T. Y. Artificial Intelligence & Data Science
Semester VI

Prof. Mandar Diwakar
Asst. Professor
Department of AI & DS
Vishwakarma Institute of Information Technology

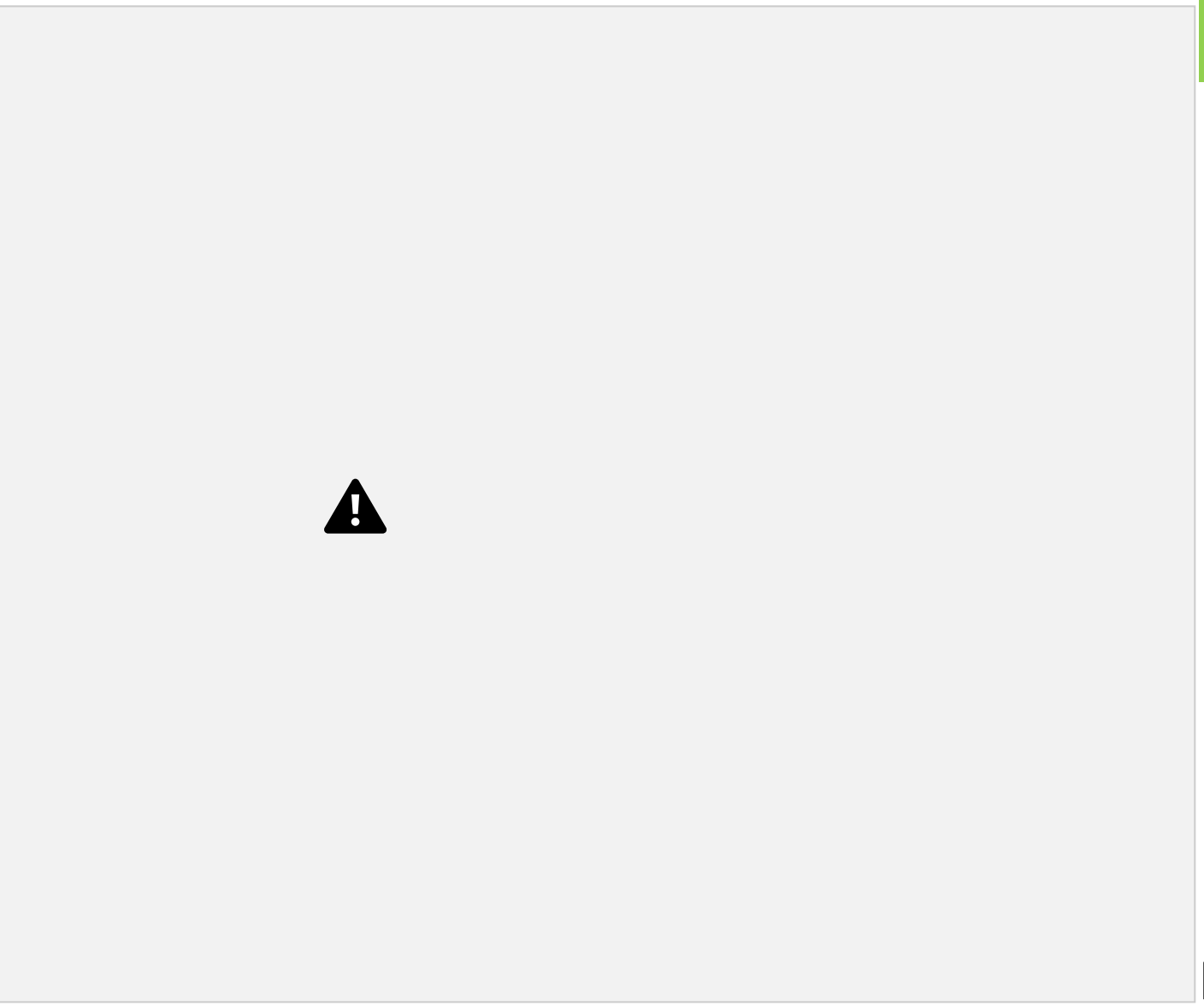
ADUA32203: Natural Language Processing

Unit I

Introduction Natural Language Processing

Unit I:	Introduction	4 Hrs
History of NLP, Generic NLP system, levels of NLP, Knowledge in language processing, Ambiguity in Natural language, stages in NLP, challenges of NLP, Applications of NLP, Approaches of NLP: Rule based, Data Based, Knowledge Based approaches **Case Studies:** Recent advances in NLP		

What Is NLP?



Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and

spoken words in much the same way human beings can.

What Is NLP?

- Humans communicate with each other using words and text. The way that humans convey information to each other is called Natural Language

- However, computers cannot interpret this data,
  which is in natural language, as they communicate in 1s and 0s.

- The data produced is precious and can offer valuable insights. Hence, ***you need computers to be able to understand, emulate and respond intelligently to human speech***.

- Natural Language Processing or NLP refers to

the branch of Artificial Intelligence that ***gives the machines the ability to read, understand and derive meaning from human languages.***

History of NLP

EARLY WORK - 1946

- Started as Machine Translation work
- Very simplistic approach
- Based on dictionary lookup
- Word ordering rules applied
- Poor results
- Language is ambiguous

History of NLP

1950S - CONFLICTING VIEWS

-Chomsky's Syntactic Structures book.
- Language is generative in nature
- Two trends emerged:
  a. Linguistics-based
  b. Statistical-based

1960S: ONE UP, THE OTHER DOWN

- Linguistics witnessed more development
- Computer scientists thought MT was possible
- Poor results
- Funding MT suspended
- MT halted altogether

History of NLP

1970S: DIVIDE BETWEEN LINGUISTICS AND CS

- Linguistics dominated by transformation model
- Transformation was very abstract
- Other attempts continued
- ELIZA and SHRDLV systems were developed

1990S: FIRST RESULTS

- Good progress achieved
- POS with good accuracy
- Electronic texts
- Fast CPUs and more memory
- The Internet

History of NLP

2000S UP TO DATE: HAND IN HAND

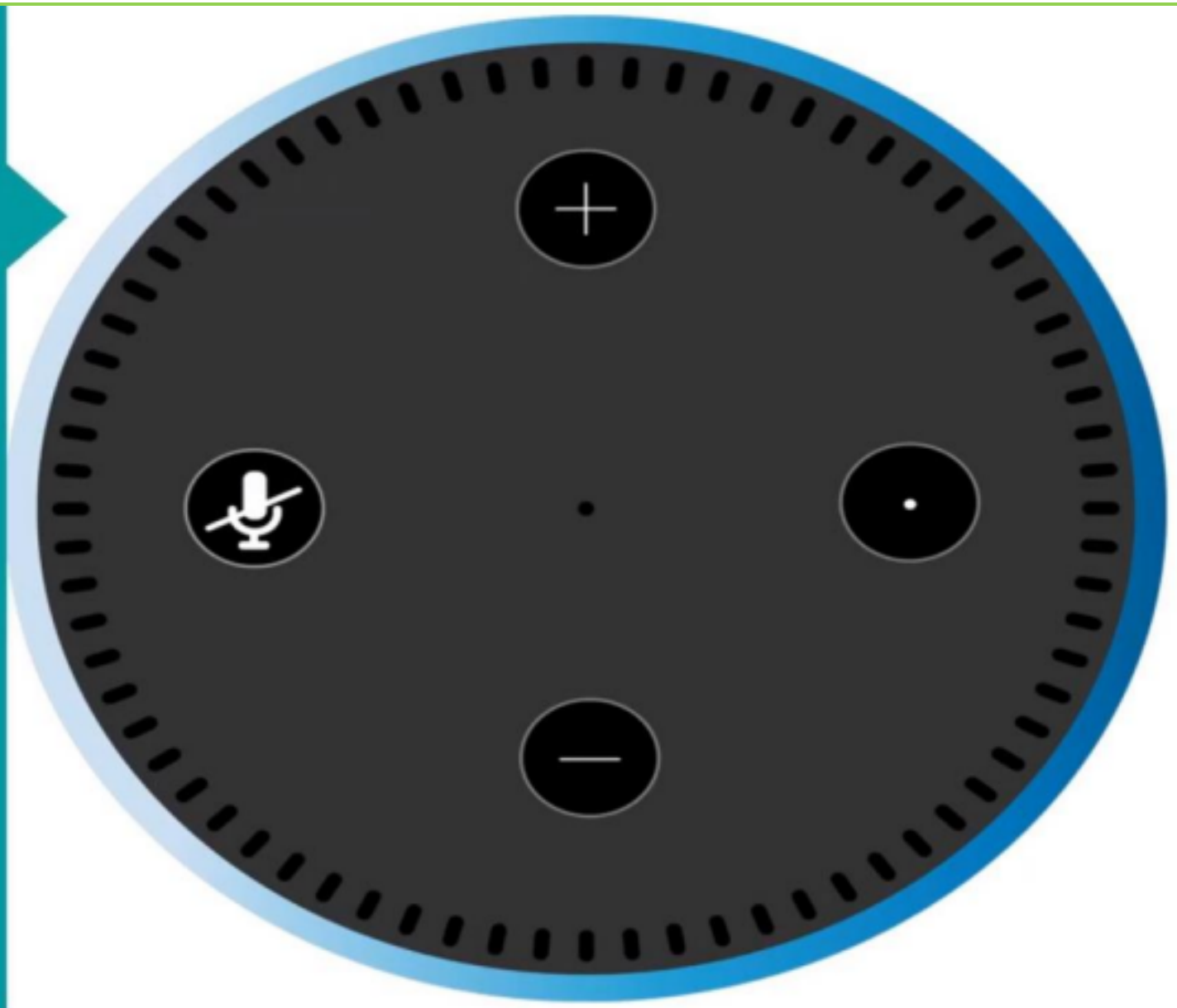
- Linguistics is critical in NLP
- Statistics and Artificial Intelligence played major roles
- Faster CPUs and cheaper storage
- Abundant resources
- Deep learning
- Amazon Alexa
- Sophia
- Yet, no 100% accuracy
- Several problems remain open

History of NLP

AMAZON ALEXA

* CHATTING
* QUESTION ANSWERING SYSTME
* TRACK APPOINTMENTS
* NOT SMART ENOUGH
* MOSTLY NOT SURE

www.lltacademia.com

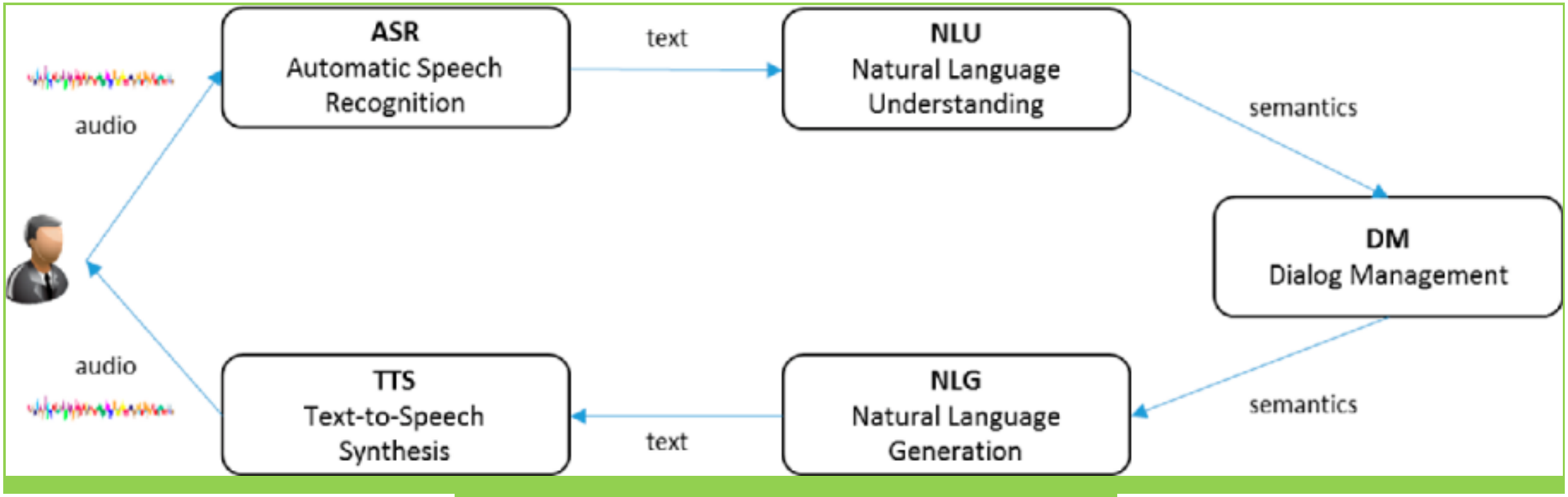


History of NLP

SOPHIA THE TALKING ROBOT

* SMARTER THAN ALEXA
* CAN ANSWER COMPLEX QUESTIONS
* CAN DO LIMITED REASONING
* STILL NEEDS IMPROVEMENTS
* SPEAKS ENGLISH ONLY

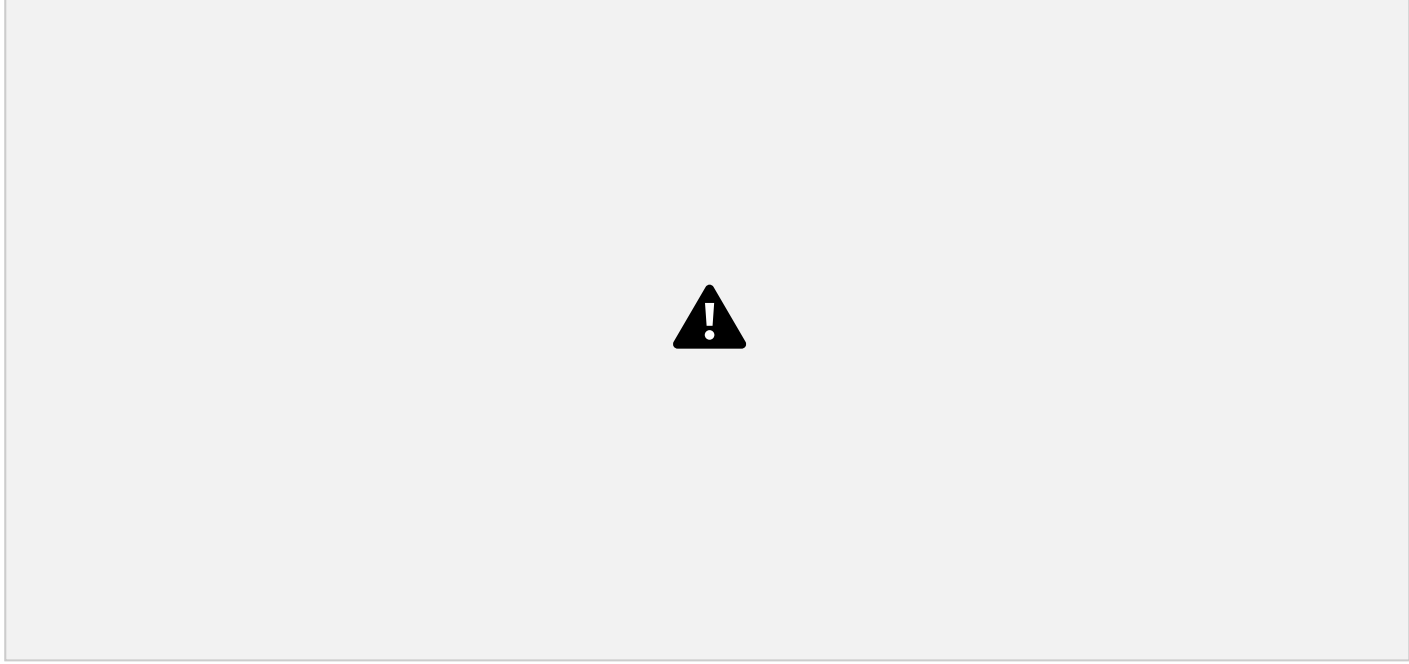
Generic NLP system



Generic NLP system

ASR & TTS

- **Automatic Speech Recognition (ASR),** or Speech to-text (STT) is a field of study that aims to transform raw audio into a sequence of corresponding words.



- **Speech synthesis,** or text-to-speech (TTS), is the computer-based creation of artificial speech from normal language text. Not to be confused with recorded audio playback, TTS is computer-generated speech formed from text.

Generic NLP system

What is natural language understanding (NLU)?

- Natural language understanding is an artificial intelligence technology who's ***main job is understanding spoken or written words and phrases***.

- ***It turns language, known technically as 'unstructured data', into a 'machine readable' format, known as 'structured data'.*** This enables other computer systems to process the data to fulfil user requests.

- Most of the time, NLU is found in **chatbots, voicebots** and **voice assistants**, but it can theoretically be used in any application that aims to understand the meaning of typed text.

Generic NLP system

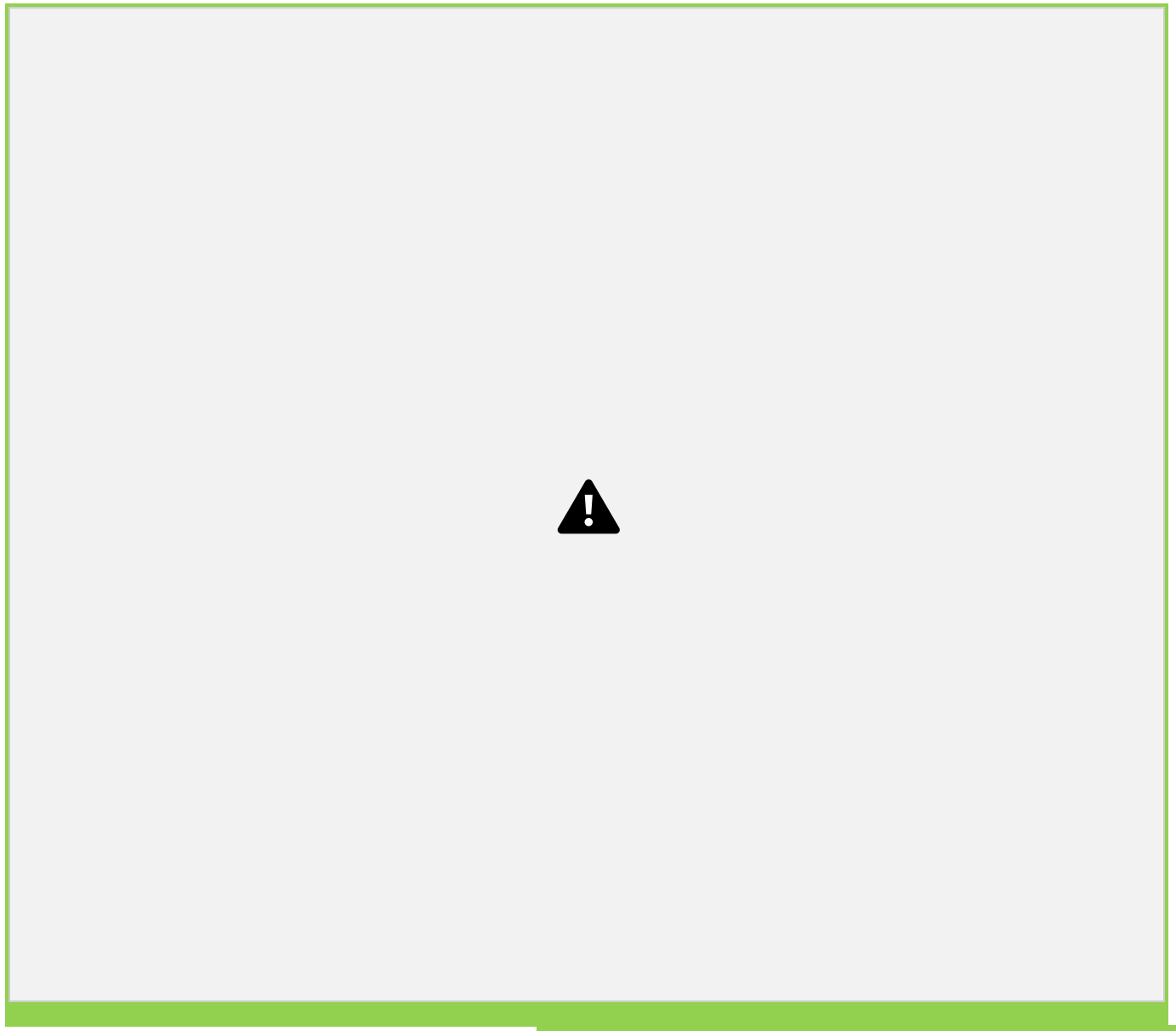
What is Natural Language Generation(NLG)?

• NLG is a process to **produce meaningful sentences** in Natural Language.

• It **explains the structured data in a manner that is easy to understand for humans** with a high speed of thousands of pages per second. •
• NLG uses **ML** and **Deep Learning** Techniques

• Some of the **NLG models** are listed below:
1. Markov chain
2. Recurrent neural network (RNN)
3. Long short-term memory (LSTM)
4. Transformer

levels of NLP

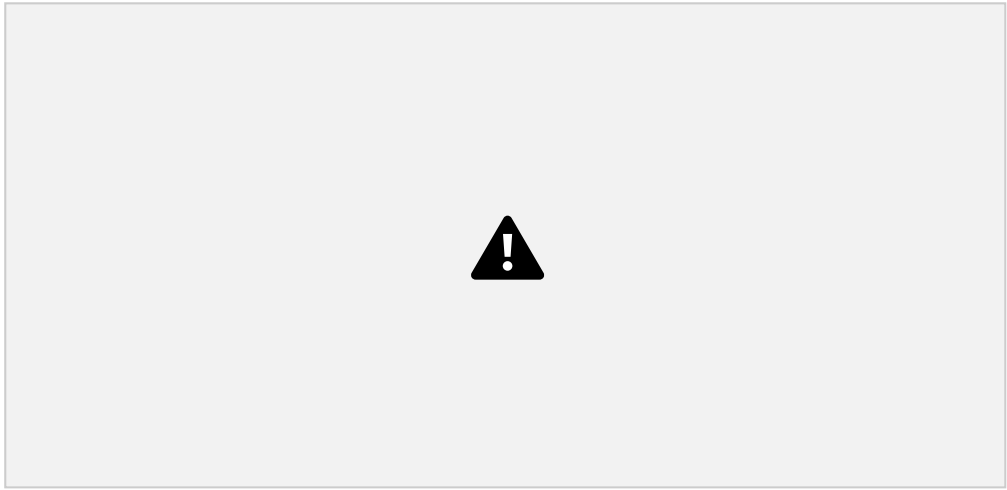


morphological level

- **The morphological level of linguistic processing deals with the study of word structures and word formation, focusing on the analysis of the individual components of words.**
- Words are formed of morphemes, which are the minimal (that is, non-decomposable) linguistics units that carry meaning

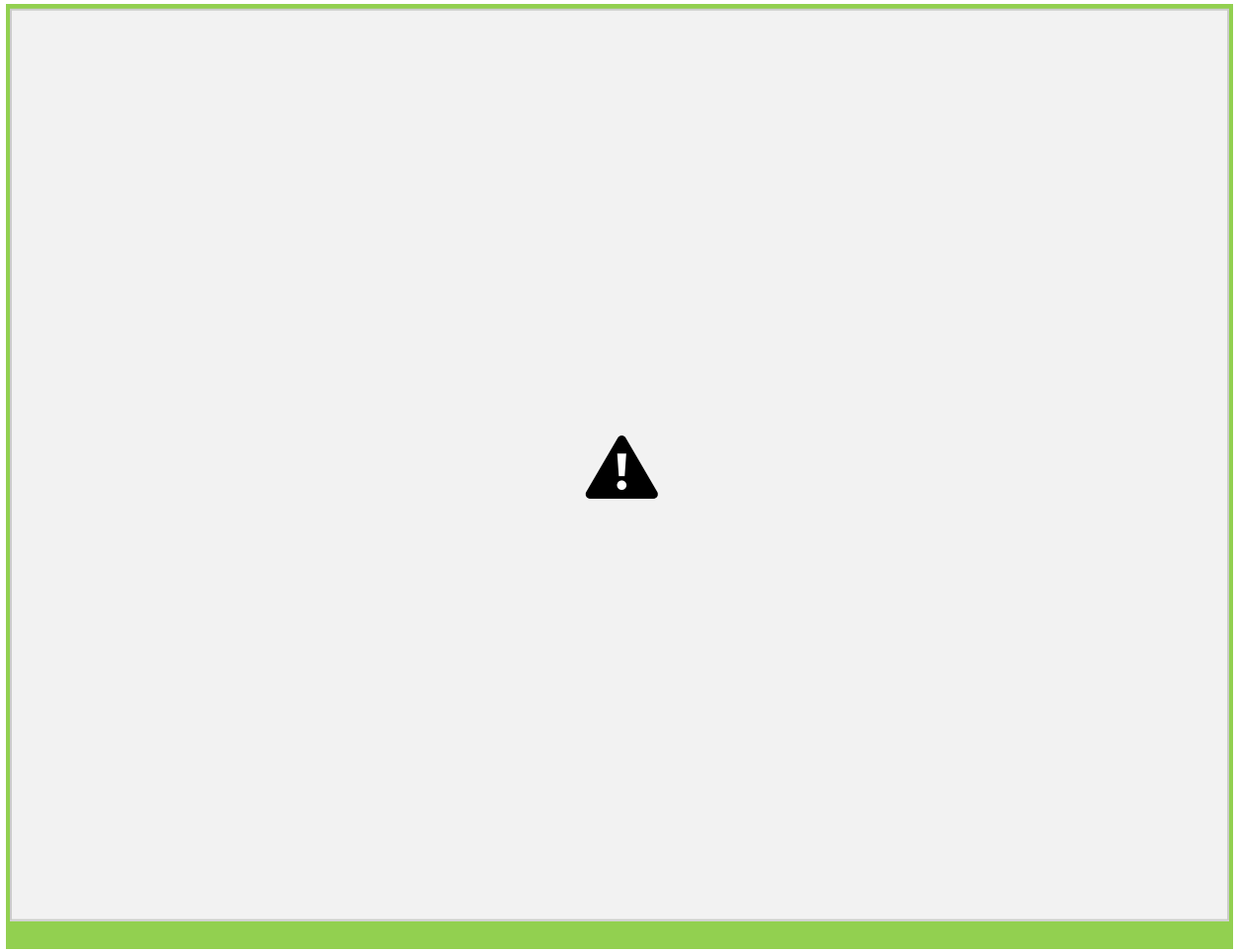
➢ Taking, for example, the word: "unhappiness".

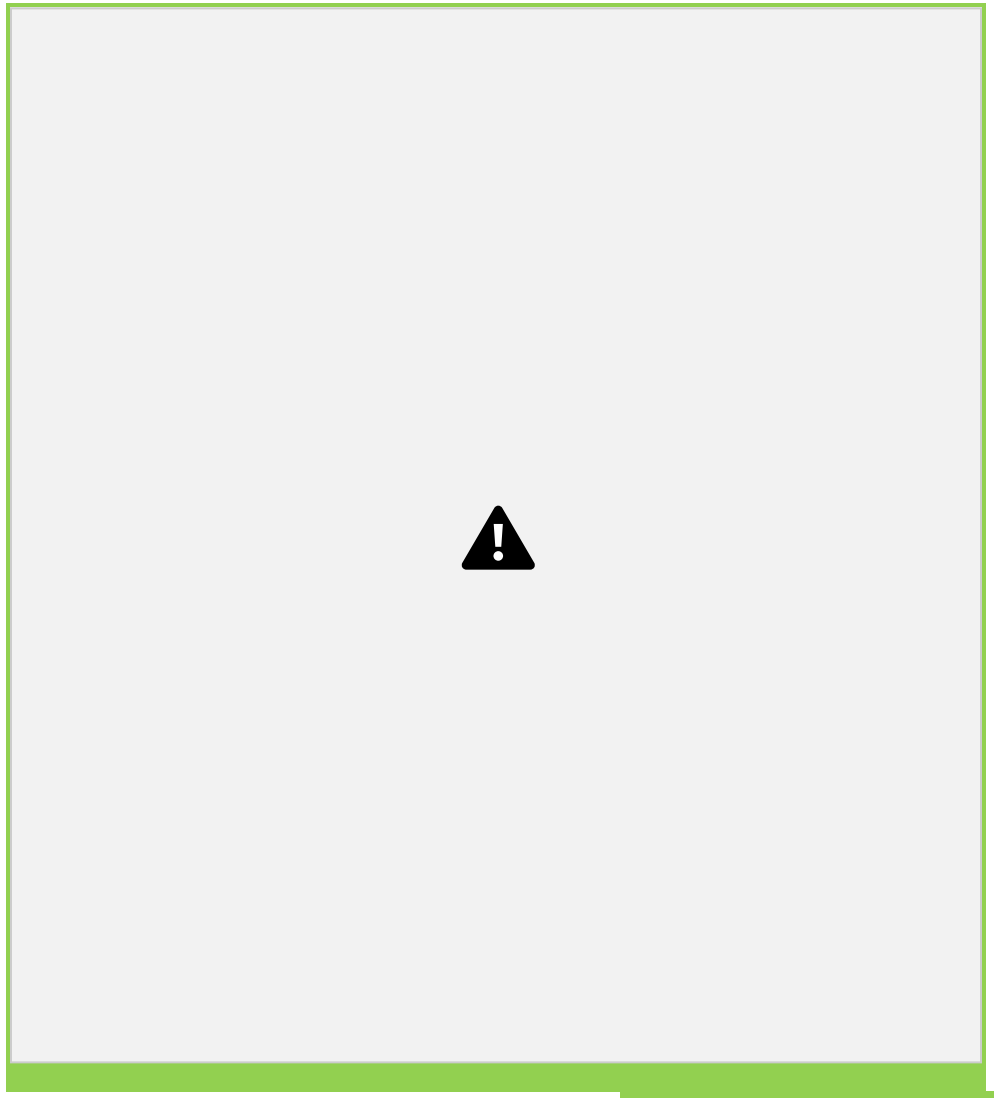
➢ It can be broken down into three morphemes (prefix, stem, and suffix), with each conveying some form of meaning: the prefix un- refers to "not being", while the suffix -ness refers to "a state of being".



➢ The stem happy is considered as a free morpheme since it is a "word" in its own right.

morphological level





lexical analysis level

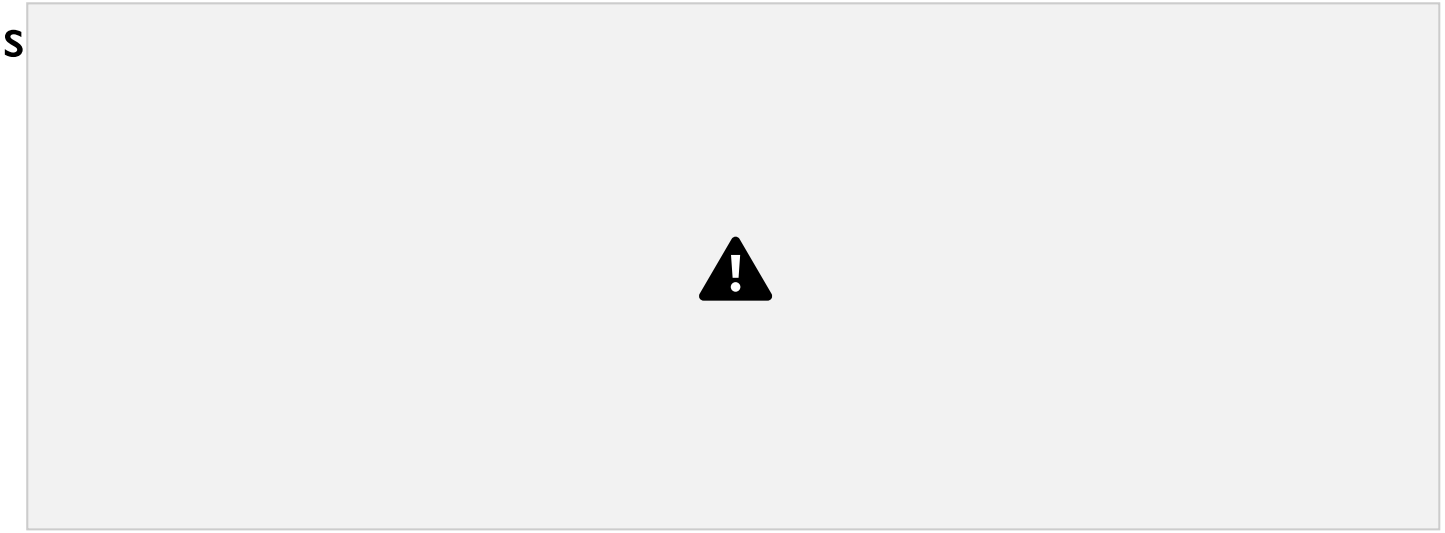
- The lexical analysis in NLP deals with the study at the level of words with respect to their lexical

meaning and part-of-speech.

- This level of linguistic processing utilizes a language's lexicon, which is a collection of individual lexemes.
- A lexeme is a basic unit of lexical meaning; which is an abstract unit of morphological analysis that represents the set of forms or "sens

My third cat **ate** its meal

My cat **ate** its third meal



Drinking too much alcohol can harm your health

Sleeping well directly affects your mental and physical health

Syntactic Analysis level

- The syntax of the input string refers to the arrangement of words in a sentence so they

grammatically make sense.

- NLP uses syntactic analysis to asses whether or not the natural language aligns with grammatical or other logical rules.

1. Lemmatization / Stemming - ***reduces word complexity to simpler forms*** that have less variation. Lemmatization uses a dictionary to reduce the natural language to its root words. Stemming uses simple matching patterns to strip away suffixes such as 's' and 'ing'.

2. Parsing -This is the process of ***undergoing grammatical analysis of a given sentence***. A common method is called Dependency Parsing, which assesses the relationships between words in a sentence.

3. Word Segmentation -This is the ***separation of continuous text into separate words***. In English this is easy because all words are usually separated by spaces, but for some languages like Japanese and Chinese they do not mark spaces for words. This is when word segmentation becomes very useful.

Syntactic Analysis level

For example, ***the cat chases the mouse in the garden***, would be represented as:

np - noun phrase

vp - verb phrase
s - sentence
det - determiner (article)
n - noun
tv - transitive verb (takes an object)
iv - intransitive verb
prep - preposition
pp - prepositional phrase
adj - adjective

Here the sentence is broken down according to the categories. Then it is described in a hierarchical structure with nodes as sentence units.

Semantic Analysis level

- Semantic analysis is concerned with the meaning representation.

- It mainly focuses on the literal meaning of words, phrases, and sentences.

- Semantic analysis starts with lexical semantics, which studies individual words' meanings (i.e., dictionary definitions).

- Semantic analysis then examines relationships between individual words and analyzes the meaning of words that come together to form a sentence.

- This analysis provides a clear understanding of words in context.

- For example, it provides context to understand the following sentences:

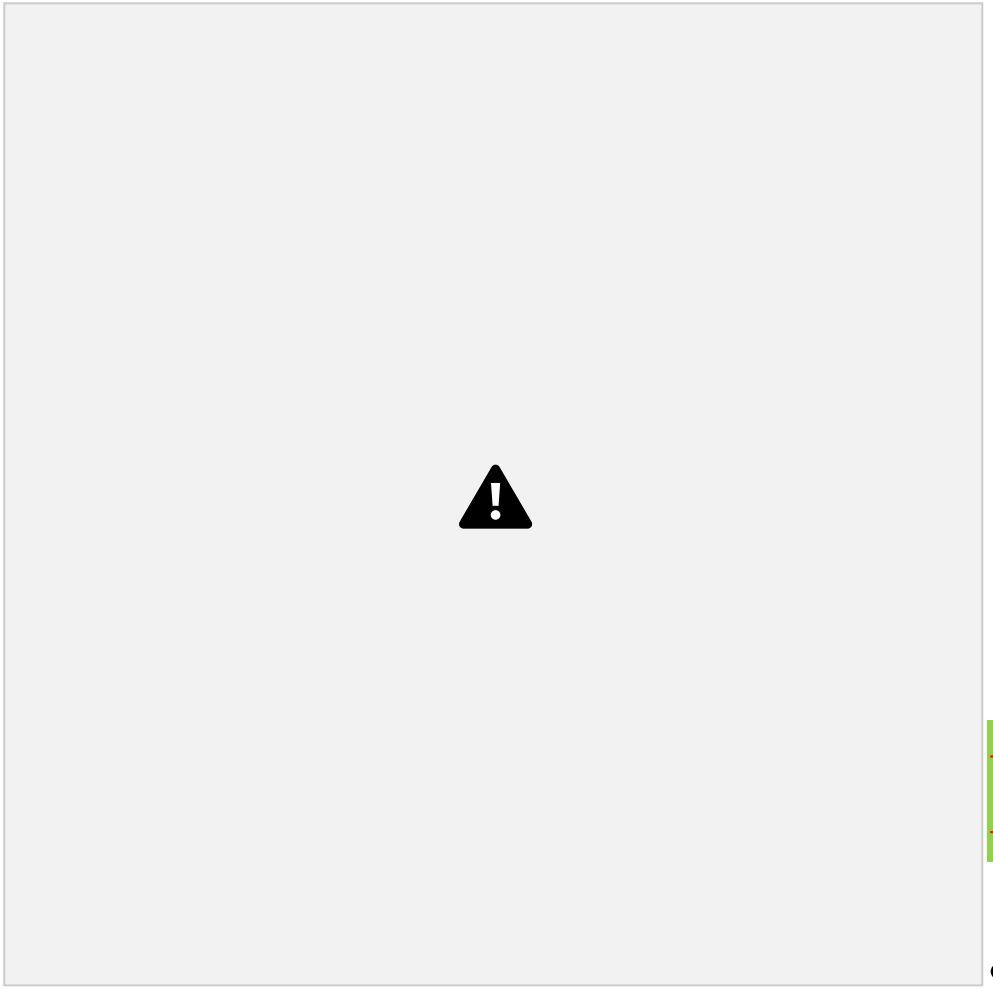
"The boy ate the apple" -→ defines an apple as a fruit.

"The boy went to Apple" -→ defines Apple as a brand or store.

Semantic Analysis level

Discourse level

It deals with the analysis of structure and meaning of text beyond a single sentence, making connections between words and sentences.



“I love dominoes pizza because they put extra cheese” , she said.

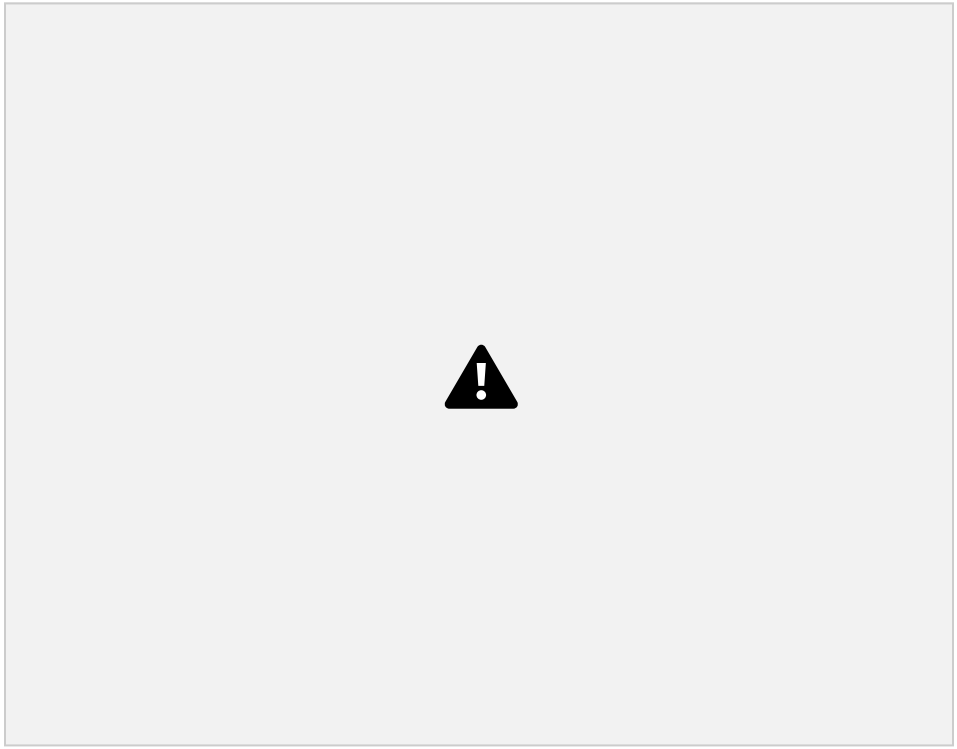
Here there are two entities she and dominoes, where she is in context of “I” and they is in context of “dominoes” so discourse will interpret this sentence has 2 entities (I and dominoes) and 2 anaphor (she and they)

Pragmatic level

- The pragmatic level of linguistic processing deals

with the ***use of real-world knowledge*** and ***understanding of how this impacts the meaning*** of what is being communicated.

• Pragmatics is the ***study of meaning in context dependent on the intentions*** of participants in a conversational exchange.



Knowledge in language processing

Knowledge required in NLP

- ❑ Phonetic and Phonological knowledge
- ❑ Morphological Knowledge
- ❑ Syntactic Knowledge
- ❑ Semantic knowledge
- ❑ Pragmatic Knowledge
- ❑ Discourse Knowledge
- ❑ Word knowledge

What Is NLP?

➢ Phonetic and Phonological Knowledge

1. Phonetics is the study of **language at the level of sounds**.
2. phonology is the study of the **combination of sounds into organized units of speech**.
3. Both deal with how words are related to the sounds that realize them.

➢ Morphological Knowledge

1. Morphology concerns **word-formation**.
2. It is a study of the patterns of formation of words by the combination of sounds into minimal distinctive units of meaning called **morphemes**.
3. Morphological Knowledge concerns **how words are constructed from morphemes**.

What Is NLP?

➢ Syntactic Knowledge

1. The syntax is the level at which we study how words combine to form phrases, phrases combine to form clauses and clauses join to make sentences.

2. It deals with how words can be put together to form correct sentences.

➢ Semantic Knowledge

1. It concerns the meaning of the words and sentences.

2. Defining the meaning of a sentence is very difficult due to the ambiguities involved.

➢ Pragmatic Knowledge

1. Pragmatics is the extension of the meanings or semantics. Pragmatics deals with the contextual aspects of meaning in particular situations.

2. It concerns how sentences are used in different situations.

What Is NLP?

➢ **Discourse Knowledge**

1. Discourse concerns connected sentences. It includes the study of chunks of language which are bigger than a single sentence.
2. It concerns inter-sentential links that is how the immediately preceding sentences affect the interpretation of the next sentence.
3. It is important for interpreting pronouns and temporal aspects of the information conveyed.

➢ **Word Knowledge**

1. It is nothing but everyday knowledge that all the speakers share about the world.
2. It includes the general knowledge about the structure of the world and what each

language user must know about the other user's beliefs and goals.

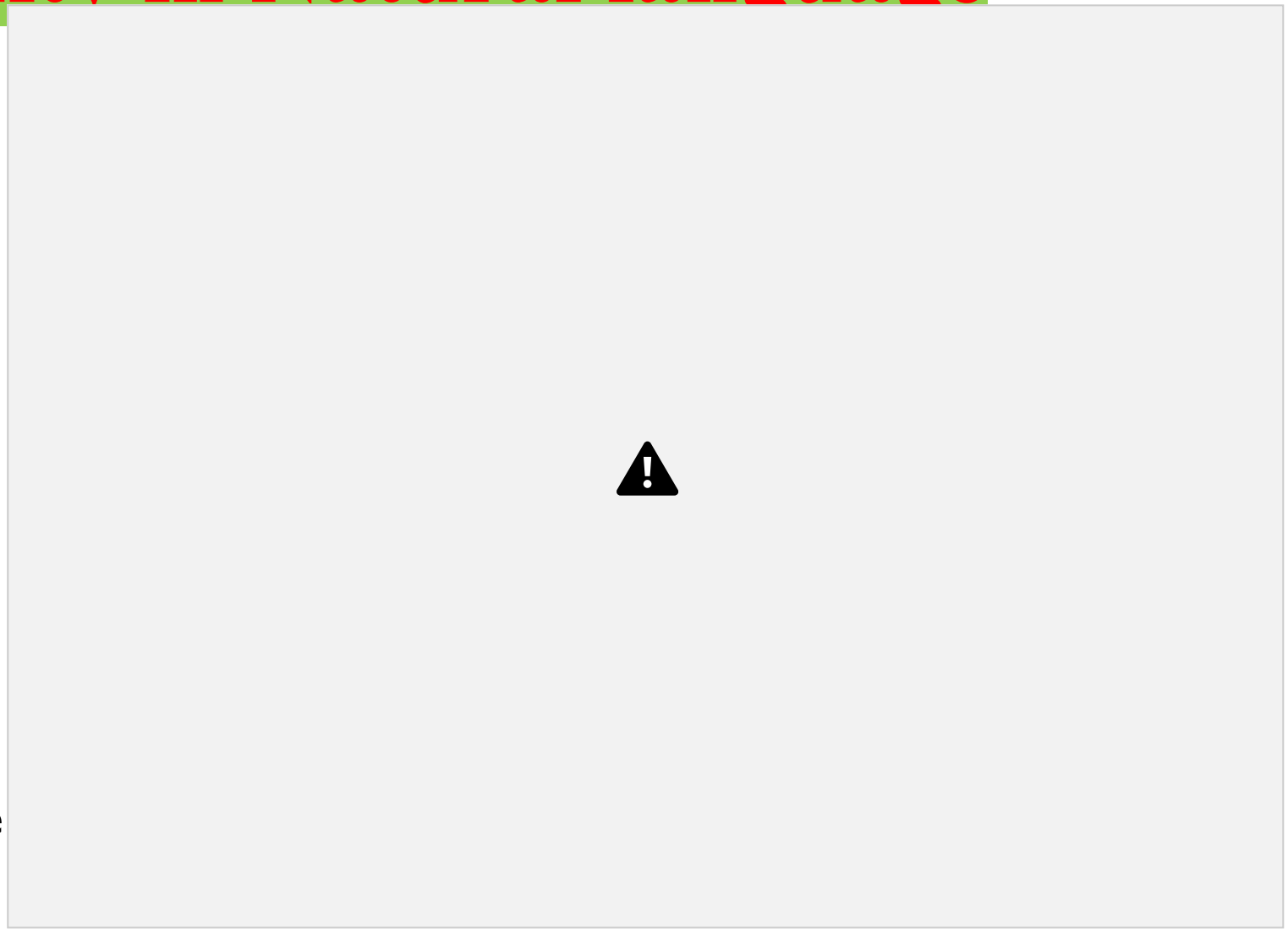
3. This is essential to make the language understanding much better.

Ambiguity in Natural language

➢ Linguistic ambiguity

- Linguistic ambiguity is a quality of language that makes speech or written text open to multiple interpretations.

- That quality makes the meaning difficult or impossible for a person or artificial intelligence (AI) program to reliably decode without some additional



information.

Ambiguity in Natural language

➢ Lexical Ambiguity

The ambiguity of a single word is called lexical ambiguity.

For example, **bank :- River Bank, Financial Institution, Power bank**

treating the word **silver** as a noun, an adjective, or a verb.

silver used as a noun:

- A lustrous, white, metallic element, atomic number 47, atomic weight 107.87, symbol Ag.
- (collectively) Coins made from silver or any similar white metal.
- (collectively) Cutlery and other eating utensils, whether silver or made from some other white metal.
- (collectively) Any items made from silver or any other white metal.
- A shiny gray color.

silver used as an adjective:

- Made from silver.
- Made from another white metal.
- Having a color like silver: a shiny gray.
- Denoting the twenty-fifth anniversary, especially of a wedding.

Ambiguity in Natural language

➢ **Syntactic Ambiguity**

This kind of ambiguity occurs when a ***sentence is parsed in different ways***. For example

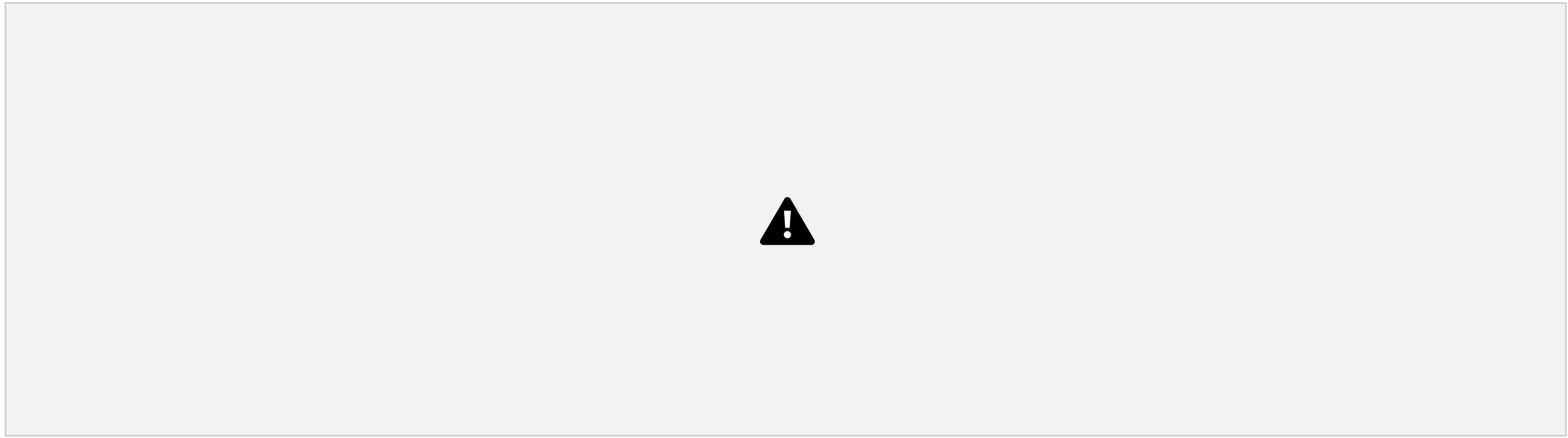
1. The professor said on Monday he would give an exam. 2. The chicken is ready to eat .

3. The burglar threatened the student with the knife. 4. Visiting relatives can be boring .

Ambiguity in Natural language

➢ **Semantic Ambiguity**

This kind of ambiguity occurs when the ***meaning of the words themselves can be misinterpreted***.



Ambiguity in Natural language

➢ **Anaphoric Ambiguity**

This kind of ambiguity arises due to the ***use of anaphora entities in discourse***

Example

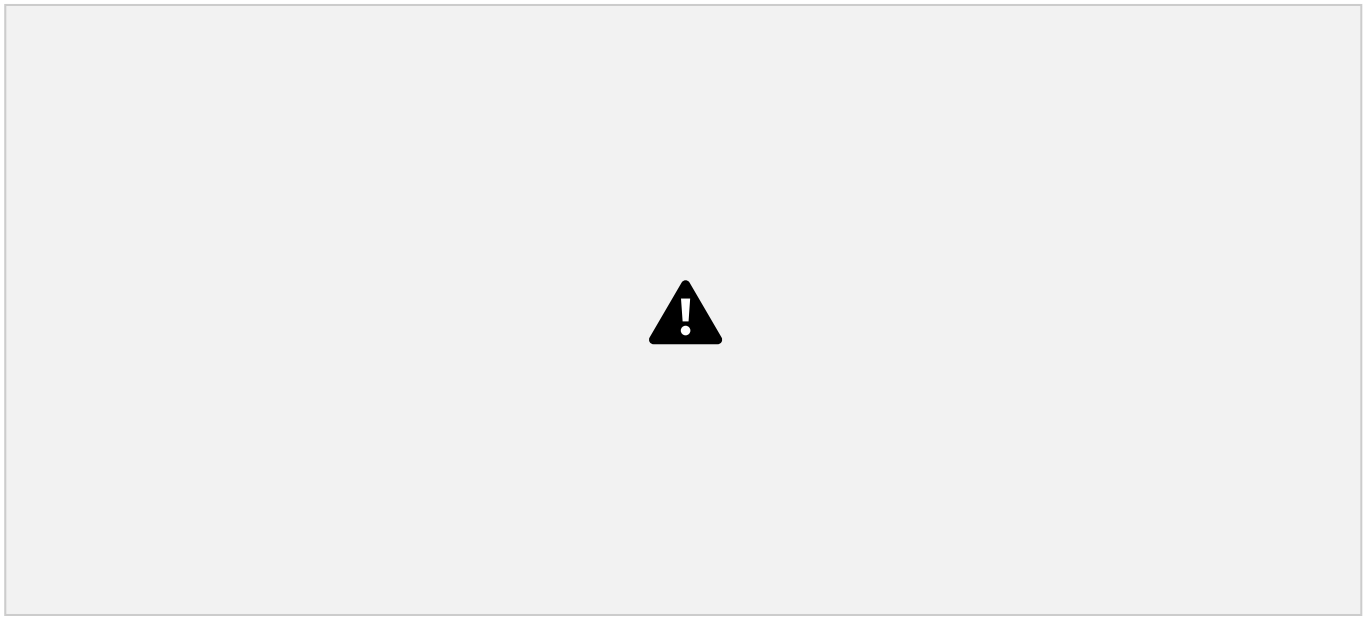
The horse ran up the hill. It was very steep. It soon got tired. The horse ran up the hill. It was very steep. It soon got tired.

Ambiguity in Natural language

➢ **Pragmatic ambiguity**

Ambiguity refers to the situation where the context of a phrase gives it multiple interpretations

pragmatic ambiguity arises when the statement is not specific.



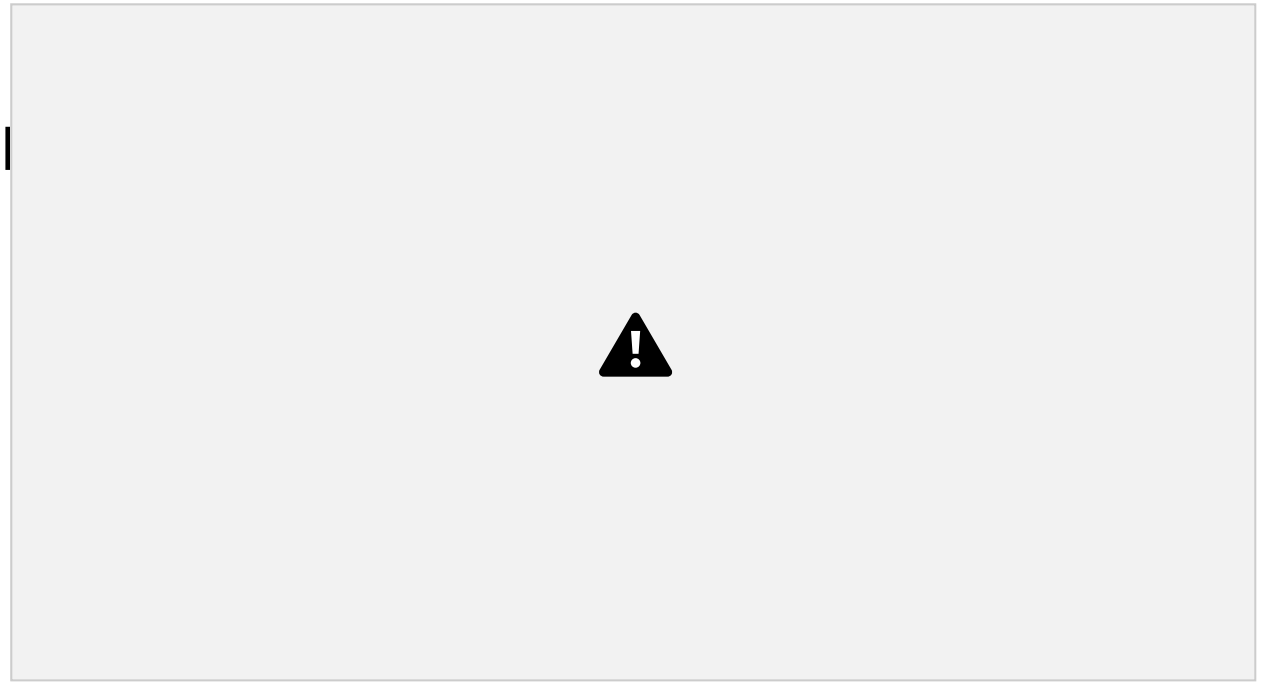
Challenges NLP

➢ **Contextual words and phrases and homonyms**

1. The **same words and phrases can have different meanings** according the context of a

sentence

2. Many words have the exact same



I **ran** to the store because
we **ran** out of milk.

Challenges NLP

➢ **Synonyms**

1. we use many different words to express the same idea.
2. Some words may convey exactly the same meaning, while some may be levels of complexity and different people use synonyms to denote slightly different meanings within their personal vocabulary.

small, little, tiny, minute

ability - capability, competence, skill

beautiful - attractive, pretty, lovely, stunning

create - generate, make, produce

Challenges NLP

➢ **Irony and sarcasm**

1. Irony is when something happens that is the opposite of what was expected.

Fire chief's house burning down (situational irony)

Saying, 'it's a great time to go for a swim,' during the winter (verbal irony)

2. Sarcasm means that the speaker intends to convey the opposite of the meaning of their words

1. Dinner rehearsal*: Yes, because we can't practice eating enough, can we?* ***2.***

After making a mistake*: You did a brilliant job.*

Challenges NLP

➢ **Errors in text and speech**

1. Misspelled or misused words can create problems for text analysis.

Autocorrect and grammar correction applications can handle common

mistakes, but don't always understand the writer's intention.

2. With spoken language**, mispronunciations, different accents, stutters,** etc., can be difficult for a machine to understand.

Challenges NLP

➢ **Slang**

Slang is vocabulary that is used between people who belong to the same social group and who know each other well. Slang is very informal language With spoken language.

GOAT : - the Greatest Of All Time."

Example: *Record for the most number of runs in T20Is with 3932 runs in 105 innings,*

*Virat Kohli really is the **GOAT** cricketer.*

Lit : - When something is lit, it's awesome or very enjoyable. When something is really lit, it's straight fire — flawless, incredible.

Example: *Last night's party was **lit***.

Challenges NLP

➢ **Colloquialism**

• Colloquial language shows up in your conversations with family and neighbors. • It's the phrases in your journal and in texts to your friends.

Deadset — True

Bloody — Very, but in a slightly profane way

Rubbish — Trash, or an exclamation meaning something is the same quality as trash

Flat out — Extremely busy

Hard to swallow = difficult to believe

You're not **gonna** change any of them

Applications NLP

➢ Email filters

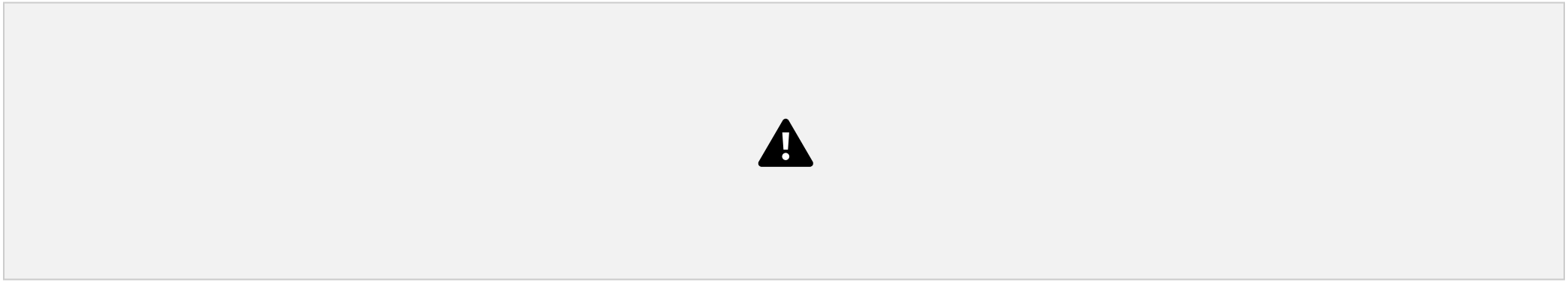
• It began with spam filters, which ***identified specific words or phrases that indicate a spam message***.

• **Gmail's** email categorization is one of the more common, newer implementations of NLP. Based on the contents of emails, ***the algorithm determines whether they belong in one of three categories (main, social, or promotional).***

• This maintains your **inbox manageable for all Gmail users**, with critical,

relevant emails

you want to see and reply to fast.



Applications NLP

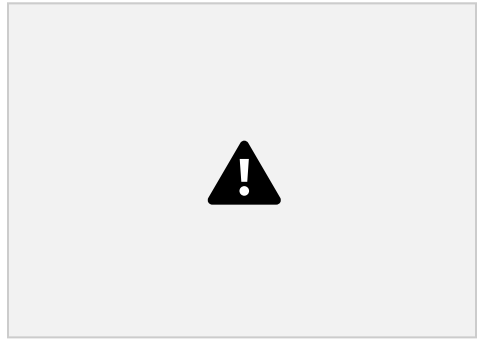
➢ Smart assistants

• Voice recognition allows smart assistants like ***Apple's Siri*** and ***Amazon's Alexa*** to identify patterns in speech, infer meaning, and offer a helpful answer.

• ***voice technology can detect spoken sounds and recognize them as***

words. ASR is the cornerstone of the entire voice experience, allowing computers to finally understand us through our most natural form of communication: speech.

• And when we talk with Siri or Alexa through ***products like the thermostat, light switches, automobiles, and more,*** we're becoming accustomed to seeing them crop up across our house and everyday lives.



Applications NLP

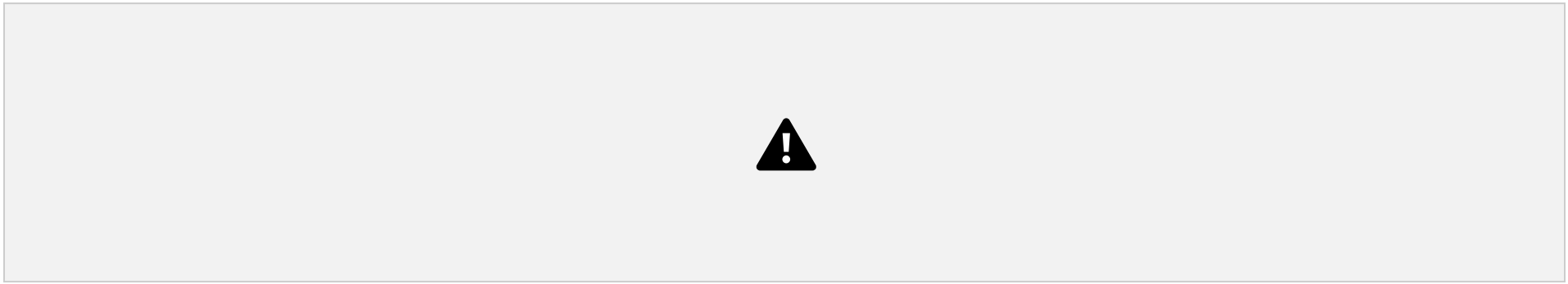
➢ **Search results**

➢ NLP process built on Bidirectional Encoder Representations from Transformers or ***BERT***.

➢ By analyzing individual words in the ***body of a text in relation to every other***

word in the same body of text, the algorithm can gain a more complete picture of the text then simply ***analyzing each word one-by-one***.

➢ Google searches were ***best at returning results that matched the structure or text of a search,*** but not necessarily the intended meaning.



Applications NLP

➢ Predictive text

- Things like ***autocorrect, autocomplete,*** and ***predictive text*** are so commonplace

on our

smartphones that we take them for granted.

- Autocomplete and predictive text are similar to search engines in that ***they predict things to say based on what you type***, finishing the word or suggesting a relevant one.
- Autocorrect will sometimes even change words so that the overall message makes more sense.

Applications NLP

➢ Social Media Monitoring

- People are increasingly using social media to express their ***opinions on a product,***

policy, or issue.

• These may *offer* ***important information about a person's preferences and dislikes***. • Companies now ***utilize a variety of NLP approaches to evaluate social media postings*** and learn what their consumers think about their products. • Companies also use social media monitoring to have a better understanding of the ***challenges and problems that their consumers*** are encountering as a result of utilizing their goods.

• It is used by the government as well as businesses to ***identify possible dangers to national security.***

Approaches- NLP

➢ **Rule-based NLP**

• Rules are considered an outdated approach to text processing.

• They're ***written manually and provide some basic automatization*** to routine tasks.

- For example,

you can write rules that will allow the system to identify ***an email address in the text because it has a familiar format***, but as soon as any variety is introduced, the system's capabilities end along with a rule writer's knowledge.

Approaches- NLP

➢ **Rule-based NLP**

Properties

• Rule-based approaches ***tend to focus on pattern-matching or parsing*** •

It can often be thought of as "fill in the blanks" methods

• Rule-based approaches are ***low precision, high recall***,

• They can have high performance in specific use cases, but often suffer performance degradation when generalized

Approaches- NLP

➢ **Machine learning-based NLP**

• A machine learning-based NLP system relies on ***more modern 'statistical inference' techniques.***

- It's ***more intelligent*** and understands speech and text in a similar way to how humans do.
- Once it's learned to understand human language in a particular environment—say, the legal world—it can infer the meaning of misspellings, omitted words, and new words without a human setting up a new rule.
- Machine-learning also learns the patterns between phrases and sentences and is constantly optimizing and evolving itself so that it's level of accuracy is getting ever closer to reality.