns how words are put together to form correct sentences and what structural role each word has.(Ex: “the dog ate my homework”)
- **Semantics:** concerns what words mean and how these meanings combine in sentences to form sentence meanings.(plant: industrial plant/ living organism)
- **Pragmatics:** concerns how sentences are used in different situations and how it affects the interpretation of the sentence.
- **Discourse:** concerns how the immediately preceding sentences affect the interpretation of the next sentence.

Lexical Disambiguation

First step: part of Speech Disambiguation

- Dog as a noun (animal)
- Dog as a verb (to pursue)

Needs word relationships in a context

The chair emphasized the need for adult education Very common in day to day communications.

Challenges in Syntactic Processing: Structural Ambiguity

- 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 (who has the telescope?)
- I saw the mountain with a telescope (world knowledge: mountain cannot be an instrument of seeing)
- I saw the boy with the pony-tail (world knowledge: pony-tail cannot be an instrument of seeing)

Semantic Analysis

- Challenge: ambiguity in semantic role labeling

(Eng) 1. Visiting aunts can be a nuisance

2. Colorless green ideas sleep furiously

Discourse

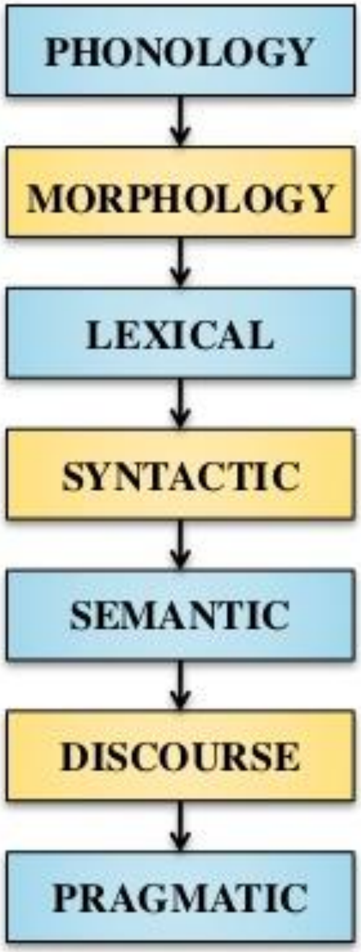
Processing of sequence of sentences

- Mother to John:

John go to school. It is open today. Should you bunk? Father will be very angry.

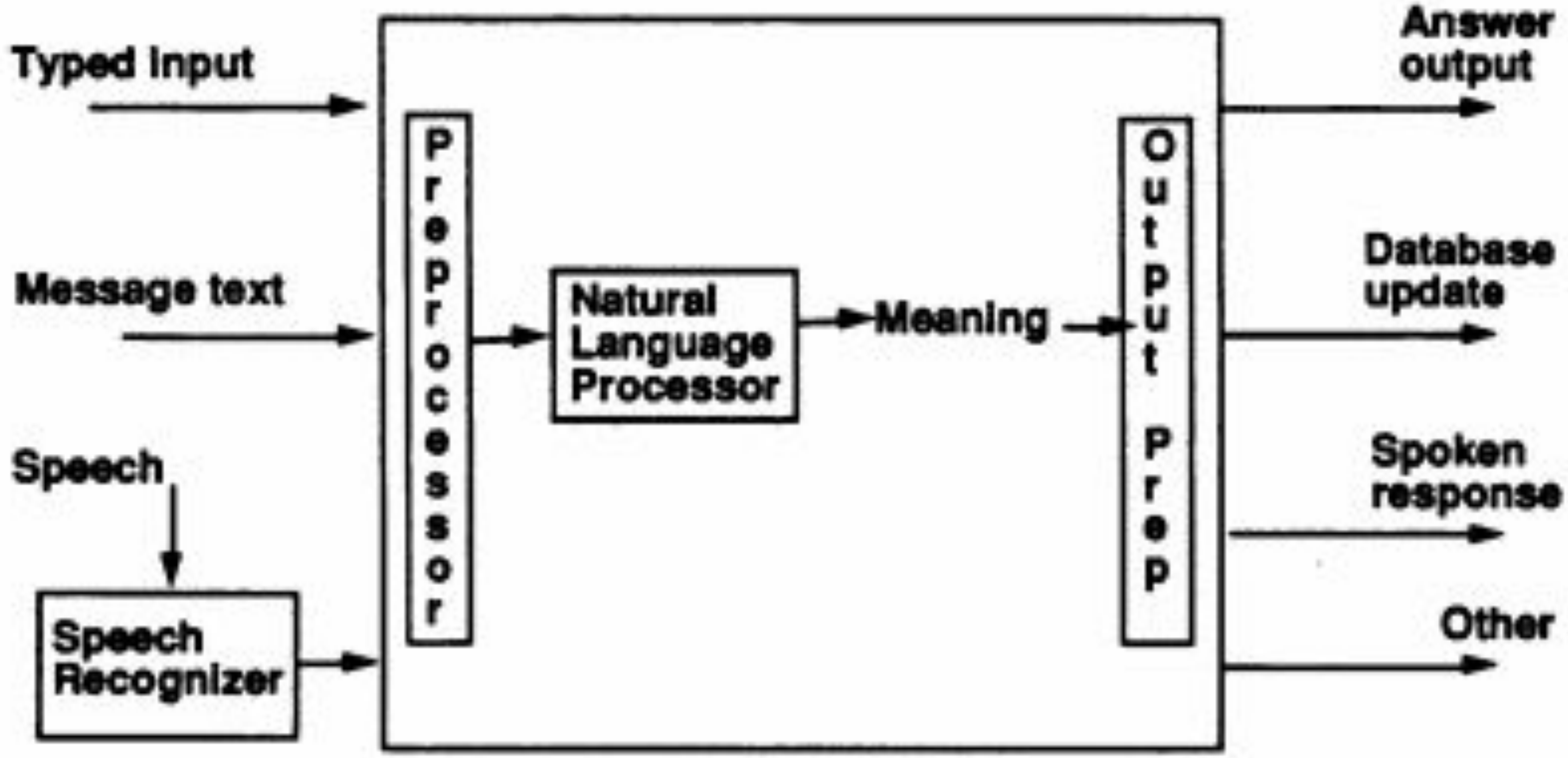
- Ambiguity of
  - open
  - bunk
- what?
- Why will the father be angry?

LEVELS



PHONOLOGY	-Interprets sounds within and across words
MORPHOLOGY	-Breaks down words into morphemes
LEXICAL	-Assign meanings to individual words
SYNTACTIC	-Check if sentence is grammatically correct
SEMANTIC	-Performs disambiguation of words
DISCOURSE	-Makes connection between sentences
PRAGMATIC	-Uses contextual and situational meanings

Generic NLP architecture



Pipeline view

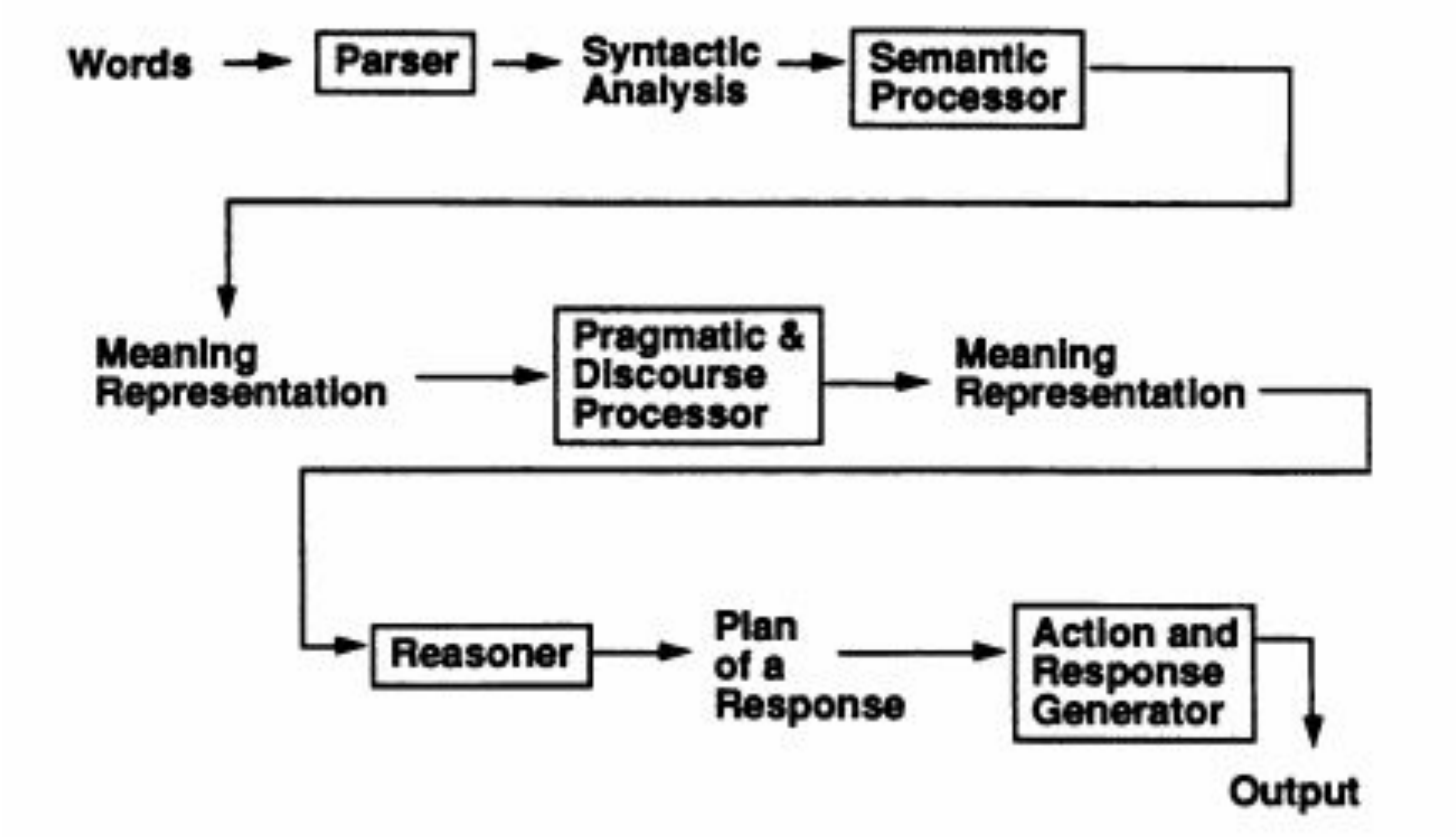
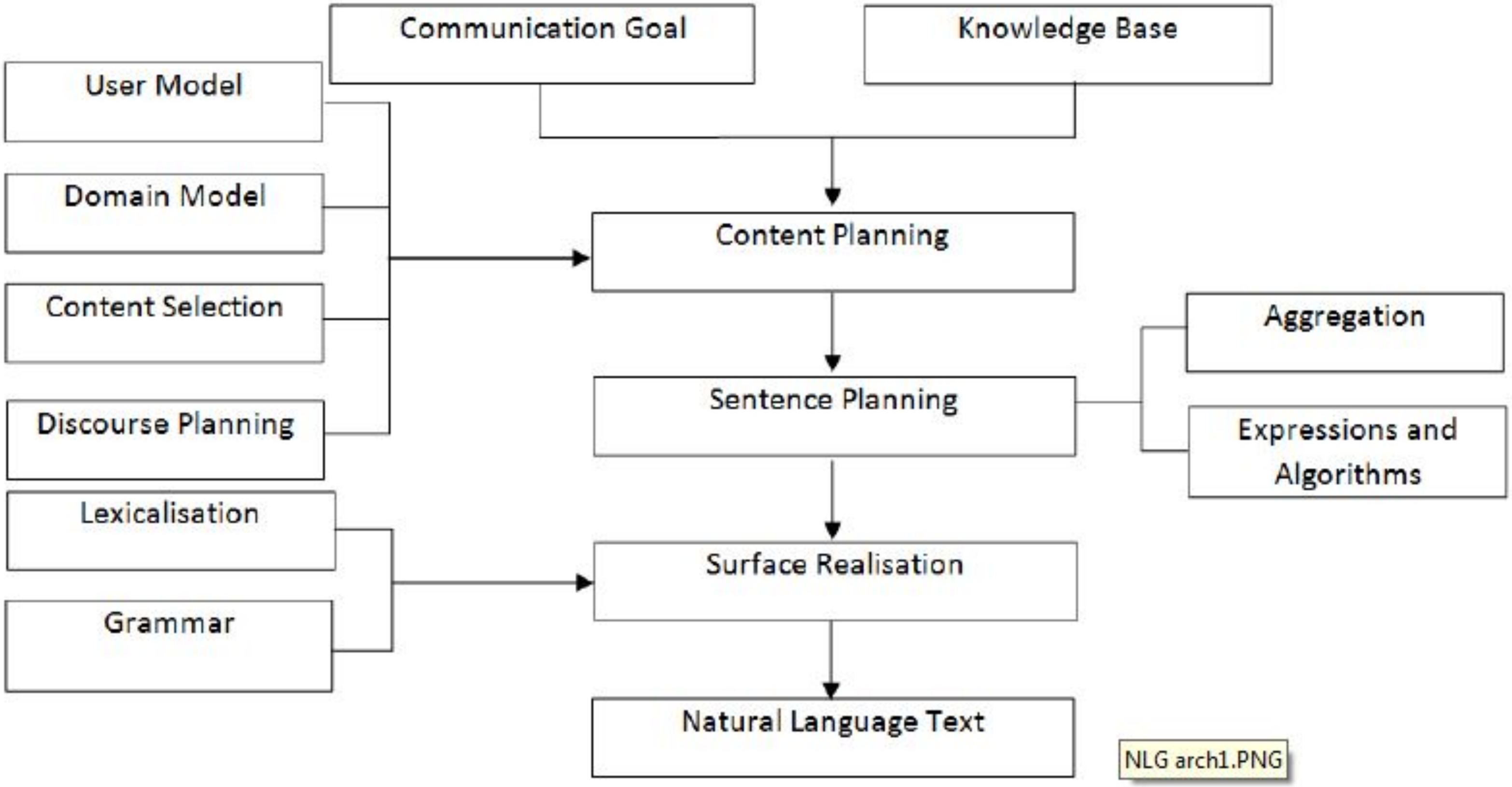


FIG. 3. A pipeline view of the components of a generic NL system.

Models of natural language understanding MADELEINE BATES

Phases of NLP architecture

Natural Language Processing: State of The Art, Current Trends and Challenges

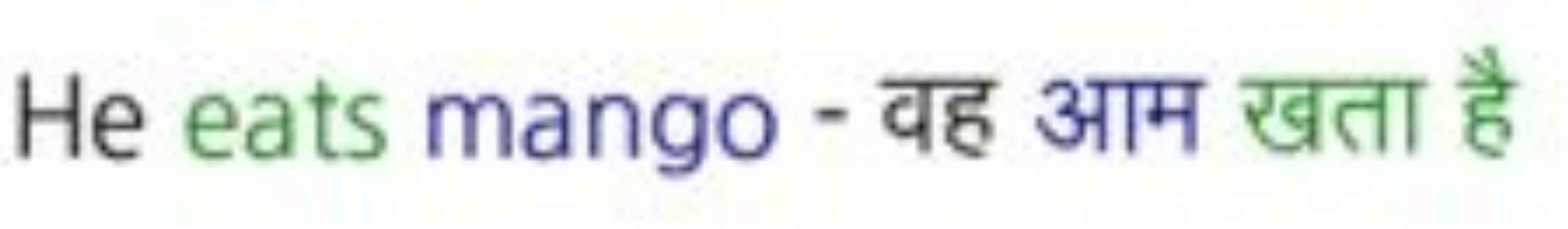


Applications of NLP

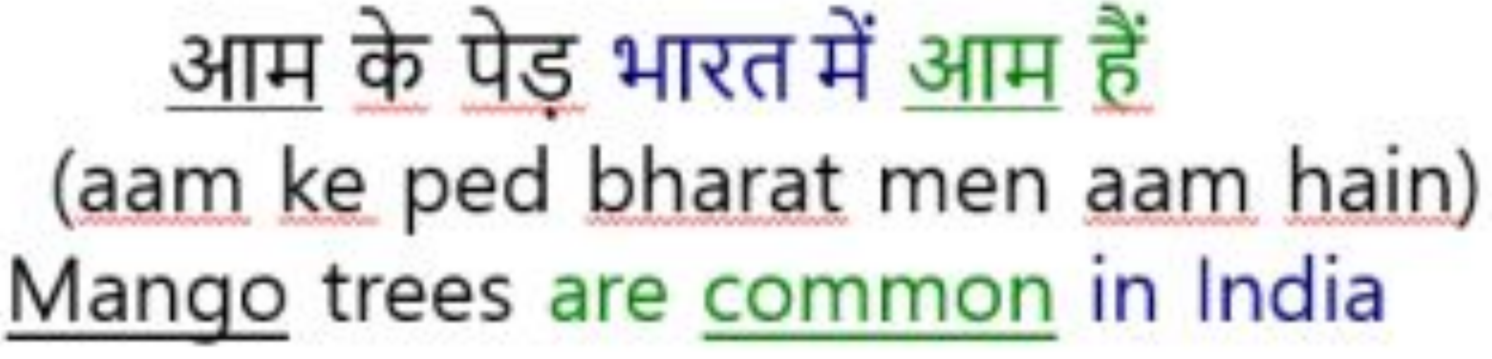
Applications

- Machine translation
- Question Answering System
- Information retrieval
- Image & Video Captioning
- Text categorization
- Text summarization
- Social Media Analysis..Sentiment Analysis

Machine Translation



He eats mango - वह आम खता है



आम के पेड़ भारत में आम हैं

(aam ke ped bharat men aam hain)

Mango trees are common in India

Named Entity Recognition

- Given a noun compound

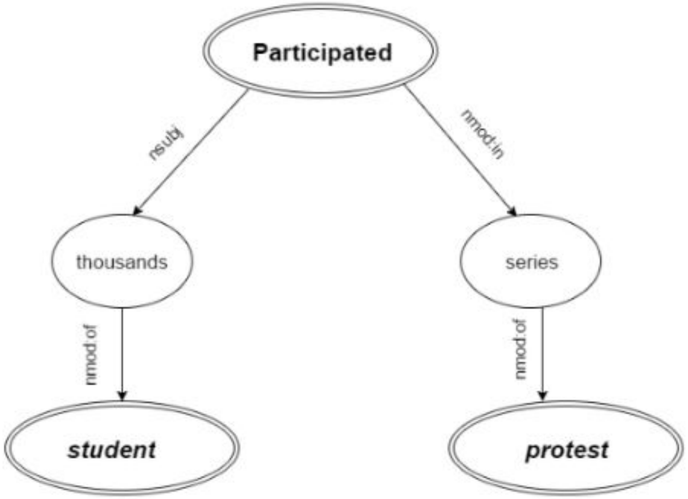
  NC: *"student protest"*

- Extracted sentences:

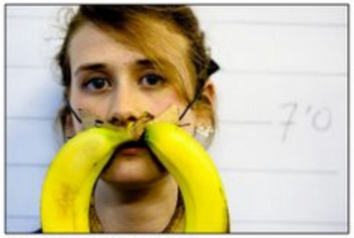
  *"Thousand of **students** have <u>participated</u> in a series of **protest** at JNU"*

- So, here..

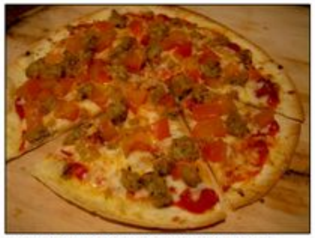
  "An **AGENT** *<u>participates in</u>* an **EVENT/ACT**"



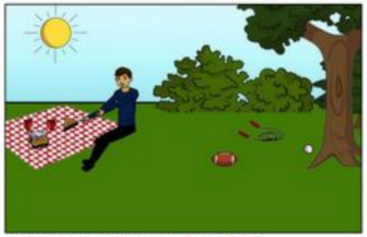
Given an image, can our machine answers the questions in natural language?



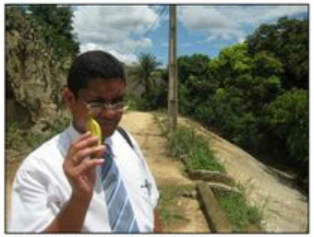
What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

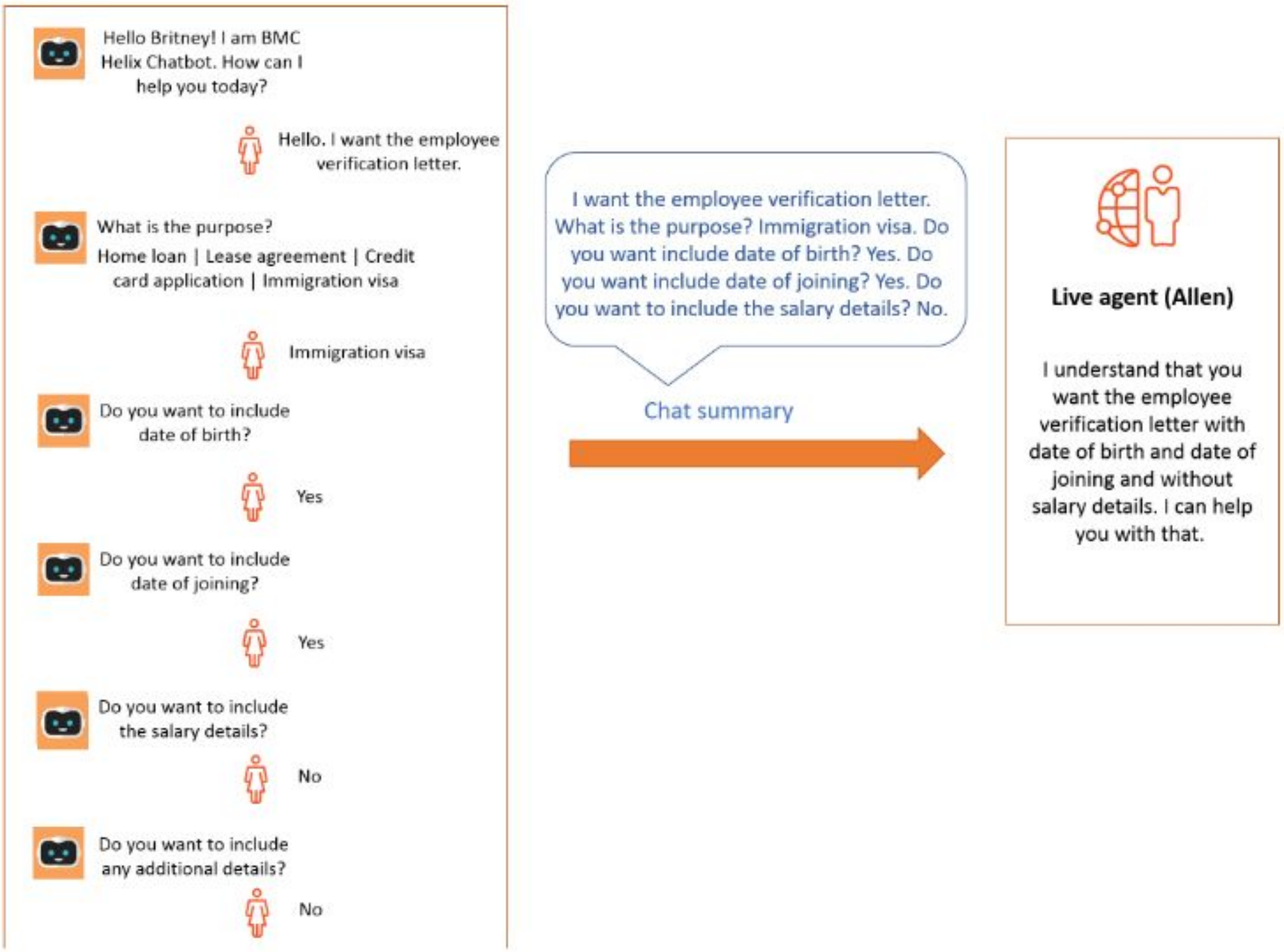


Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

NLG Pic Credit: https://docs.bmc.com/docs/helixplatform/support-for-text-summarization-in-yourapplication-888488610.html



Hello Britney! I am BMC Helix Chatbot. How can I help you today?

Hello. I want the employee verification letter.

What is the purpose?

Home loan | Lease agreement | Credit card application | Immigration visa

Immigration visa

Do you want to include date of birth?

Yes

Do you want to include date of joining?

Yes

Do you want to include the salary details?

No

Do you want to include any additional details?

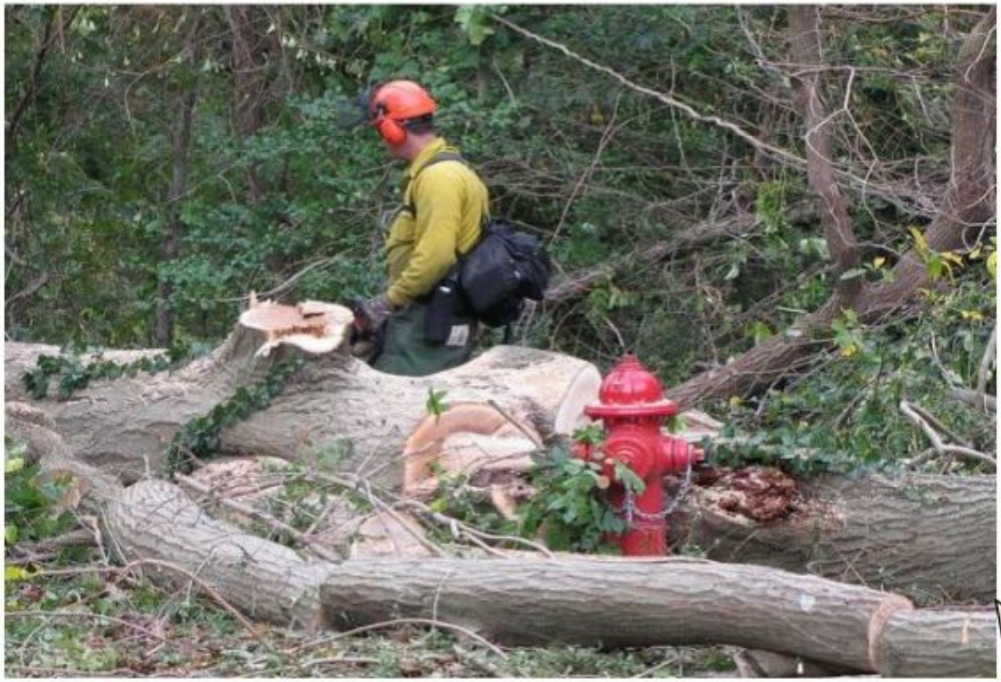
No

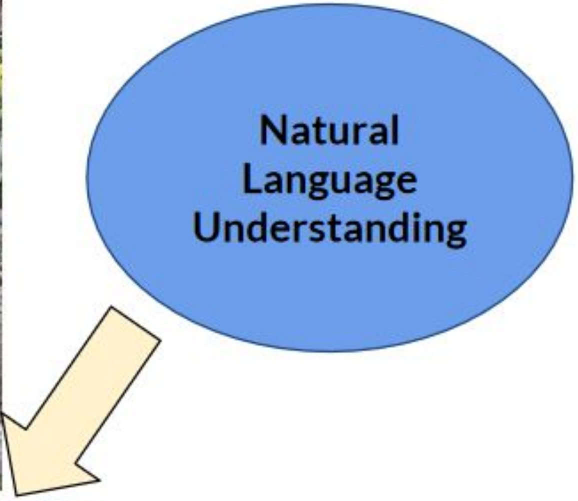
I want the employee verification letter. What is the purpose? Immigration visa. Do you want include date of birth? Yes. Do you want include date of joining? Yes. Do you want to include the salary details? No.

Chat summary

Live agent (Allen)

I understand that you want the employee verification letter with date of birth and date of joining and without salary details. I can help you with that.





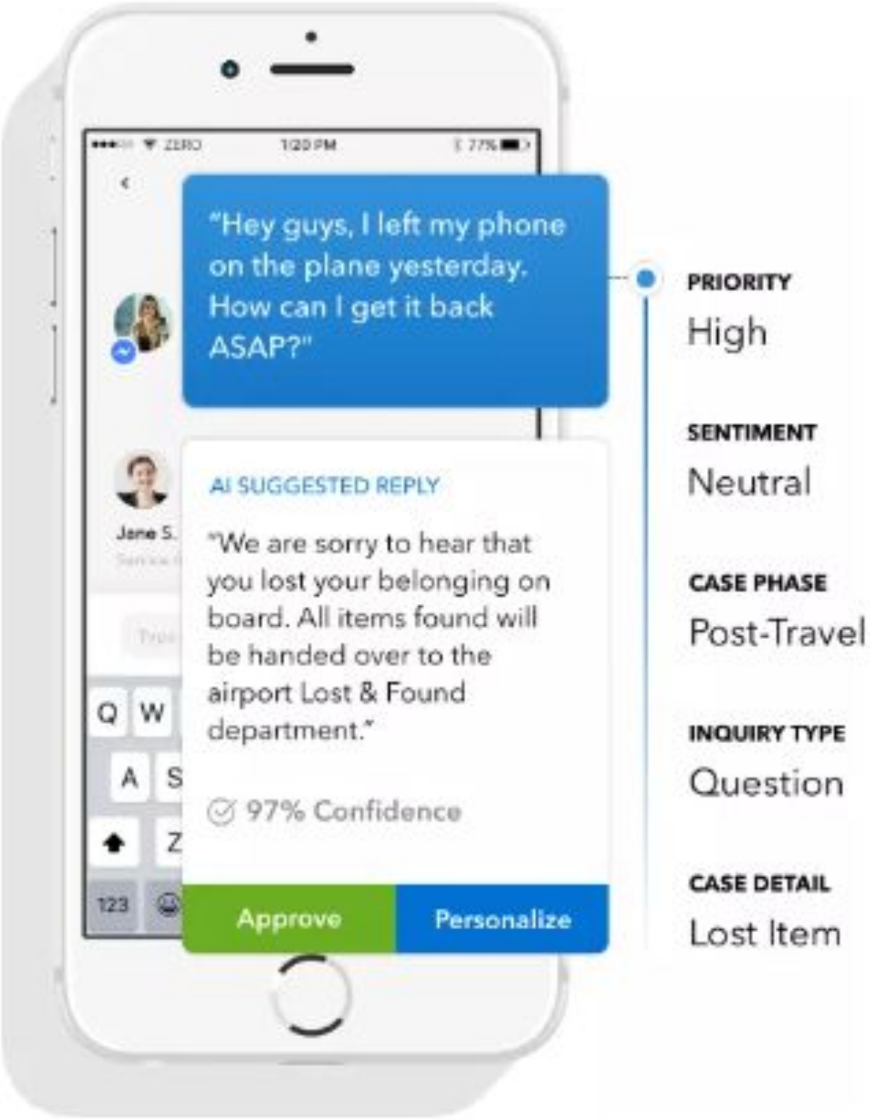
Natural Language Understanding

Question: What can the red object on the ground be used for ?

Answer: Firefighting

Wang et. al., 2016

Customer service automation



"Hey guys, I left my phone on the plane yesterday. How can I get it back ASAP?"

AI SUGGESTED REPLY

"We are sorry to hear that you lost your belonging on board. All items found will be handed over to the airport Lost & Found department."

97% Confidence

PRIORITY
High

SENTIMENT
Neutral

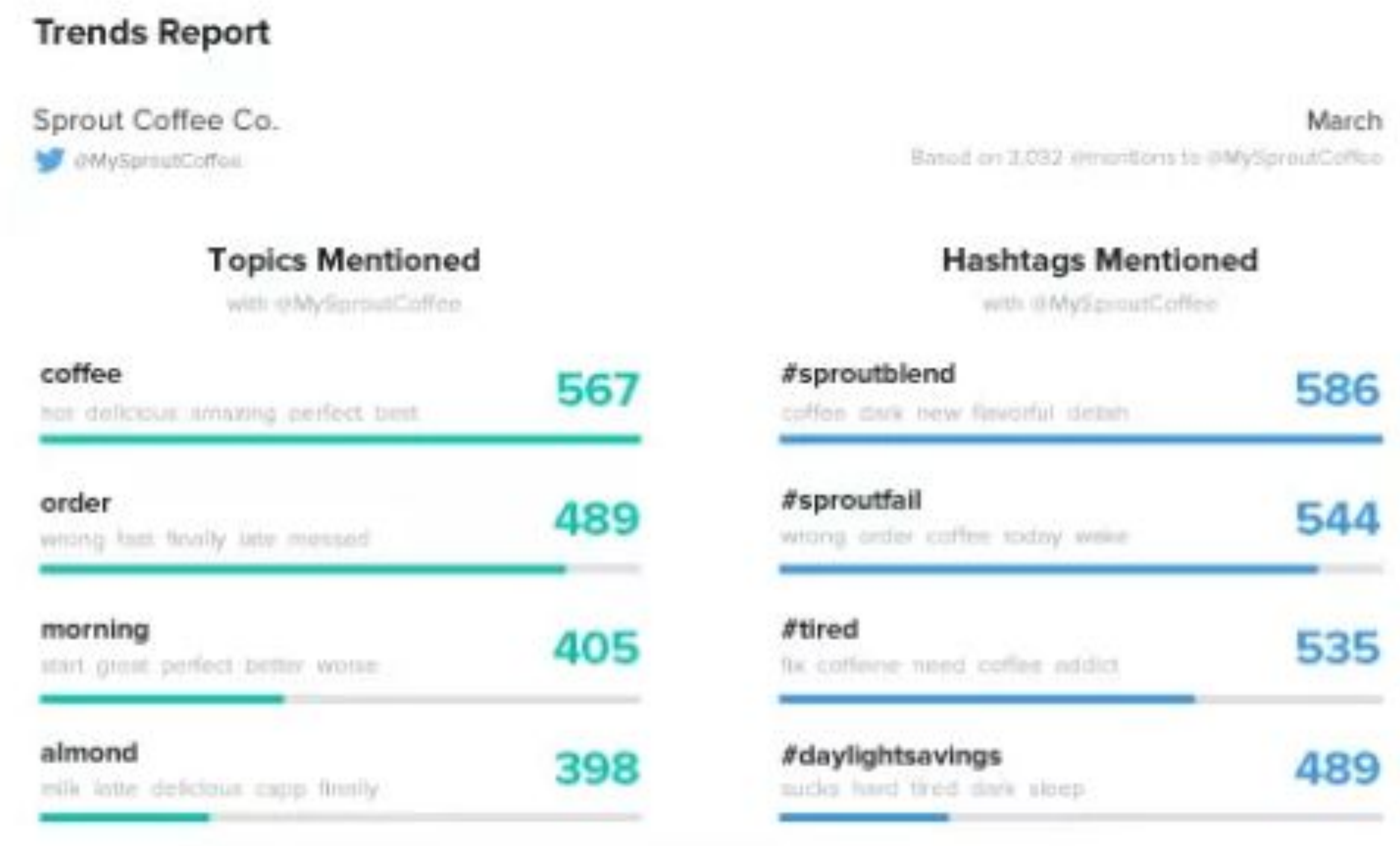
CASE PHASE
Post-Travel

INQUIRY TYPE
Question

CASE DETAIL
Lost Item

Social media trends

Sprout Social



Trends Report

Sprout Coffee Co.
@MySproutCoffee

March
Based on 3,032 @mentions to @MySproutCoffee

Topics Mentioned

with @MySproutCoffee

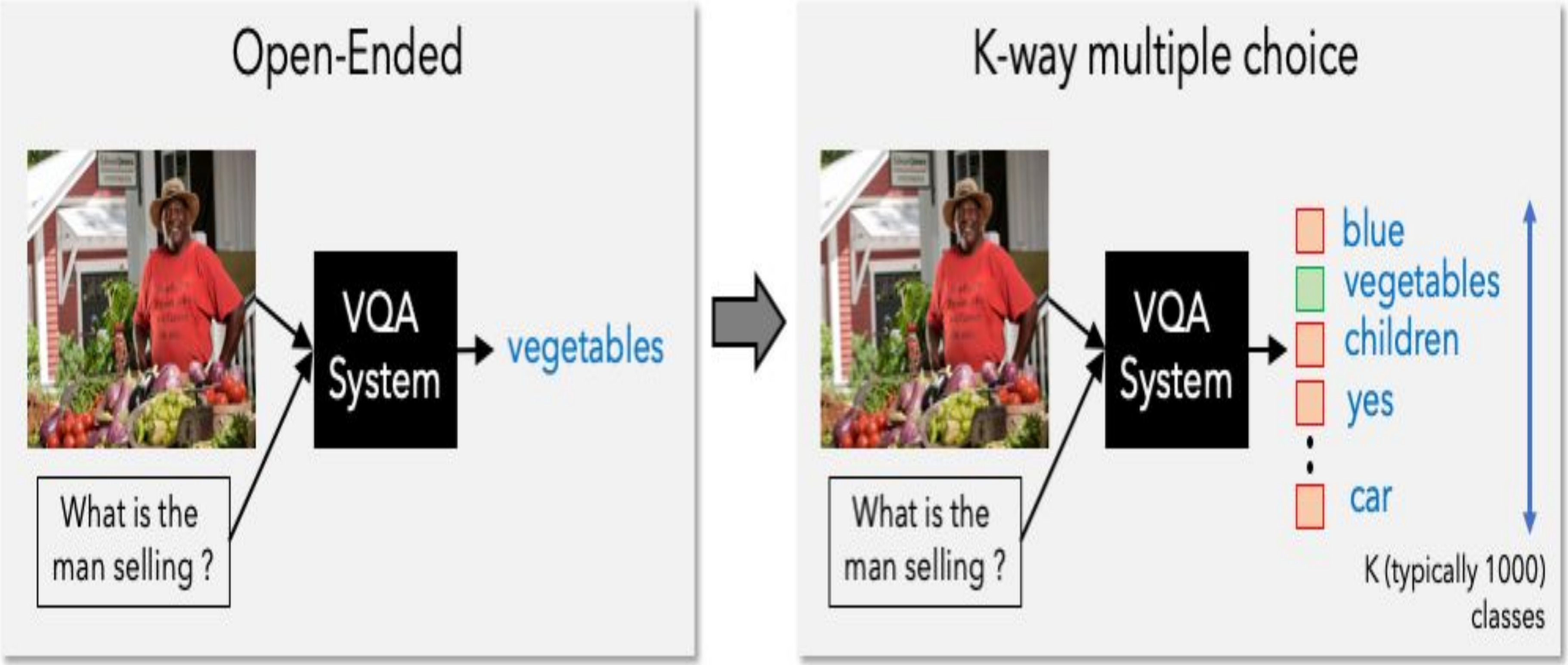
coffee hot delicious amazing perfect best	567
order wrong fast finally late messed	489
morning start great perfect better worse	405
almond milk latte delicious capp finally	398

Hashtags Mentioned

with @MySproutCoffee

#sproutblend coffee dark new flavorful delish	586
#sproutfail wrong order coffee today wake	544
#tired fix caffeine need coffee addict	535
#daylightsavings sucks hard tired dark sleep	489

Multiple choice vs open-ended settings



https://visualqa.org/

Medicine

The LSP-MLP helps enabling physicians to extract and summarize information of **any signs or symptoms, drug dosage and response data with aim of identifying possible side effects of any medicine** while highlighting or flagging data items.

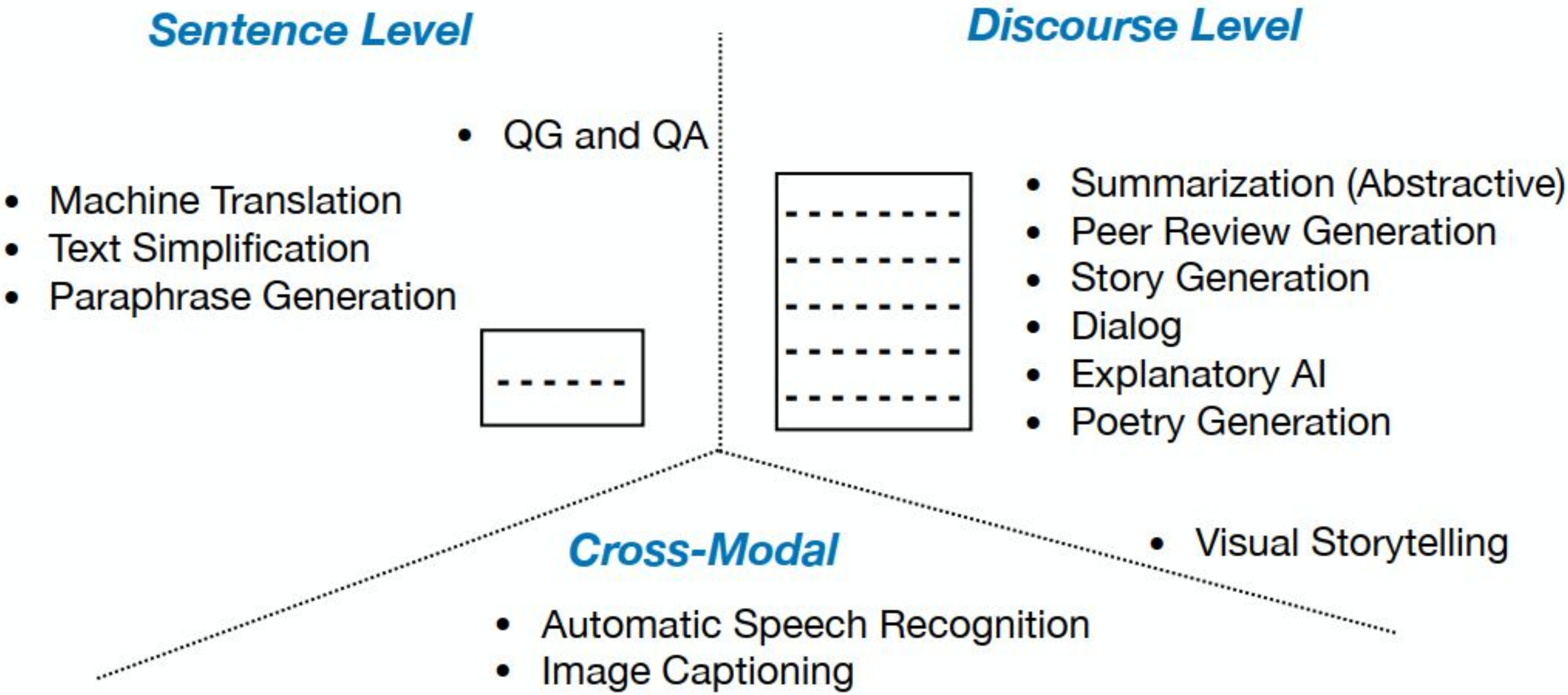
The Columbia university : NLP system called MEDLEE (MEDical Language Extraction and Encoding System) that identifies **clinical information in narrative reports and transforms the textual information into structured representation.**

https://pubmed.ncbi.nlm.nih.gov/28269895/

NLP in education

- Query expansion
- Summarization
- Translation
- Chatbot and Personal assistance

Applications of NLG



Sentence Level

- QG and QA

- Machine Translation
- Text Simplification
- Paraphrase Generation

Discourse Level

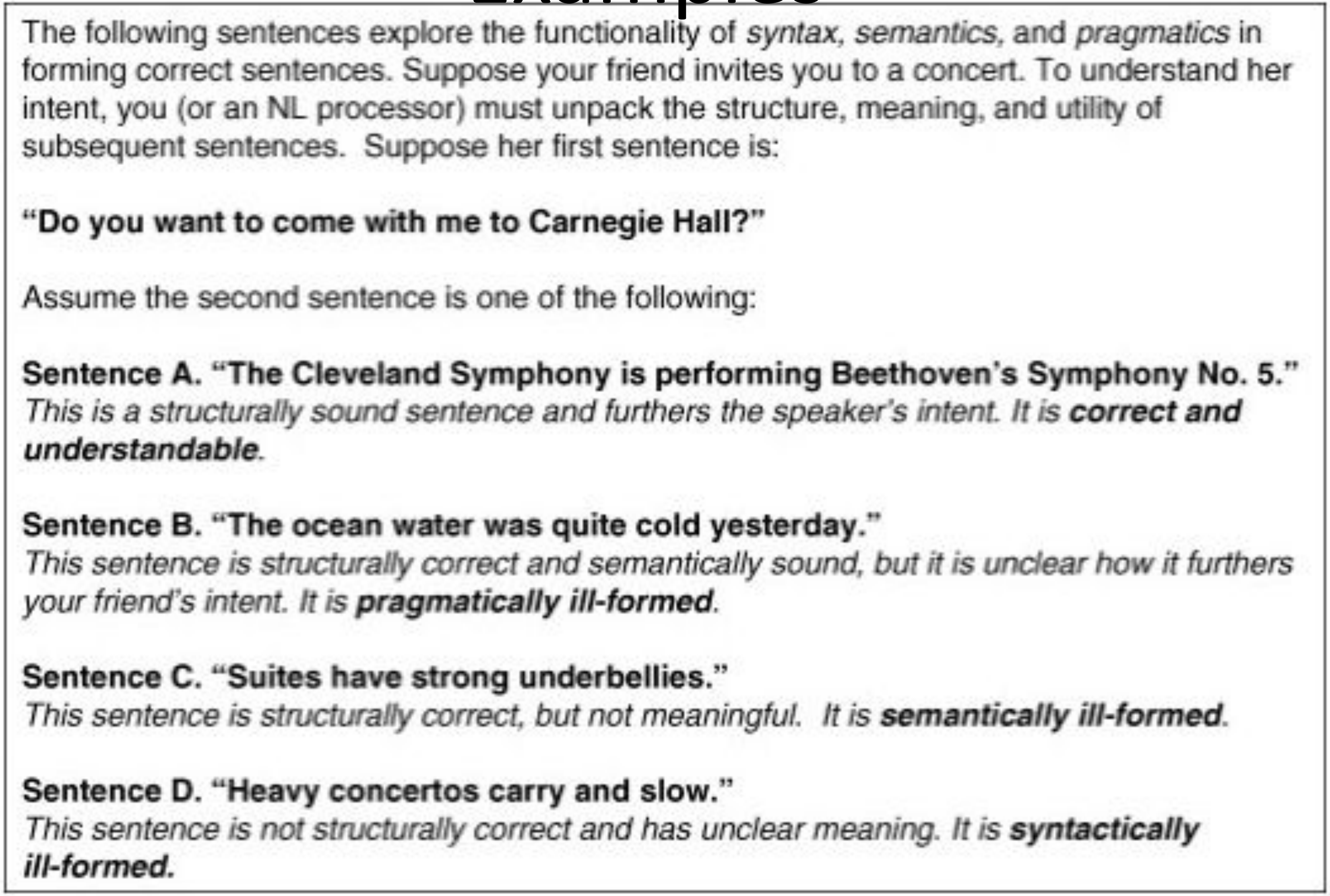
- Summarization (Abstractive)
- Peer Review Generation
- Story Generation
- Dialog
- Explanatory AI
- Poetry Generation

- Visual Storytelling

Cross-Modal

- Automatic Speech Recognition
- Image Captioning

Examples



The following sentences explore the functionality of *syntax, semantics,* and *pragmatics* in forming correct sentences. Suppose your friend invites you to a concert. To understand her intent, you (or an NL processor) must unpack the structure, meaning, and utility of subsequent sentences. Suppose her first sentence is:

"Do you want to come with me to Carnegie Hall?"

Assume the second sentence is one of the following:

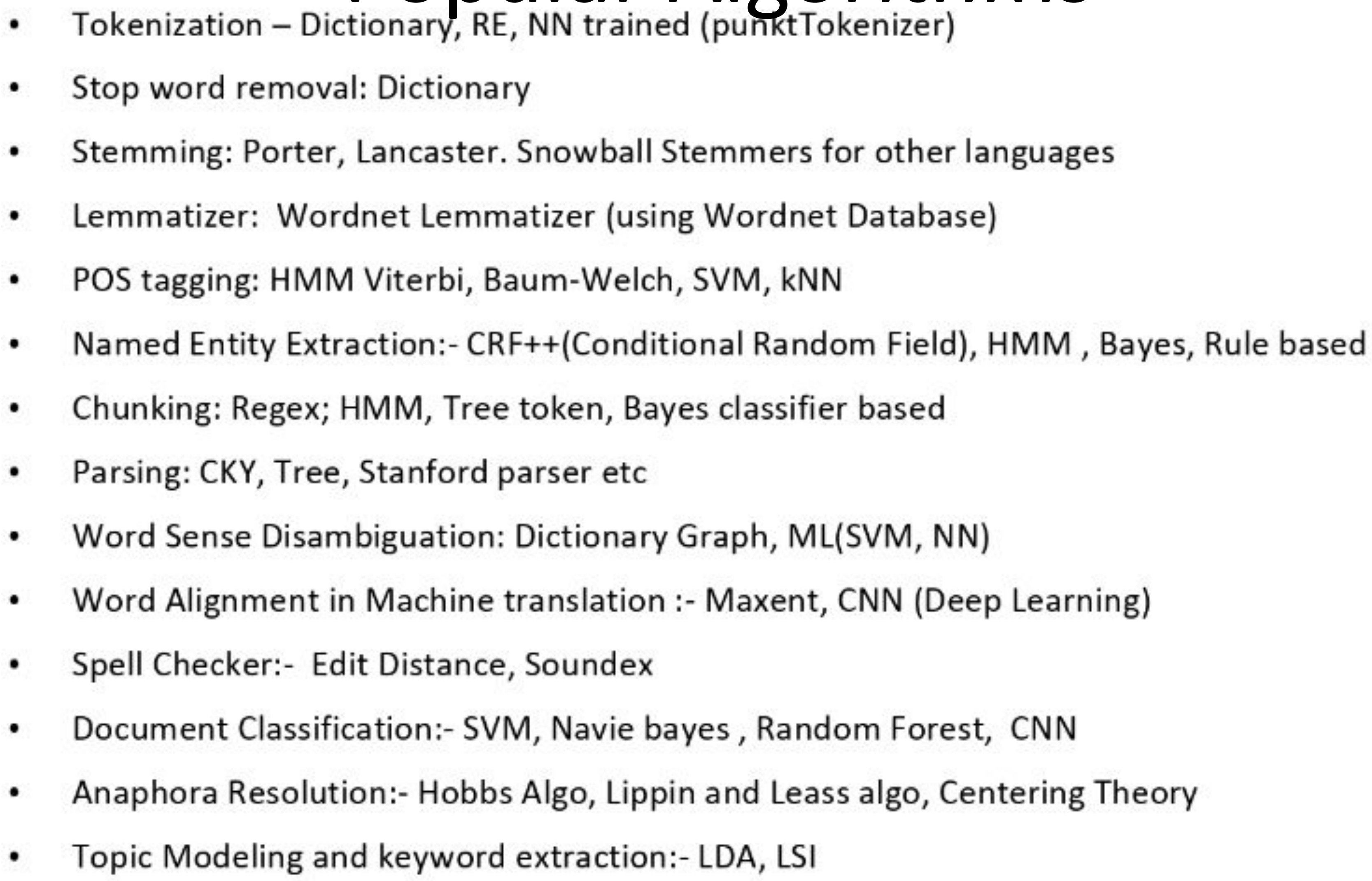
Sentence A. "The Cleveland Symphony is performing Beethoven's Symphony No. 5."
*This is a structurally sound sentence and furthers the speaker's intent. It is **correct and understandable**.*

Sentence B. "The ocean water was quite cold yesterday."
*This sentence is structurally correct and semantically sound, but it is unclear how it furthers your friend's intent. It is **pragmatically ill-formed**.*

Sentence C. "Suites have strong underbellies."
*This sentence is structurally correct, but not meaningful. It is **semantically ill-formed**.*

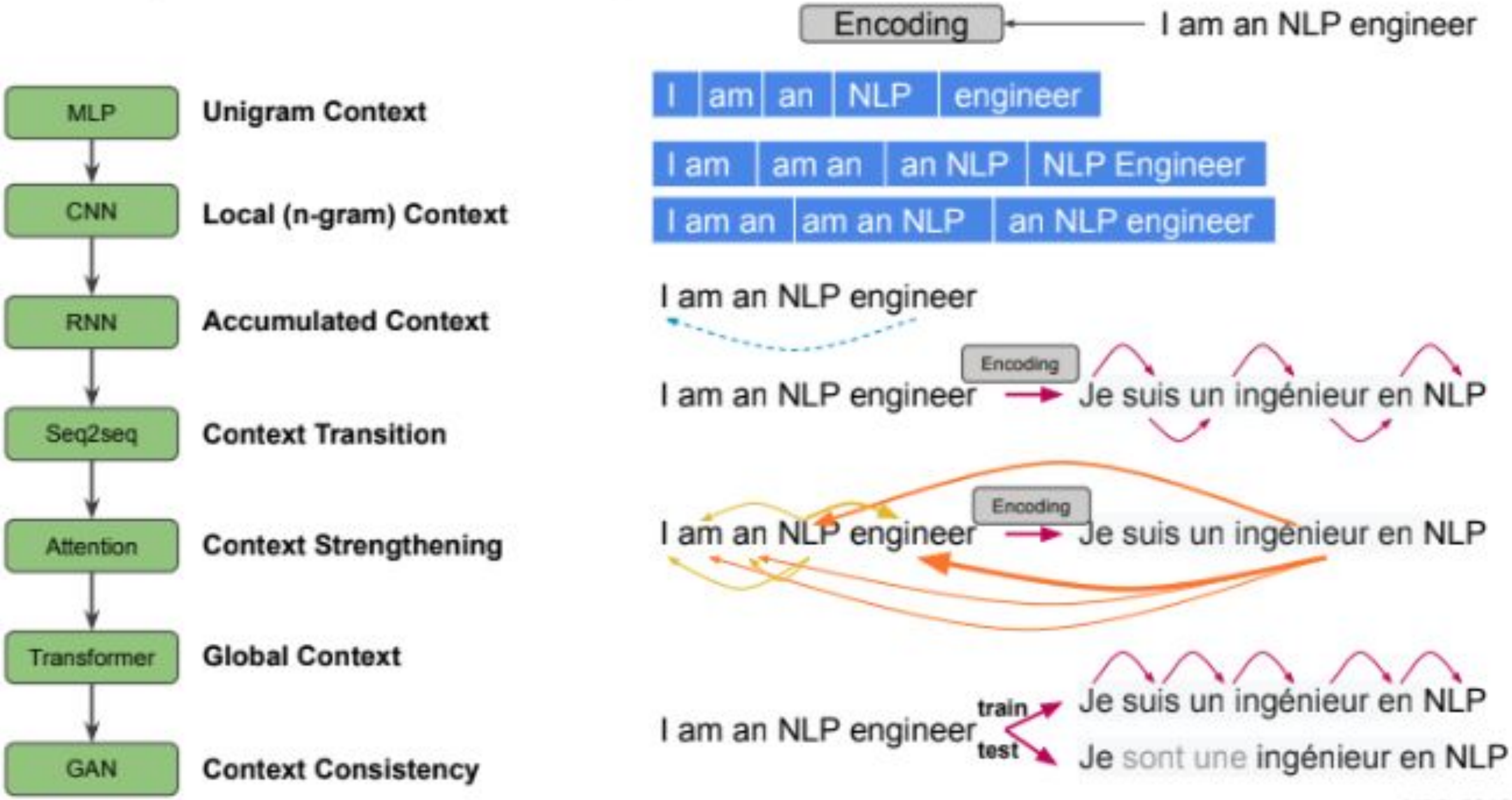
Sentence D. "Heavy concertos carry and slow."
*This sentence is not structurally correct and has unclear meaning. It is **syntactically ill-formed.***

Popular Algorithms



- Tokenization – Dictionary, RE, NN trained (punktTokenizer)
- Stop word removal: Dictionary
- Stemming: Porter, Lancaster. Snowball Stemmers for other languages
- Lemmatizer: Wordnet Lemmatizer (using Wordnet Database)
- POS tagging: HMM Viterbi, Baum-Welch, SVM, kNN
- Named Entity Extraction:- CRF++(Conditional Random Field), HMM , Bayes, Rule based
- Chunking: Regex; HMM, Tree token, Bayes classifier based
- Parsing: CKY, Tree, Stanford parser etc
- Word Sense Disambiguation: Dictionary Graph, ML(SVM, NN)
- Word Alignment in Machine translation :- Maxent, CNN (Deep Learning)
- Spell Checker:- Edit Distance, Soundex
- Document Classification:- SVM, Navie bayes , Random Forest, CNN
- Anaphora Resolution:- Hobbs Algo, Lippin and Leass algo, Centering Theory
- Topic Modeling and keyword extraction:- LDA, LSI

Sequential Modeling on Semantic Context



Encoding ← I am an NLP engineer

MLP → Unigram Context: I | am | an | NLP | engineer

CNN → Local (n-gram) Context: I am | am an | an NLP | NLP Engineer; I am an | am an NLP | an NLP engineer

RNN → Accumulated Context: I am an NLP engineer

Seq2seq → Context Transition: I am an NLP engineer → Encoding → Je suis un ingénieur en NLP

Attention → Context Strengthening: I am an NLP engineer → Encoding → Je suis un ingénieur en NLP

Transformer → Global Context

GAN → Context Consistency: I am an NLP engineer; train → Je suis un ingénieur en NLP; test → Je sont une ingénieur en NLP