ECII/ECSI 3206:
Artificial Intelligence [and expert systems]
Topic 8: Natural Language Processing

By: Edgar Otieno

Introduction to NLP

- **NLP-**defines a method of communication with A.I system using natural language such as English . It may be in the form of speech , written text.
- It involves making computers perform useful tasks with the natural language human use.
- **Text mining**-deriving meaningful insight from normal language text [emails, sms, wordmaps]
- NLP advanced over the years by combining the power of artificial intelligence (AI), computational linguistics, and computer science.

Why NLP

- Tones of data are generated Within Digital systems : (https://www.internetlivestats.com/one-second/#email-band accessed on 17/5/2021)
  - **3,036,757** Emails sent in 1 second
  - **90,135** YouTube videos viewed in 1 second
  - **9,520** Tweets sent in 1 second
  - **1,079** Instagram photos uploaded in 1 second
  - **5,744** Skype calls in 1 second
  - **92,878** Google searches in 1 second
  - **120,587GB** of Internet traffic in 1 second

Cont..

- NLP technology can be used to extracting information from **unstructured** text such as emails, newspaper articles, and user reviews into structured text. **Structured** text mostly takes the form of tables or values in a structured form.
- **Entity extraction** refers to extracting entities from the text such as organizations, people, locations and so on. For example, the World Health Organization, IBM, Sara, John, Paris, US.
- **Relation extraction** refers to identifying the relationship between entities, for example, “Bob Collymore was a CEO at Safaricom”; “Salesio is the Director of SCIT”.
- **Text summarization** refers to the technique of shortening long pieces of text. Automatic text summarization is a common use case in machine learning and natural language processing.

Applications of NLP

- Auto complete[for google searches,]
- Optical character recognition (OCR) from documents (text that is scanned from actual documents)
- Spam mail detection
- Predictive typing[texting systems]
- Spell checkers
- Chatbots
- Machine translation
- Natural language assistants, such as Siri and Alexa
- Sentiment Analysis[Analysis of social media content to help determine public opinion]

Selected Application Examples

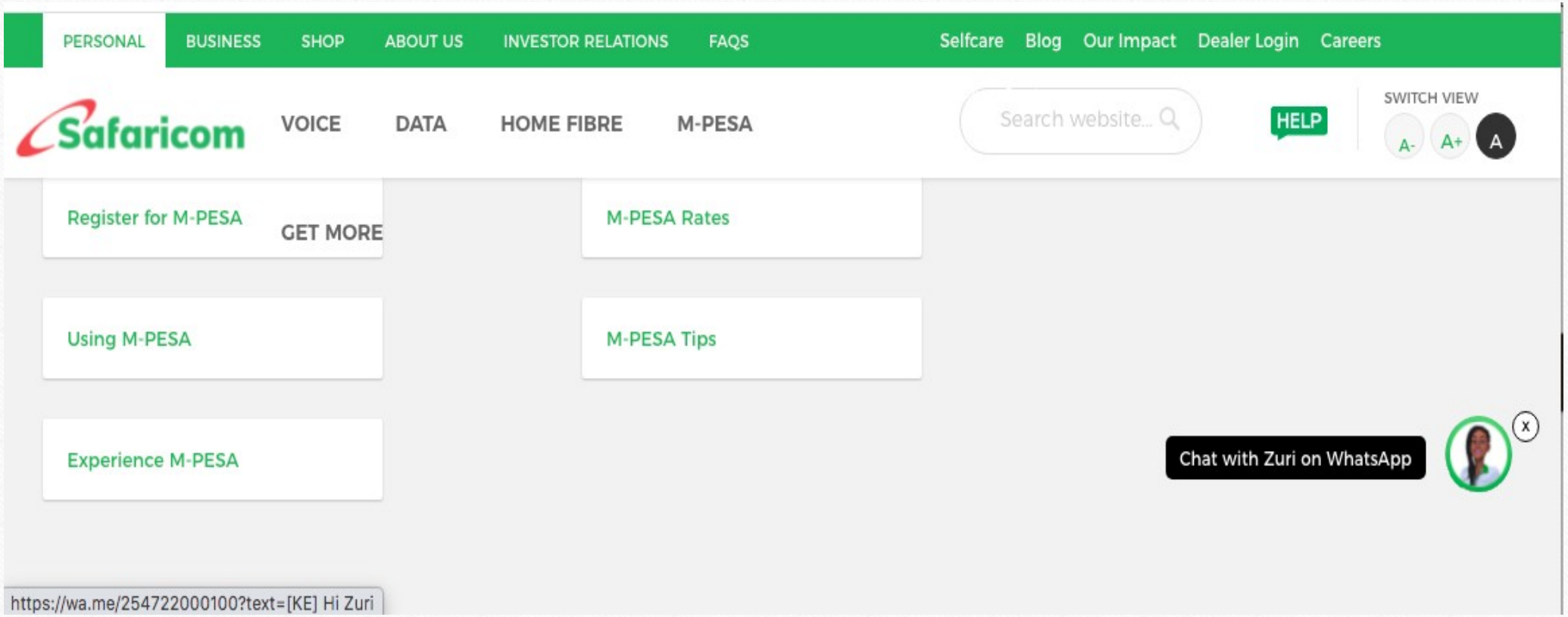
1. Sentiment Analysis:

- The process of identifying emotions or opinions that are expressed in user input.
- Sentiment analysis answers various questions, such as how people feel about your product or whether your customers are satisfied with your customer service .
- This analysis is used in marketing, retention plans, and emotional intelligence for chatbots

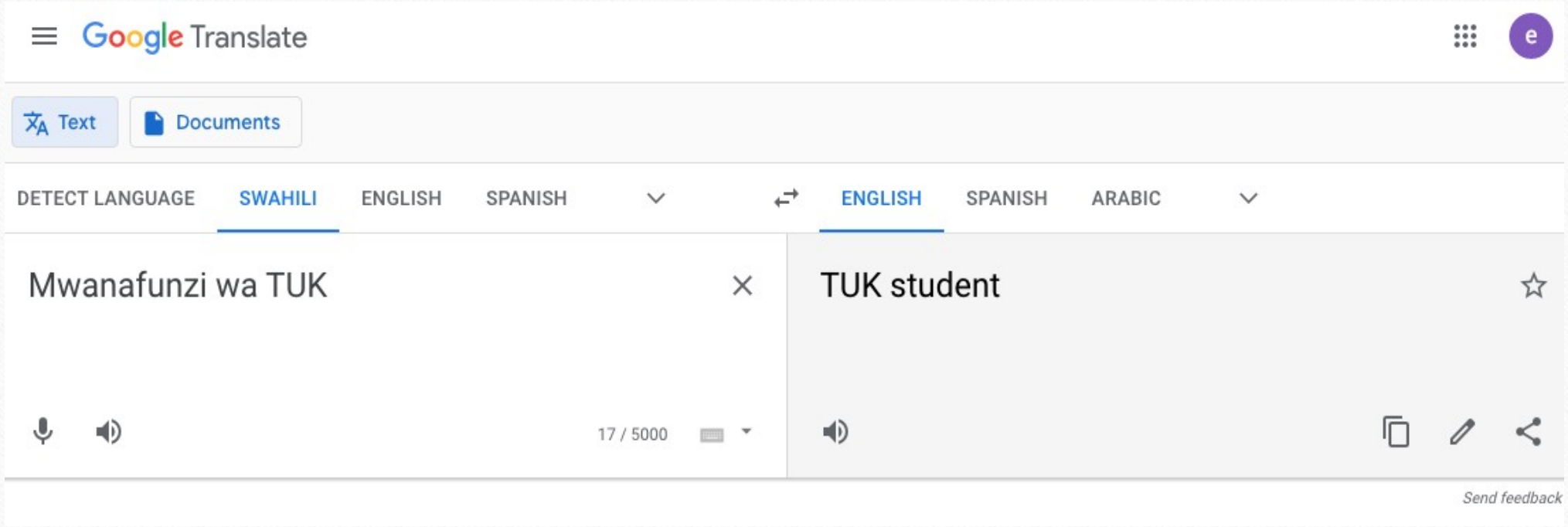
Cont…

2. Chatbots

- Building solutions that can answer questions that are asked by humans in natural language. A question and answering system can be used for the following tasks:
  - Retrieving answers from forums.
  - Building a Frequently Asked Questions (FAQs) system.
  - Training chatbots


PERSONAL
BUSINESS
SHOP
ABOUT US
INVESTOR RELATIONS
FAQS
Selfcare
Blog
Our Impact
Dealer Login
Careers
Safaricom
VOICE
DATA
HOME FIBRE
M-PESA
Search website...
HELP
SWITCH VIEW
A-
A+
A
GET MORE
Register for M-PESA
M-PESA Rates
Using M-PESA
M-PESA Tips
Experience M-PESA
Chat with Zuri on WhatsApp
https://wa.me/254722000100?text=[KE] Hi Zuri

Machine translation



NLP Terminology

- **Morphology**-It is a study of construction of words
- **Semantics**-It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
- **Phonology**-It is study of organizing sound systematically.
- **Pragmatics**-It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.
- **Discourse**-How current sentence can affect meaning of the next one.
- **Stop words**-commonly used words in a language.
- **Document term matrix-** A tool used to show the frequency of words in a particular document/set of documents.

Cont....

- **Syntax**-It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases
- **World of knowledge**-It includes the general knowledge about the world.
- **Lexicon**-Collection of phrases and words in a language.
- **Tokenization**-Breaking down a sentence into chunks(tokens) for easier analysis
- **Stemming**-normalize a word by removing prefix or suffix to have it in its root/base form
- **Lemmitization**-Finding meaning of a word before stemming
- **Ambiguity**-State of uncertainty in meaning

Cont..

- **Synonyms**: are words that are written differently but are similar in meaning. For example: Clever and smart ,Begin and start, Beautiful and pretty
- **Antonyms:** are words that have meanings that are opposite to each other. For example: Clever and stupid, Begin and end, Beautiful and ugly, Sad and happy
- **Polysemy:** Words that have the same written form and a related meaning. For example: You must face your fear. Vs. Her face is beautiful.
- **Hyponymy:** A word is a hyponym of another word if it represents a subclass of the other word. For example : Orange is a hyponym of fruit and Yellow is a hyponym of color
- **Text normalization:** is the process of transforming text into a single form, which ensures consistency before operations are performed on it

Cont…

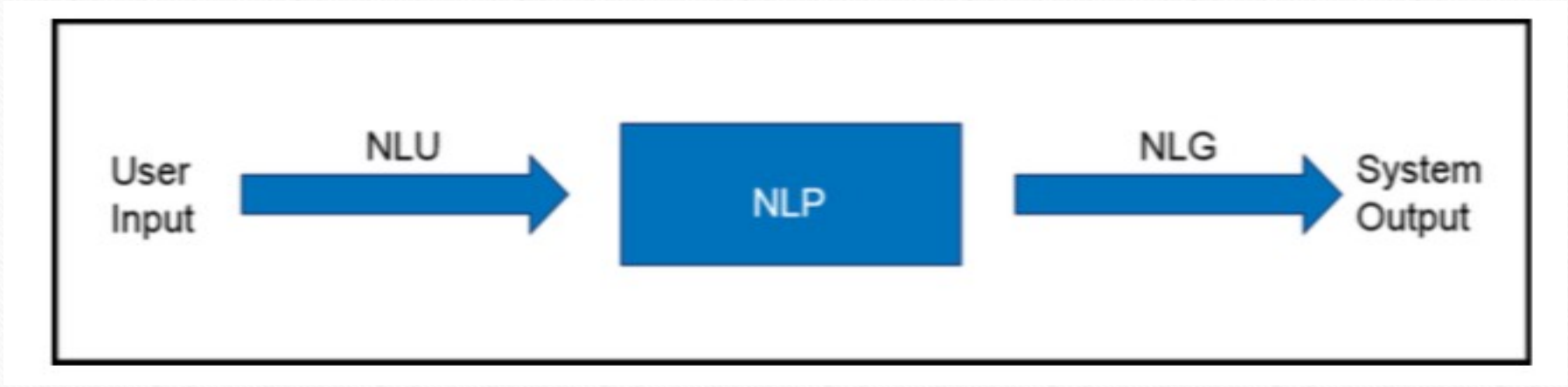
- Here are some examples of normalization:
- Examples:
  - Case folding: Child → child
  - Duplication removal: Hiiiiii → Hi
  - Acronyms processing: WHO → World Health Organization
  - Format normalization: Ksh100 → 100 Shillings
  - Value normalization: 2 July 1980 → DATE

Cont…

- **Homonyms:** Words that have the same written form but have unrelated meanings. There are two types of homonyms:
  - *Homographs:* Words that have the same written form. For example:
    - This answer is **right**.
    - The building is on the **right** side of the river.
    - You have the **right** to remain silent.
    - Come here **right** now.
  - *Homophones:* Words that sound similar when spoken but have different meanings and spellings. For example:
    - “right” and “write”.

Components of NLP

- Natural Language Understanding [NLU]
- Natural Language Generation [NLG]



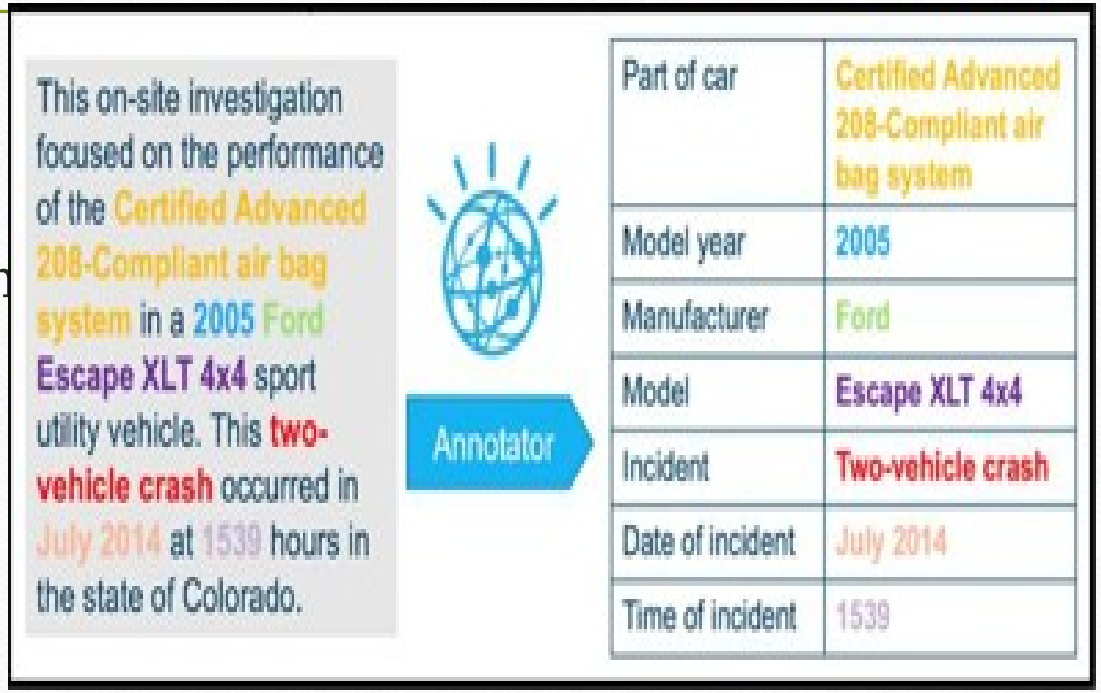
Components of NLP

1. Natural Language Understanding [NLU]- Involves mapping input in N.L into useful representation and analyzing different aspects of the language.

- When doing this , some Ambiguities may arise as follows:
  - Lexical Ambiguity e.g consider the word Board , fly, jam, current ,Mine, orange , mto
  - Syntax level Ambiguity e.g He lifted the boy with the red rope
  - Referential Ambiguity/Anaphoric Ambiguity e.g she said I am tired

Cont…

- **NLU applications**
  - Unstructured to structured
  - Question and answer systen
  - Sentiment analysis



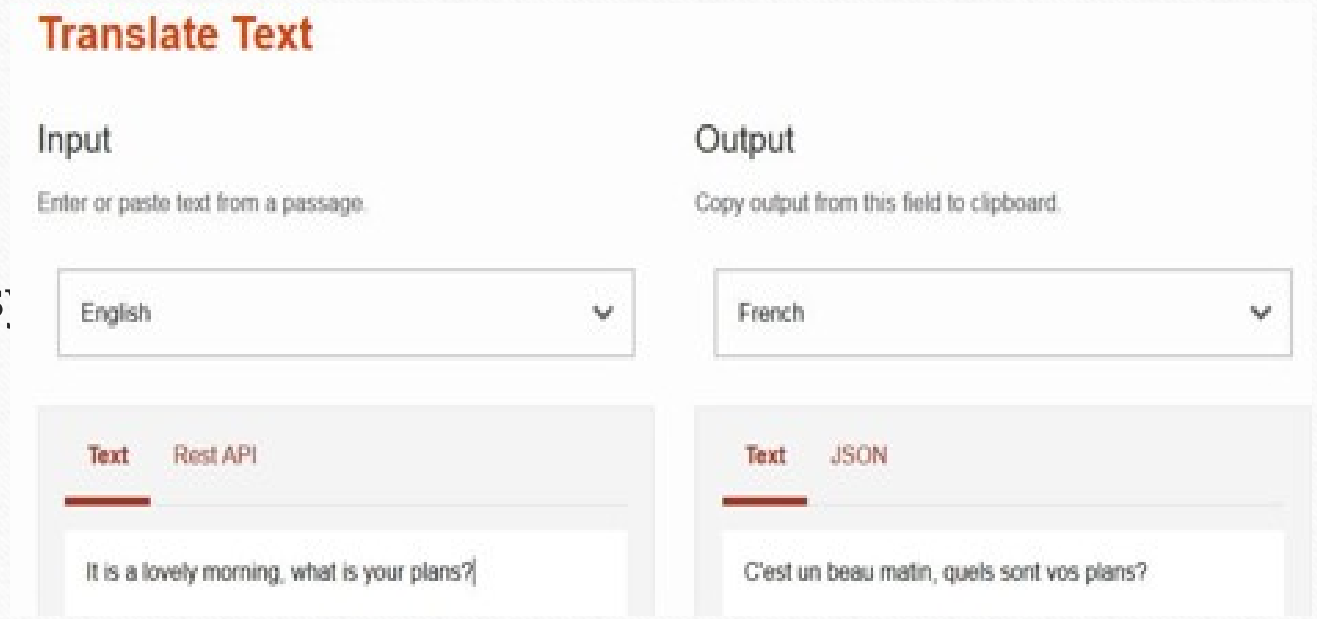
Cont…

2. Natural Language Generation [NLG]- Involves producing meaningful phrases and sentences in the form of natural language .It involves the following:

- Text planning[retreving content from a K.B]
- Sentence planning[choosing the required words to make a sentence]
- Text realization[mapping sentence plan into sentence structure]

Cont.....

- **NLG applications**
  - Machine translation
  - Text summarization
  - Weather forecasting s



Steps in NLP

- 1. **Lexical Analysis −** It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.
- 2. **Syntactical Analysis −** It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

- 3. **Semantic Analysis −** It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.
- 4. **Discourse Integration −** The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

- 5. **Pragmatic Analysis −** During this, what was said is re-interpreted on what it actually meant[interpratation of actual meaning]. It involves deriving those aspects of language which require real world knowledge.

NLP pipeline

- A **pipeline** is a way to design a program in which the output of one module feeds into the input of the next module.
- Using the NLP pipeline divides the tasks of NLU, which makes NLU less complicated.

Example ..

- We can use NLP pipeline to understand the following sentence: “Yes, I received your invitation, and I will happily attend your party.”

[This statement is understandable by a human but not a computer. We use the pipeline to walk though each stage to analyze the statement]

1. Sentence segmentation:[Detect sentence boundaries i.e. where sentence starts and Ends.]

There are two sentences in the example that are separated by the conjunction “and”:

“Yes, I received your invitation, and I will happily attend your party.”

Apply this process to the example:

“Yes, I received your invitation.”

“I will happily attend your party.”

2.Tokenization:[Breaks a sentence into tokens]

- Tokenization uses delimiters, for example, a space ” “.Apply tokenization to the example:
  - First sentence: “Yes””,” “I” “received” “your” “invitation”
  - Second sentence: “I” “will” “happily” “attend” “your” “party”

3.Parts of speech (POS) tagging:[The process of tagging each token with grammatical representation, such as noun, verb, or adjective).]

POS helps the computer to
understand language and
grammar and derive meaning from the input sentence.

Example of POS tagging

Tag	Description
JJ	Adjective
NN	Noun Single
NNS	Noun Plural
RB	Adverb
RP	Particle
VB	Verb, base form
VBD	Verb, past tense

Apply POS tags to both sentences.

- First sentence: “yes/UH” “,/,” ”I/PRP” “got/VBD” “your/PRP$” “invitation/NN” “,/,”
- Second sentence: “I/PRP” “will/MD” “happily/RB” “attend/VB” “your/PRP$” “party/NN”

Tools used in NLP

- **NLTK**-Natural Language Tool Kit :A Python library that provides modules for processing text, classifying, tokenizing, stemming, tagging, parsing, and more.
- **Stanford Core NLP**:A suite of NLP tools that provide part-of-speech tagging, a named entity recognizer, a co-reference resolution system, sentiment analysis, and more.
- **WordNet :**One of the most popular lexical databases for the English language. Supported by various API and programming languages
- **Apache Open NLP** : Provides tokenizers, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, co-reference resolution, and more.

Services used in NLP

- Instead of using low-level libraries, you can use many cloud services that accomplish high-level NLP tasks, for example, IBM Cloud, Microsoft Cloud (Azure), and Google Cloud.
- IBM offers its AI services through IBM Cloud. The NLP services that are provided include the following ones (among others):
  - Watson Natural Language Classifier for text classification
  - Watson Natural Language Understanding for entity identification and relation extraction

Challenges to NLP

- There are always some challenges that need to be tackled for any case. In NLP, here are the most popular challenges:
  - Domains: Higher accuracy for specific domains compared to generic domains.
  - Language: English gets the most attention because it is an international language.
  - Medium: Processing speech is more difficult than processing text.