Welcome to:

Classical Approaches of NLP

Unit objectives

After completing this unit, you should be able to:

- Understand various concepts of Natural Language Processing Activities
- Gain knowledge on Text Processing methods
- Learn about Lexical Analysis and its methodologies
- Gain an insight into perform Syntactic Parsing
- Understand concepts of Semantic Analysis and its methods
- Gain knowledge on Natural Language Generation

Introduction (1 of 2)

- Computational linguistics:
  - Computational linguistics is an interdisciplinary field.
  - Concerned with the statistical or rule-based modeling.
  - Theoretical part is more relevant to the basic knowledge and foundation of linguistics.
  - Understanding the grammar of the languages and the morphology.

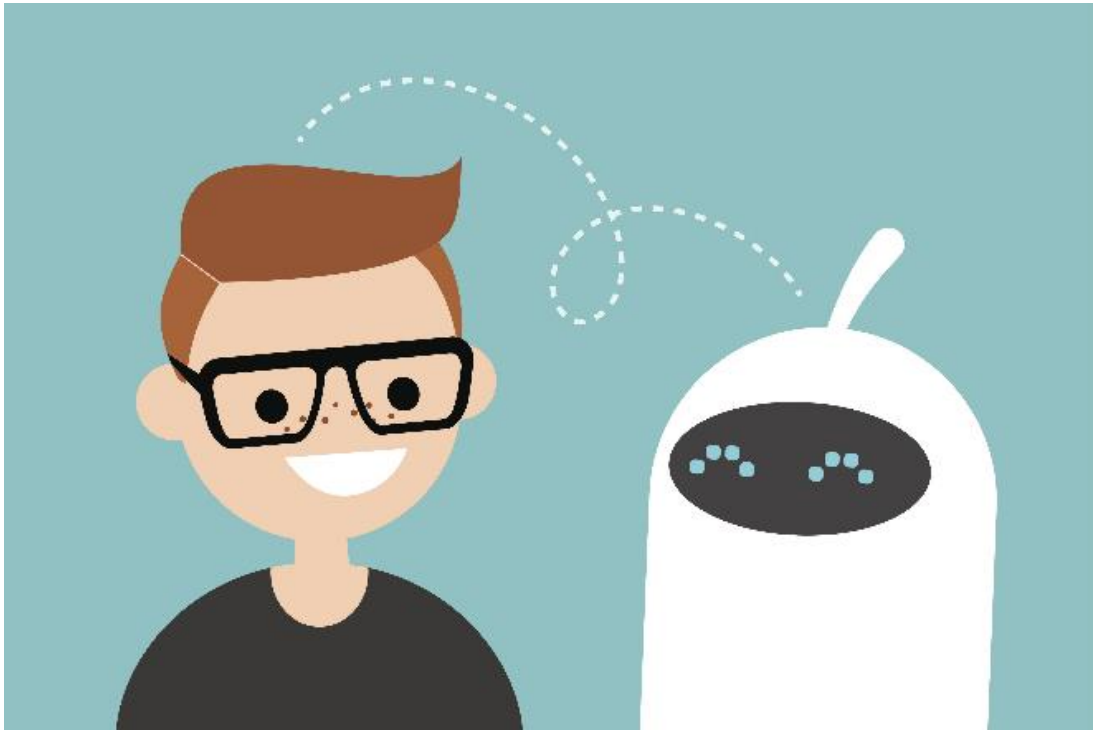


Figure: Computational Linguistics

Source: https://chatbotsmagazine.com/how-computational-linguists-help-your-chatbot-understand-humans-8ec95ab903f6

Introduction (2 of 2)

- Natural language processing:
  - Natural Language Processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, how to program computers to process and analyze large amounts of natural language data.
  - Sixth Sense activities.
  - Speech recognition.
  - Language understanding.
  - Language generation.
  - Techniques.
  - Tagging.
  - Hidden Markov models.
  - Decision trees.
  - Probabilistic values.
  - Language models.
  - Speech recognition.

Classical approaches to natural language processing

- Goals of computational linguistics:
  - Formulation of grammatical framework which can check the semantics of the language.
  - Formulated by implementing syntactic and semantic analysis.

- Growth of computational linguistics:
  - System that can comprehend, recognize speech, tagging and parsing activities.
  - Neural network approach for the activity.
  - Good in the processing power.
  - Process information faster.
  - Both syntactic and semantic processing methodologies.
  - Extraction of clusters from any language, identifying the relational tuples, paraphrase sets are all important aspects of text corpora.
  - Computational linguistics and natural language processing algorithms are implemented in machine learning.

Approaches to natural language processing (1 of 2)

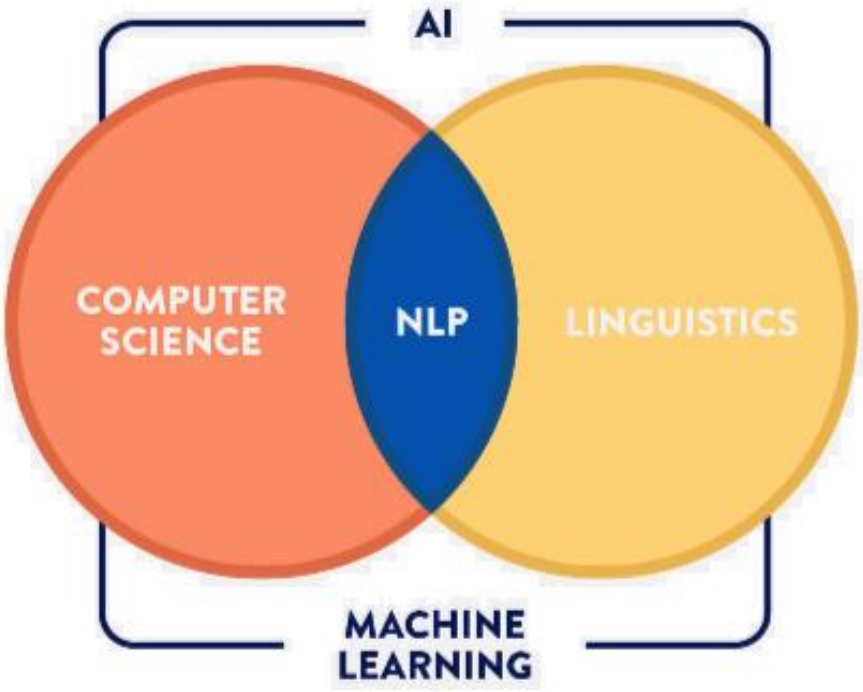
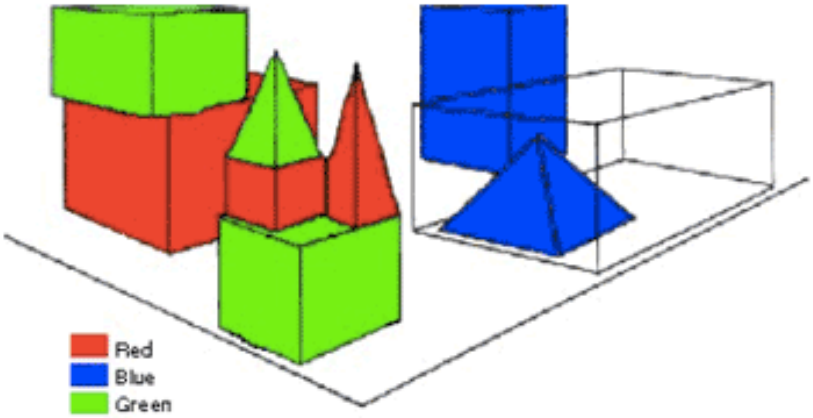


Figure: Approaches to natural language processing

Source: https://towardsdatascience.com/introduction-to-natural-language-processing-nlp-323cc007df3d



Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Figure: Approaches to natural language processing

Source: https://www.topbots.com/4-different-approaches-natural-language-processing-understanding/

Approaches to natural language processing (2 of 2)

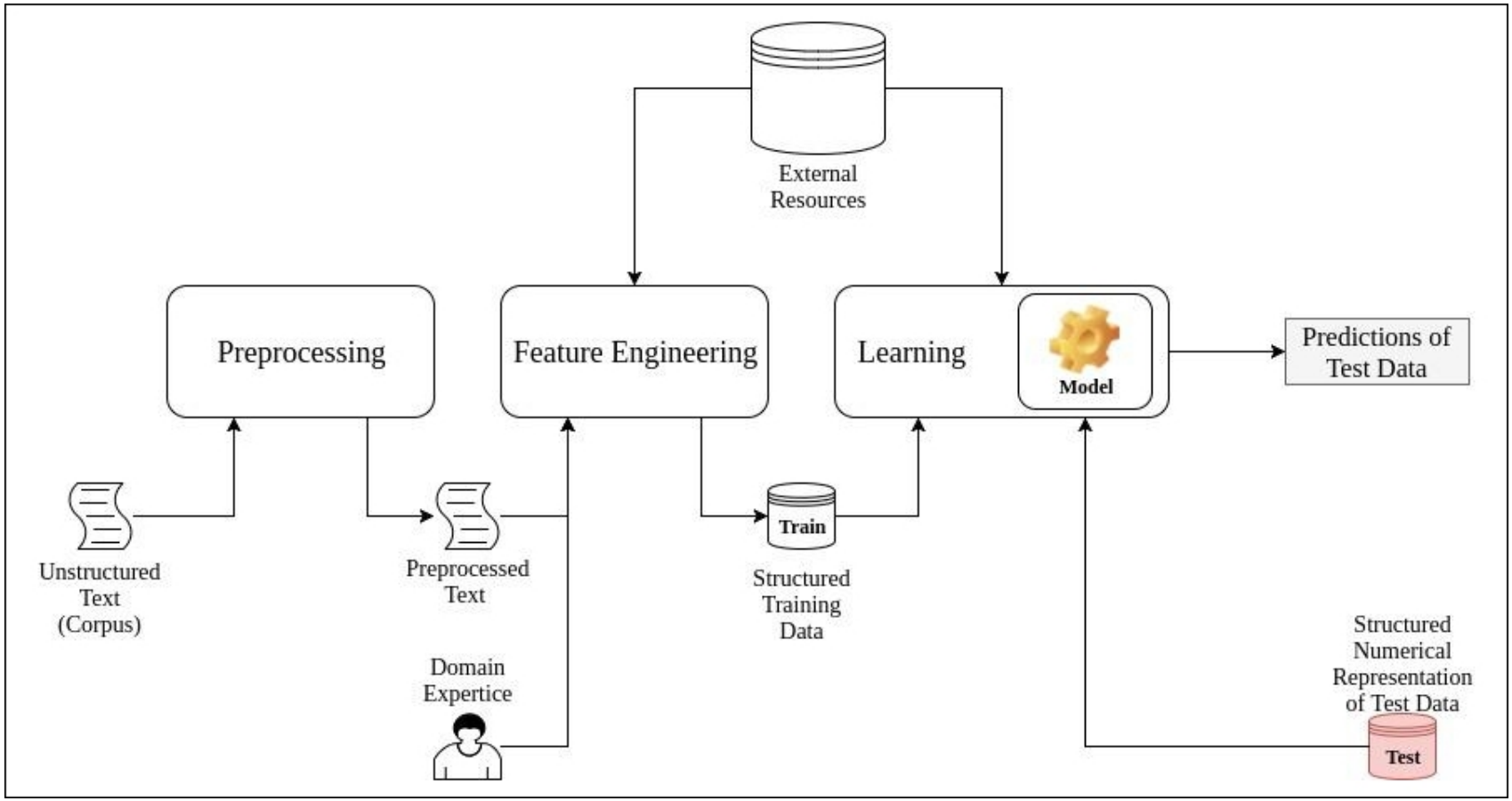


Figure: Shows the Classical approach to Natural Language Processing

Source: https://subscription.packtpub.com/book/application_development/9781788478311/1/ch01lvl1sec12/the-traditional-approach-to-natural-language-processing

Understanding linguistics

- Basic ideology starts by making the computer identify the different stages of learning a natural language.

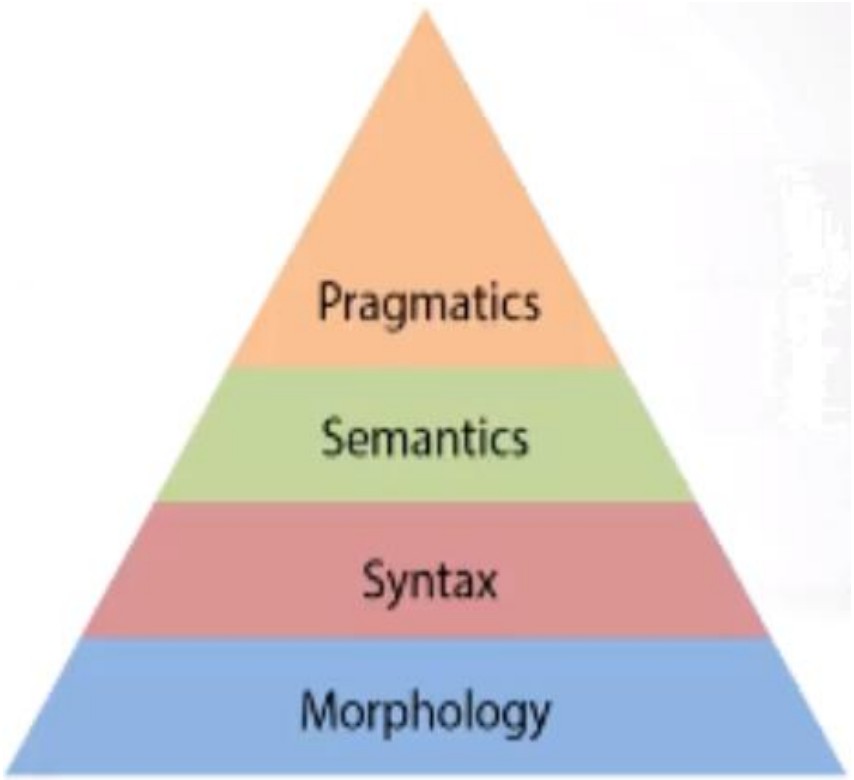


Figure: Major levels in linguistics analysis

Source: https://towardsdatascience.com/linguistic-knowledge-in-natural-language-processing-332630f43ce1

Level 1: Morphology

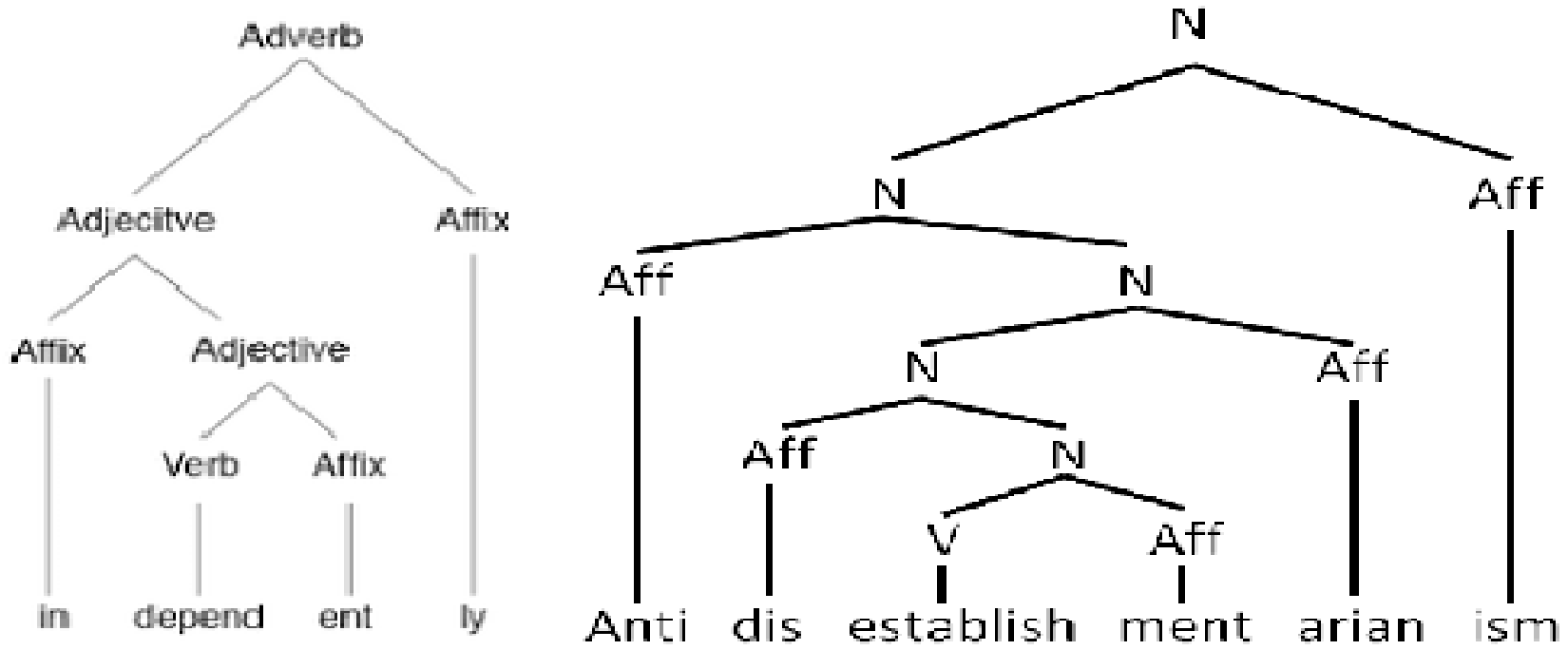


Figure: Morphology makes up the basic constructs

Source: https://en.wikipedia.org/wiki/Morphology_(linguistics)

Level 2: Syntax

- Syntax follows the grammar of the language.

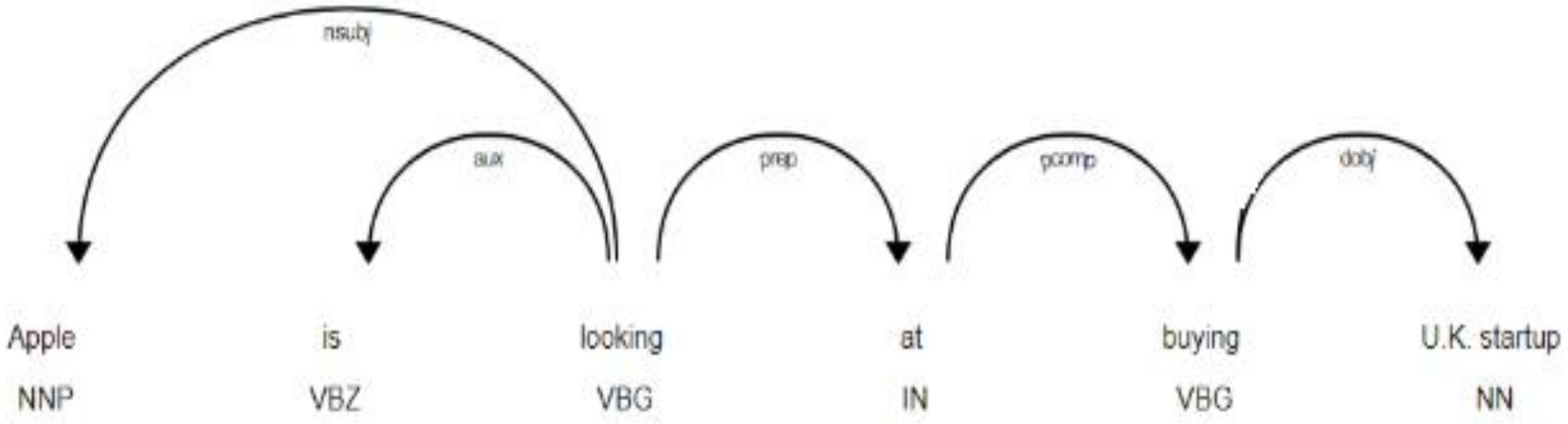


Figure: Syntactic analysis is done at the statement level

Source: https://all-about-linguistics.group.shef.ac.uk/branches-of-linguistics/morphology/what-is-morphology/

Level 3: Semantics

- Deals with the meaning conveyed by creating sentences in that language.
- Semantics include tasks like named entity recognition and relationship extraction.

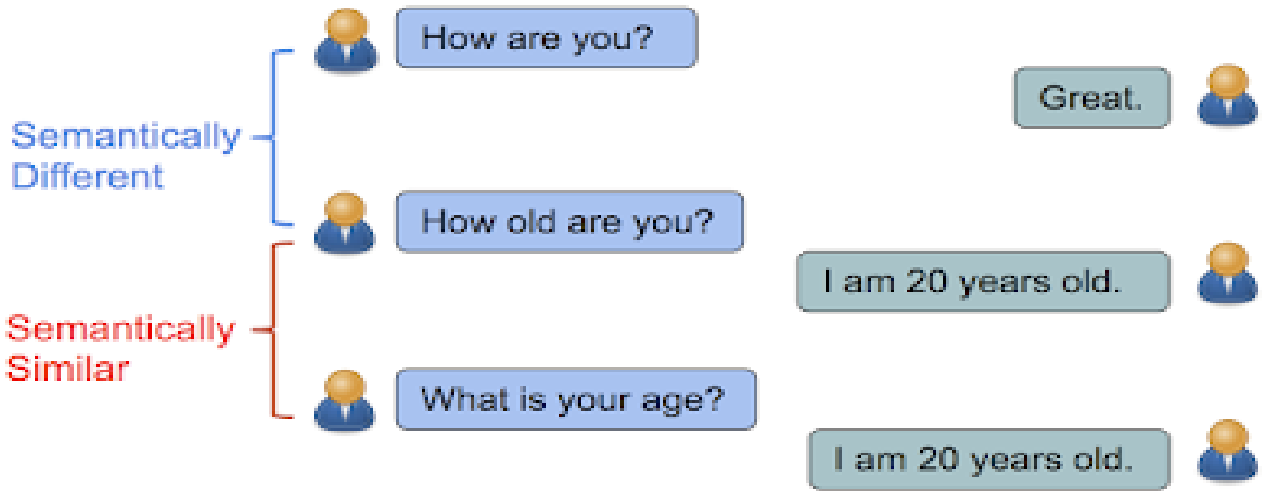


Figure: Semantics

Source: https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html

Level 4: Pragmatics

- Understanding the sentences created and to understand the conveyed meanings.

- Common problems that are associated with pragmatics:
  - Co-Referencing.
  - Summarization.
  - Modelling.
  - Question and answering.

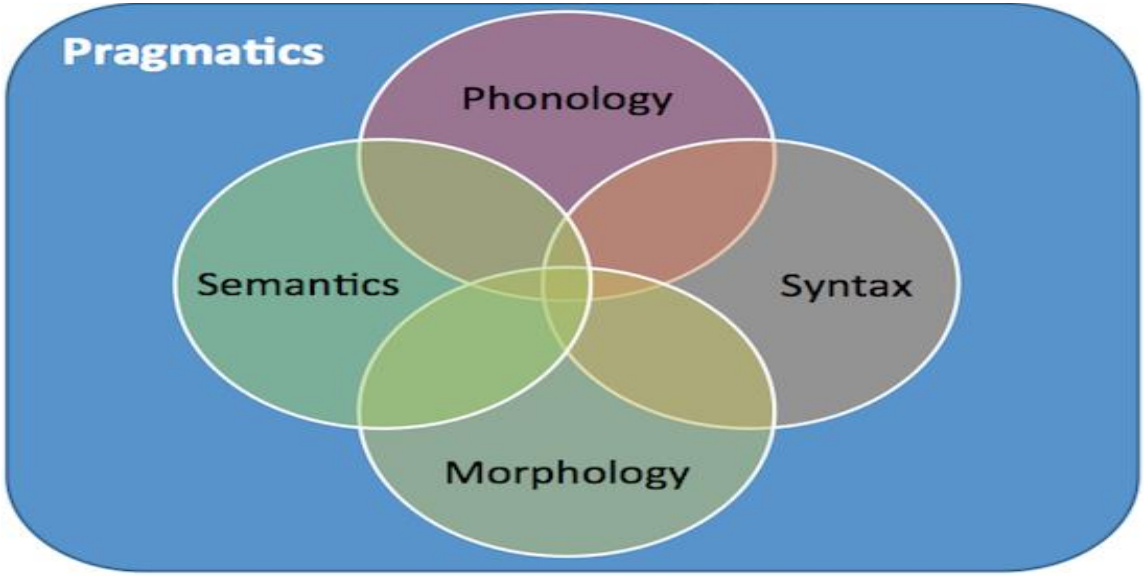


Figure: Pragmatics

Source: https://medium.com/@paulomalvar/pragmatics-the-last-frontier-9d64351eea6f

Understanding linguistics

- The syntax and semantics go hand in hand and framing a sentence:
  - Hyponymy is used to convey the relationship between a general term and a specific instance. (Crocodile is an amphibian).
  - Meronymy used to convey that one part of a sentence is a part of another(fish has gills).

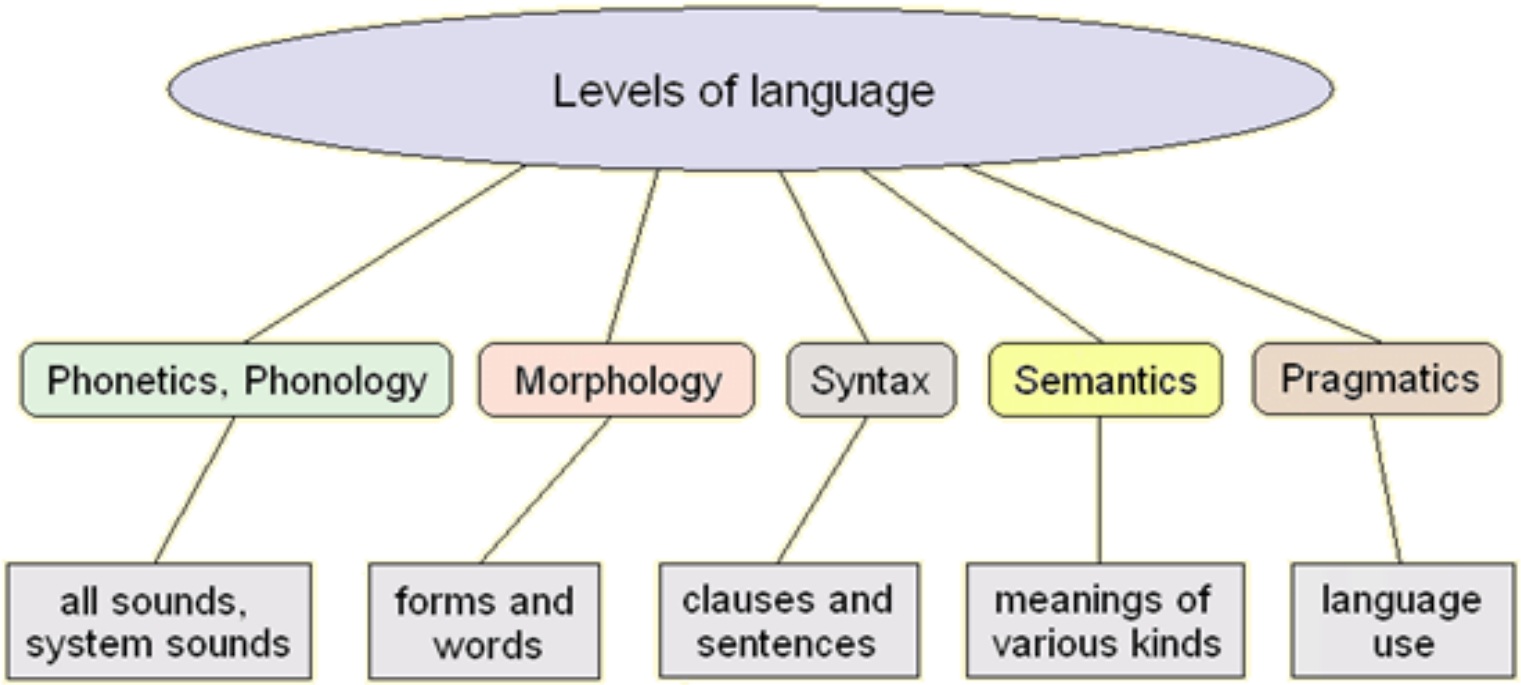


Figure: Linguistics

Source: https://www.uni-due.de/SHE/REV_Levels_Chart.htm

Traditional approach (1 of 2)

- Processing is considered as a sequence of steps.
- Separate and distinct processes take place.
- Preprocessing: Removal of unwanted data.
- Feature engineering: Understanding the numeral representation of the textual data.
- Machine learning algorithms: Learning the language using the training data.
- Predicting outputs: Identify the prediction with test data.

Traditional approach (2 of 2)

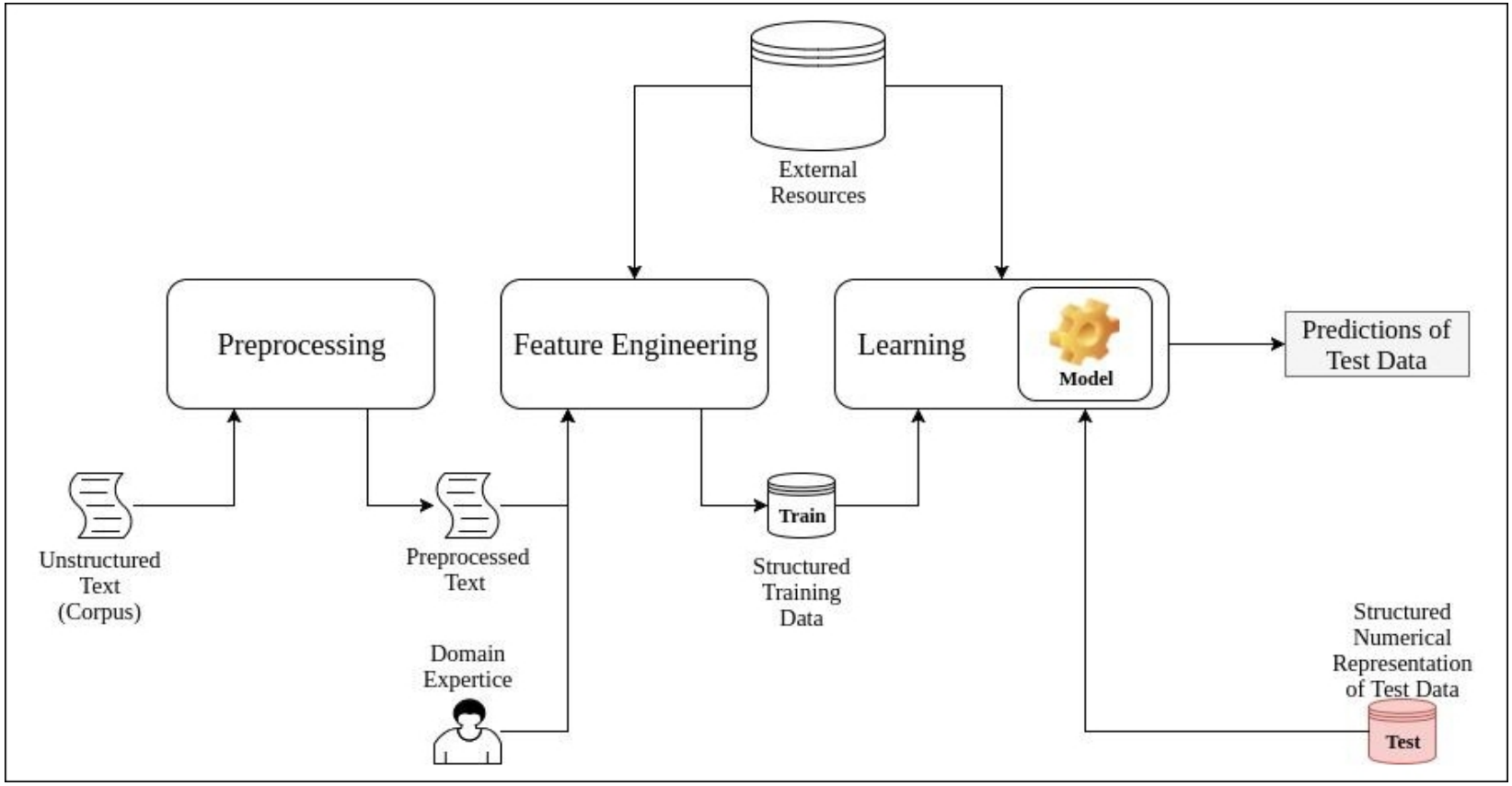


Figure: Traditional approach

Source: https://subscription.packtpub.com/book/application_development/9781788478311/1/ch01lvl1sec12/the-traditional-approach-to-natural-language-processing

Example: Automatic summarization using NLP

- During a game let the NLP activity be automatic generation of the summary of the game. Will have multiple sets of statistics like scoring, penalties etc.

- The data collected contains the most relevant sentence to create the summary for every statistical parameter. Natural language processing algorithm should now create a summary of the game.

- Pre-processing steps:
  - Stemming: Choosing root verbs.
  - Removing distractions: Eliminating the punctuation.
  - Tokenization: Identifying simple words.

Drawbacks

- Loss of information.
- Lengthy and tedious process.
- Domain expertise.
- External resources.
- Identification of external resource.

Text processing

- Theory and practice of automating the creation or manipulation of electronic text.
- Text: Alphanumeric characters specified on the keyboard.
- Processing: Automated or mechanized processing.
- Representation of data:
  - Text.
  - Images.
  - Audio.
  - Videos.
- Analyzing the data which may be structured or unstructured to obtain structured information.

What Is text processing?

- The textual information: Processed, analyzed and manipulated-machines learn.
- Text extraction and text classification.
- Extracting individual and small bits of information from large text data is called as text extraction.
- Assigning values to the text data depending upon the content is called as text classification.

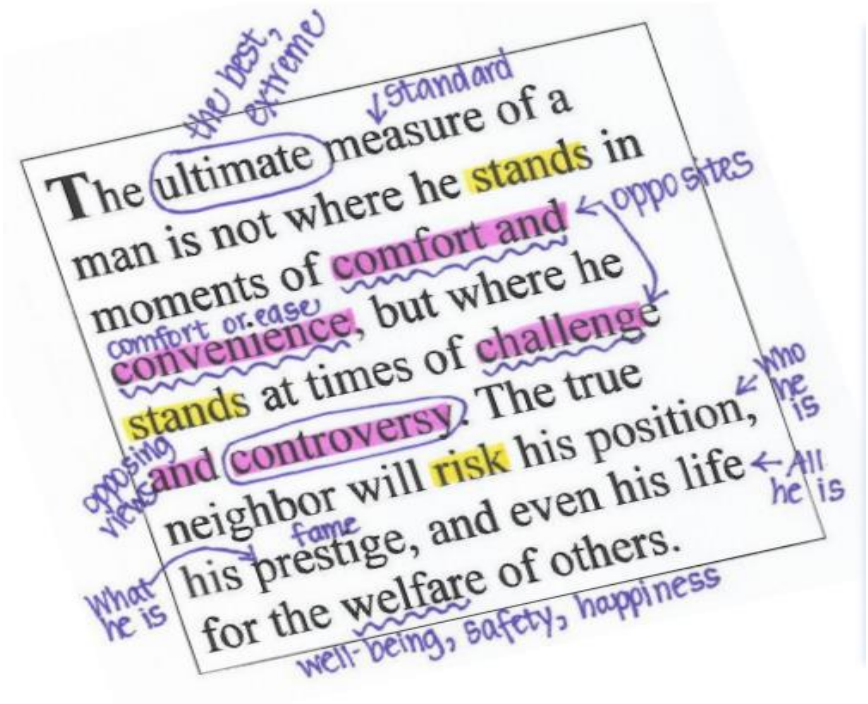


Figure: Text processing

Source: https://www.wcpss.net/Page/8911

Text analysis vs. Text mining vs. Text analytics

- Used to obtain data by statistical pattern learning.
- Both text analysis and text mining are qualitative processes.
- Text Analytics is quantitative process.

- Example:
  - Banking service: Customer satisfaction.
  - Text analysis: Individual performance of the customer support executive. Text used in the feedback like "good", "bad".
  - Text analytics:
    - Overall performance of all the support executives.
    - Graph for visualizing the performance of the entire support team.
  - Text analytics for overall count of issues resolved.

Tools and methodologies: Statistical methods

- Statistical methods:
  - Word frequency: Identify the most regularly used expressions or words that is present in a specific text.
  - Collocation: Method for identifying the common words that appear together.
  - Concordance: Methodology to provide context to the natural language.
  - TF-IDF: Identifies the importance of words in a document.

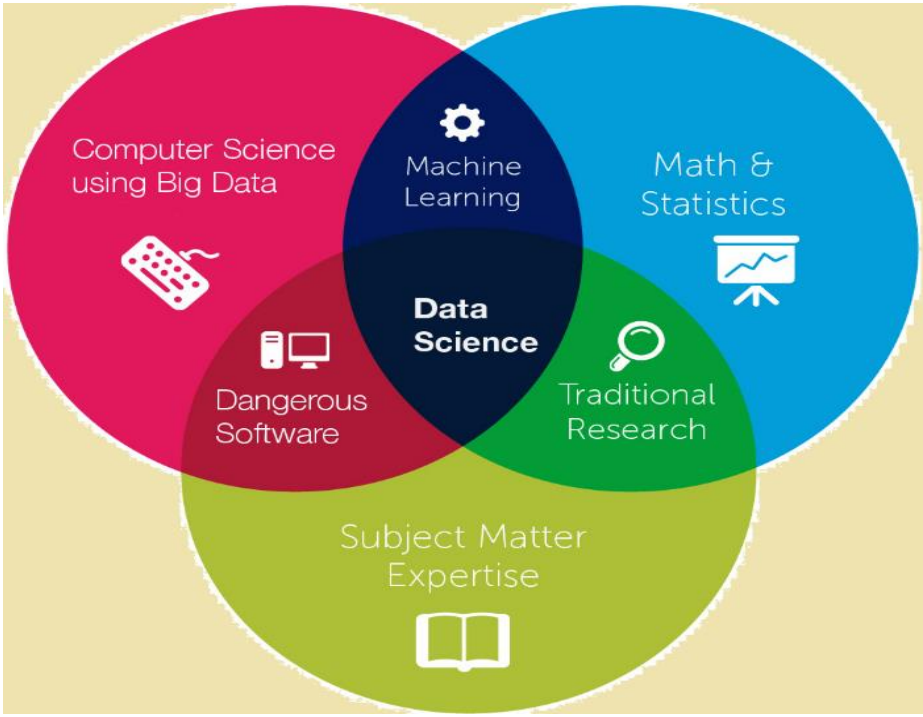


Figure: Statistical methods

Source: http://grjenkin.com/articles/category/data-science/106322/big-data-data-science-and-machine-learning-explained

Tools and methodologies: Text classification (1 of 2)

- Text classification:
  - Content is analyzed and classified into multiple predefined groups based upon the analysis.

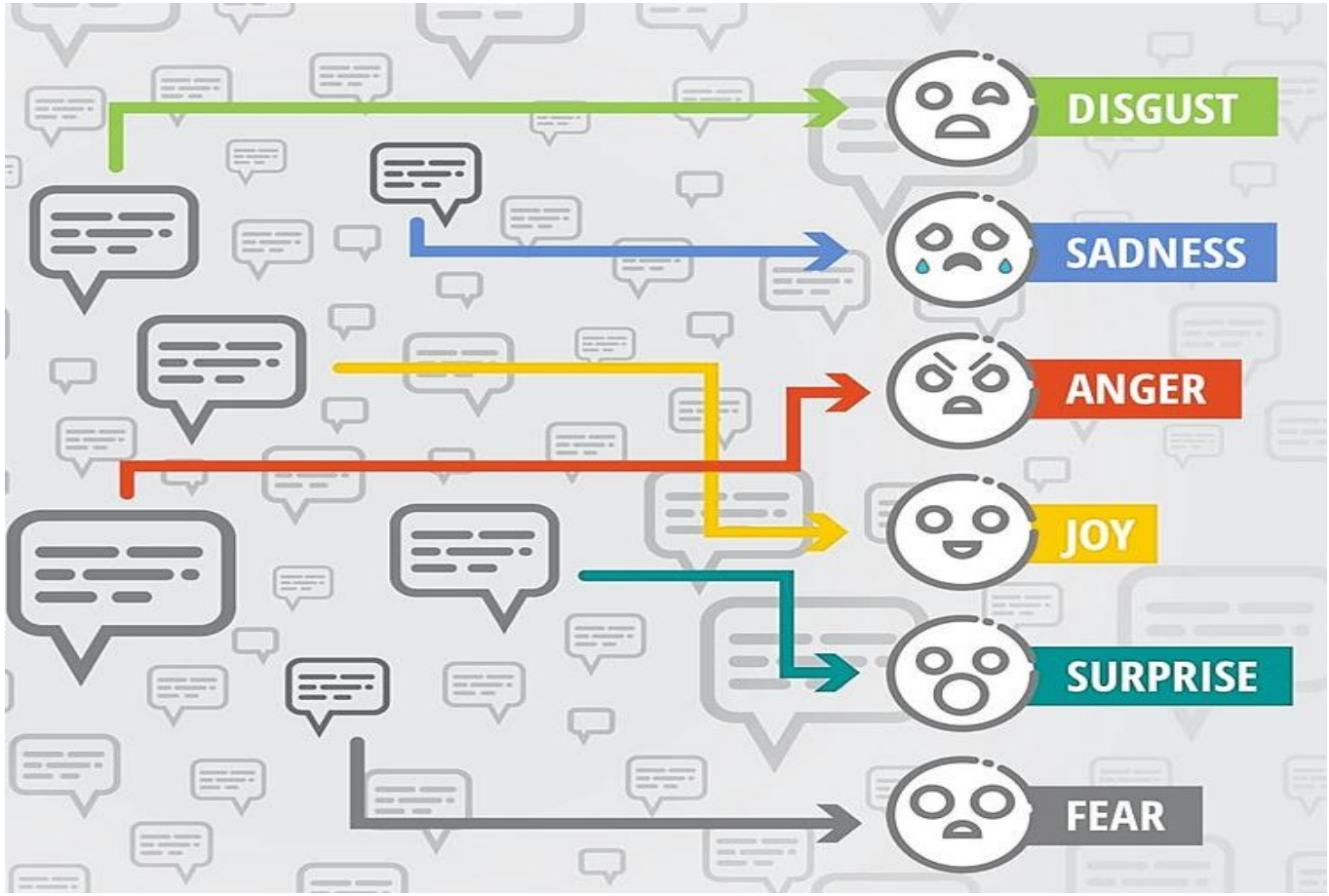


Figure: Text classification

Source: https://hackernoon.com/text-classification-simplified-with-facebooks-fasttext-b9d3022ac9cb

Tools and methodologies: Text classification (2 of 2)

- Topic analysis: Identify and interpret large collection of text according to the individual topics assigned.
- Sentiment analysis: Understanding the emotional feel represent in a textual message.

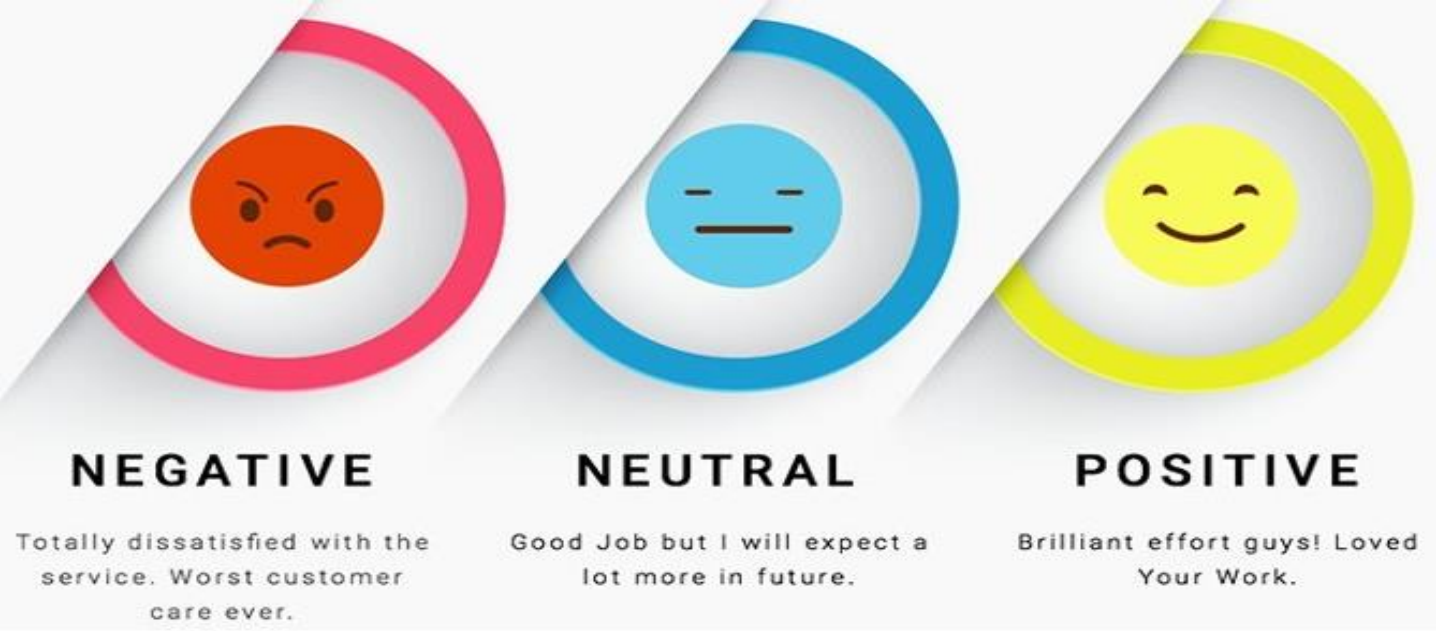


Figure: Language classification

Source: https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html

Tools and methodologies: Text extraction

- Text extraction: Process of gathering valuable pieces of information present within the text data.

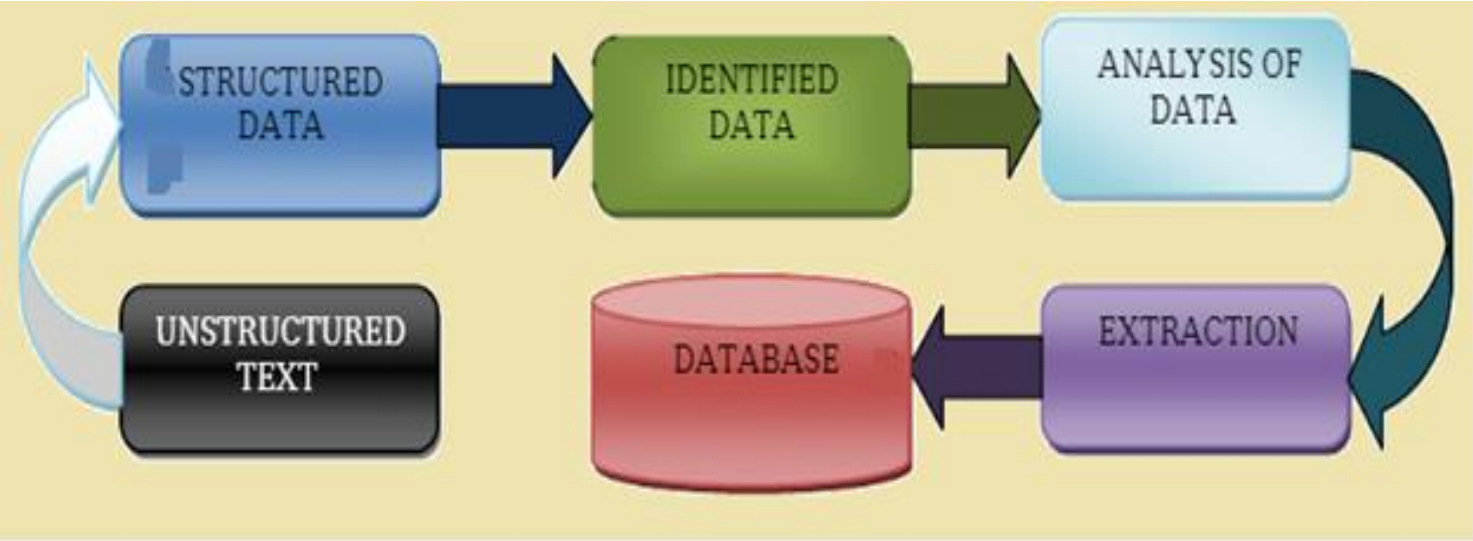


Figure: Text extraction

Source: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/

- Keyword extraction: Identifying and detecting the most relevant of the words inside a text.

- Entity extraction: Useful for gathering information on specific relevant elements and to discard all the other irrelevant elements.

Tools and methodologies: Example

- Search engines use clustering.
- Web pages: Set of similar words are grouped and clustered.
- Pulling up the pages: High count of words relevant.

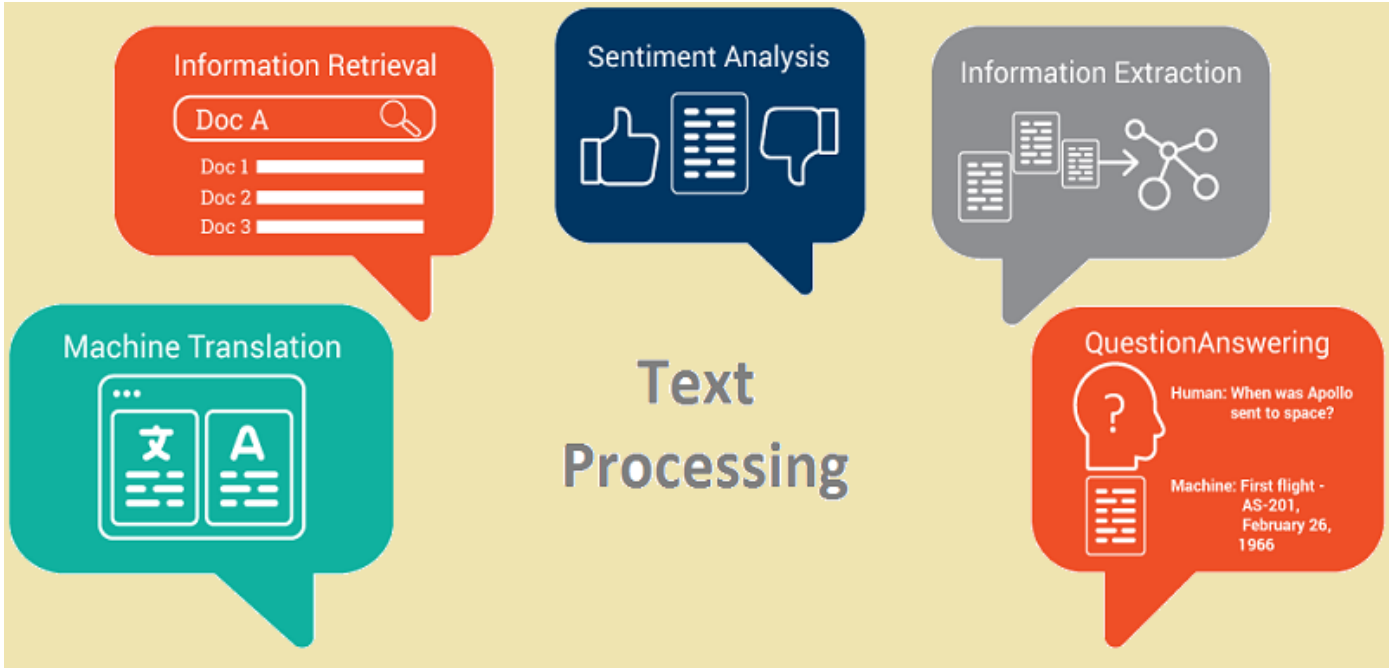


Figure: Test processing

Source: https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

Scope of text analysis/processing

- Large documents:
  - Refer for a context.
  - Cross examine multiple documents.
- Individual sentences:
  - Gathering specific information.
  - Identify the emotional or intentional activities.
- Parts of the sentences:
  - Sentiments of the words can be analyzed.
  - Better understanding of the natural language.
  - Provided for machine to analyze and understand.

Importance of text analysis

- Business growth:
  - Extraction of information to identify the customer.
- Real time analysis:
  - Urgent requirements or complaint handled on a real-time basis.
  - Categorized as priority.
  - Require multiple analysis.
- Checking for consistency:
  - Detect latest models.
  - Analyzing.
  - Understanding.
  - Sharing of the available data accurately.

Working principles of text analysis

- Data gathering.
- Data preparation.
- Data analysis.

Data gathering (1 of 2)

- Text analysis: Gathering the required data that need to be analyzed.

- Internal data:
  - Email.
  - Chat messages.
  - CRM tools.
  - Databases.
  - Surveys.
  - Spreadsheets.
  - Product analysis report.

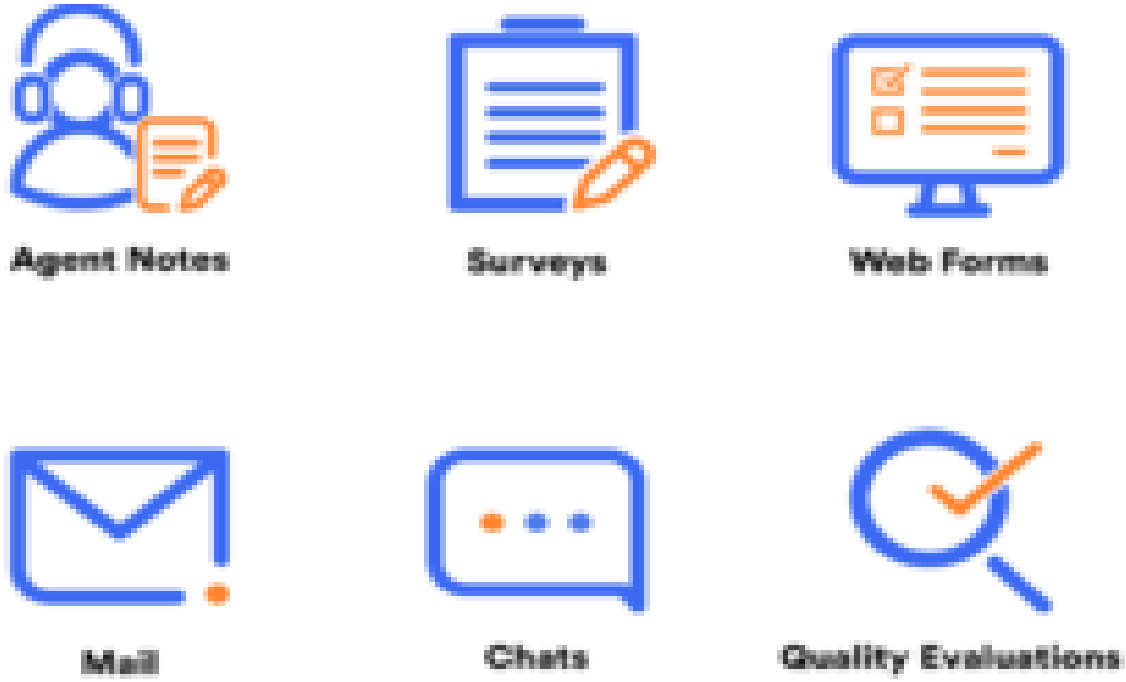


Figure: Text analysis

Source: https://voziq.com/customer-retention/improving-customer-retention-strategies-with-unstructured-customer-data/attachment/common-sources-of-unstructured-data/

Data gathering (2 of 2)

- External data: The external data do not belong to the organization and are available freely through other sources.

- Web scraping tools.
- Open data.

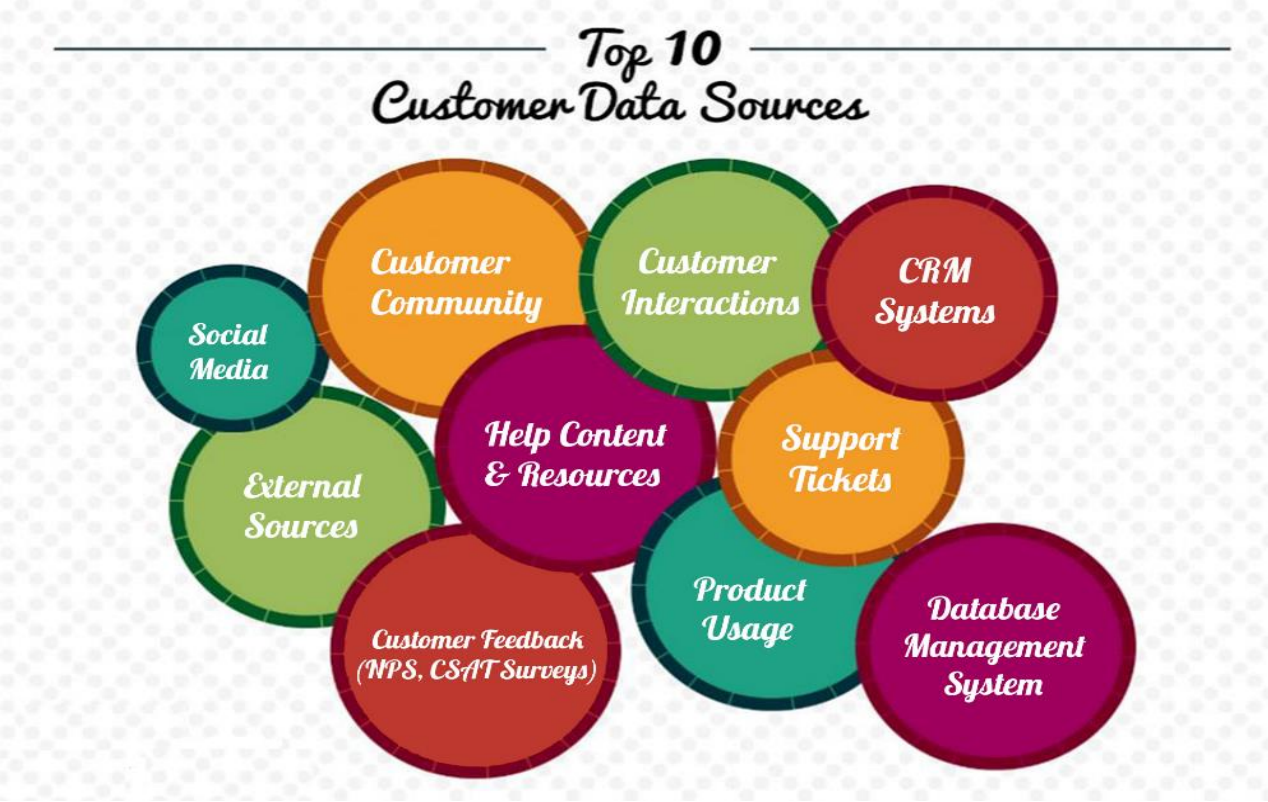


Figure: Web scraping tools

Source: https://strikedeck.com/top-10-customer-data-sources/

Data preparation (1 of 2)

- Before text is analyzed by any machine learning algorithm, it needs to be prepared.

- Tokenization:
  - Identify and recognize the unit of text.
  - Process of breaking up text characters into meaningful elements.
  - Analyze the meaningful parts of the text and discarding the meaningless sections.
  - Removes all the frequent words that can be found in a sentence.

- Stemming:
  - Used to reduces a word to its root to convey meaning.
  - Unnecessary character removal like prefix, suffix etc.

- Lemmatization:
  - Identify parts of the speech not needed and removes the inflection.

Data preparation (2 of 2)

- Constituency parsing:
  - Uses syntactic structures: Abstract notes associated to words and abstract categories.

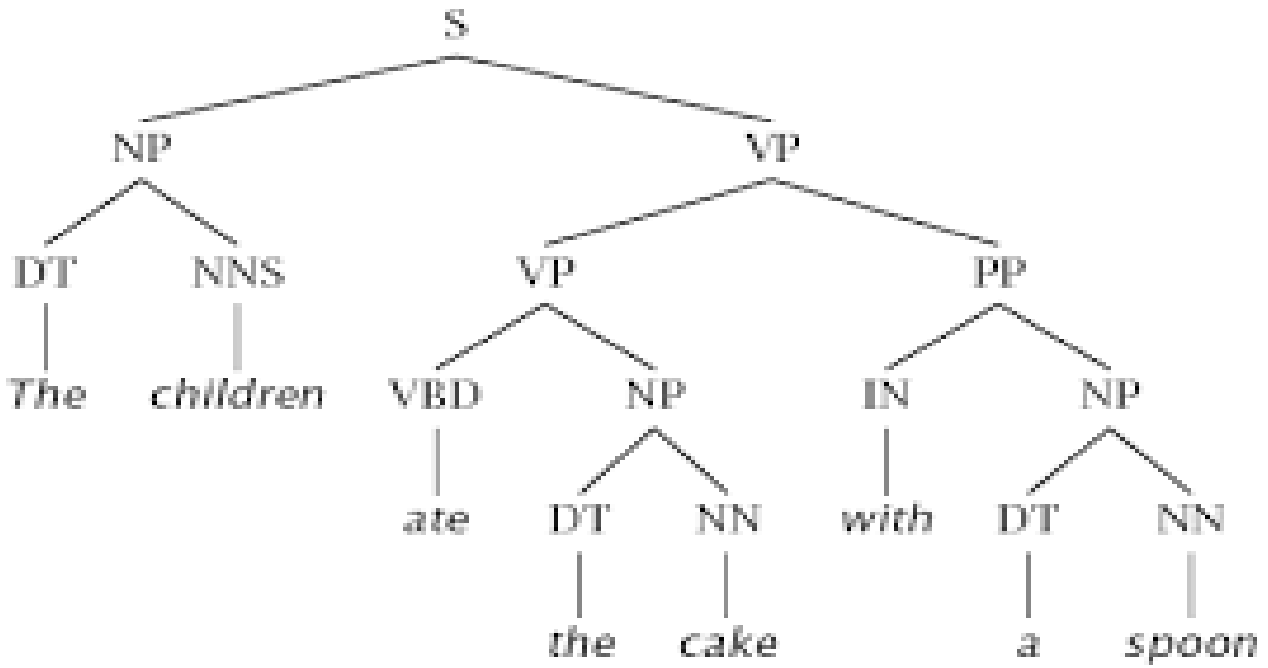


Figure: Constituency parsing

Source: http://www.cs.cornell.edu/courses/cs5740/2017sp/lectures/13-parsing-const.pdf

Data preparation steps

- Step 1: Recognize the Tokens:
  - "The sky is blue"
  - Token 1:["The S","ky is", " blue"] Token 2:["The", "sky", "is", "blue"]

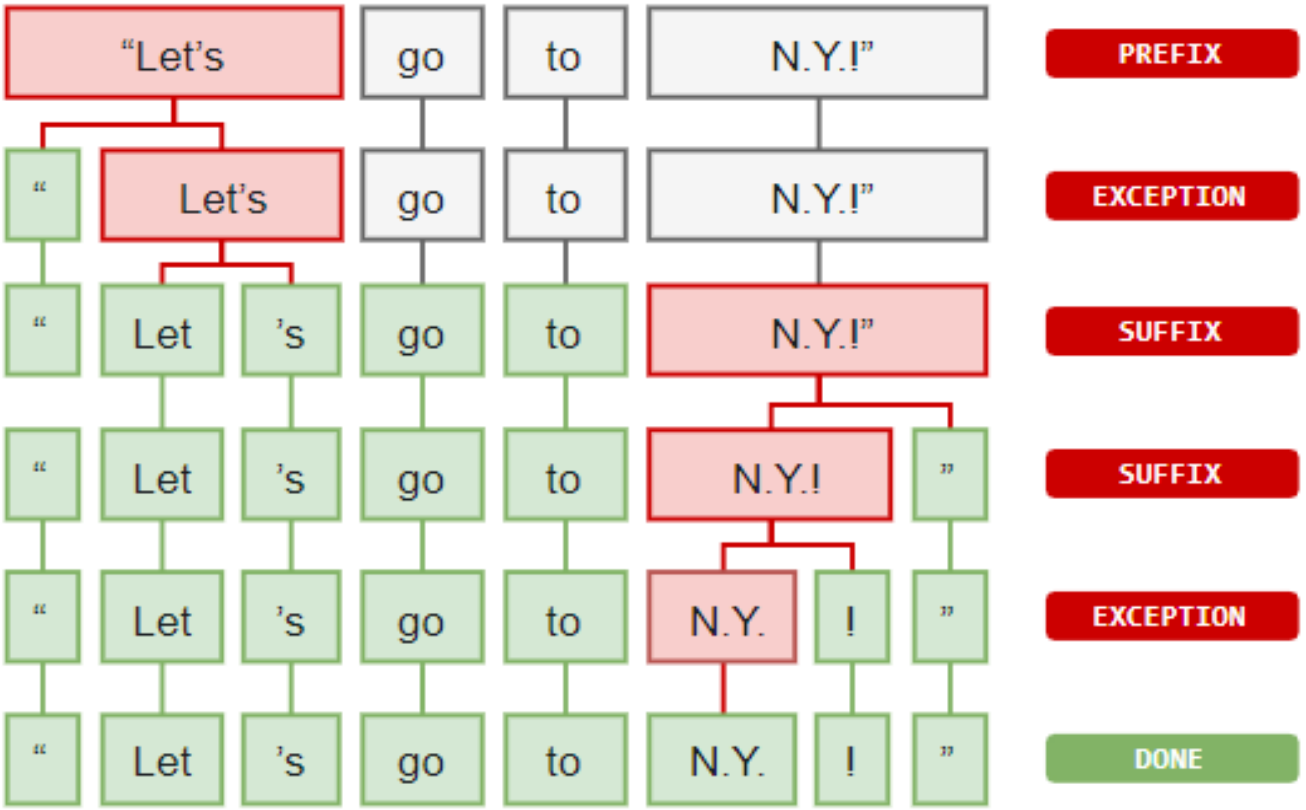


Figure: Recognize the tokens

Source: https://medium.com/@makcedward/nlp-pipeline-word-tokenization-part-1-4b2b547e6a3

Data analysis (1 of 2)

- Deals with the unstructured text that has been source from multiple locations.
- Two major processes: Text classification and Text extraction.

- Text classification:
  - Tags are assigned to the text based upon the content.

- Rule-based systems.
- Detect linguistic patterns in the text.
- Tag based upon the detection rules.

- Example:

(HDD | RAM | SSD | Memory): Hardware

- The classification algorithm will take any text that contains the word HDD, RAM, SSD, Memory and classify it under the common tag hardware.

Data analysis (2 of 2)

- Machine learning based systems.
- Predict based upon the past observations.
- Require multiple samples of text and tags.
- Converted into vectors before training.
- Vectors: Extract the features that are relevant.

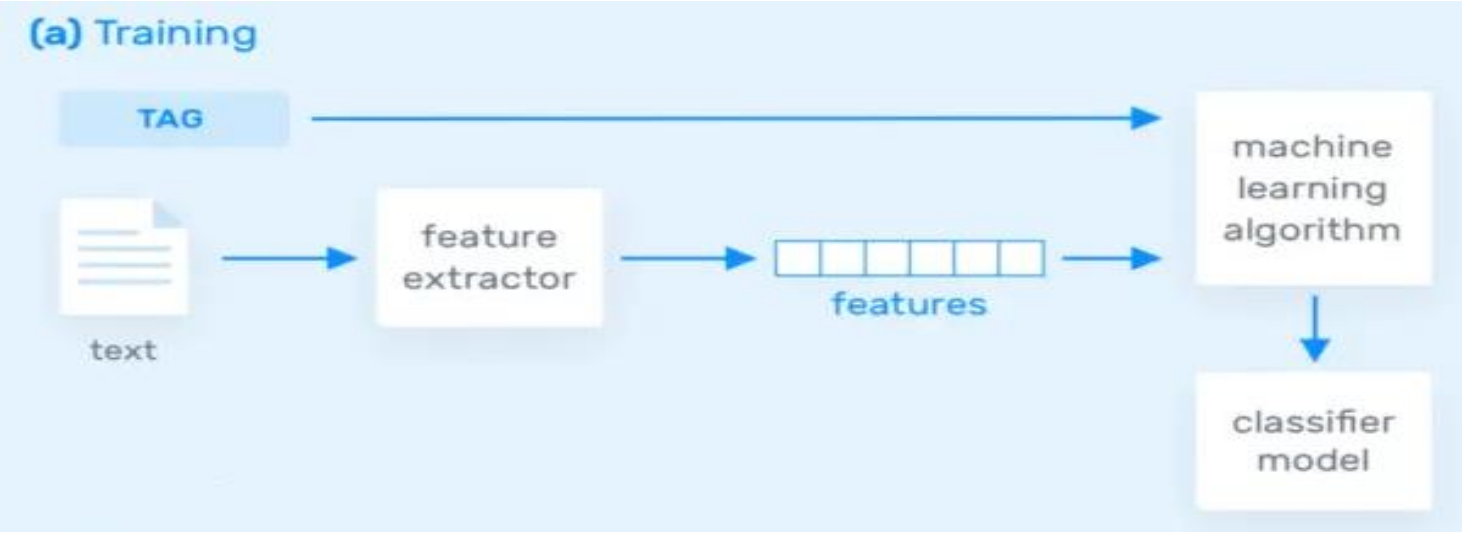


Figure: Training

Source: https://monkeylearn.com/text-analysis/

Evaluation of text classification process

- Sample testing data: Separate test data set.
- Cross validation method: Splits training and testing data.
- Accuracy: Number of correct predictions against the total number of predictions.
- Precision: Total number of correctly predicted output against the total number of outputs predicted.

Evaluation metric	Formula
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
Accuracy	(TP + TN)/N
F Measure	2*Recall*Precision/(Recall + Precision)

Figure: Evaluation metric formula

Source: https://www.researchgate.net/figure/Text-Classification-Evaluation-Metrics_tbl1_329910331

Text extraction

- Technique that can identify and gather valuable information from the data available inside any text.

- Keyword extraction:
  - Extracting the most relevant words or expressions that are available inside any text.
  - Identify and detect multiple words or expressions.
  - Extract content about any keyword.

- Entity extraction:
  - Identify the major entities that are relevant to any context.

Analysis in test extraction

- Regular expressions:
  - Identify any pattern of characters that can we search inside a large text data.

- Very fast.
- Results: Straight forward and almost accurate.
  - Complexity of the pattern leads to difficulty in coding.

```
(?i)\b(?:[a-zA-Z0-9_-.]+)@(?:(?:[[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.)|(?:(?:[a-zA-Z0-9-]+.)+))(?:[a-zA-Z]{2,4}|[0-9]{1,3})(?:]?)\b
```

- Conditional random fields.
- Algorithm is trained to learn patterns that are to be extracted from the given source.
- Complex patterns can be used for encoding higher dimensional information.
- Machine learning algorithm can follow on any source data.

Evaluation of text extraction process

- Metrics:
  - Accuracy.
  - Precision.
  - Recall.
  - F1 Score.

- Must be a perfect match of the extracted segment.
- Should be able to capture partial results also.

- Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric methodology: Identifying the performance of text extractors.

- Calculate the length and sequence numbers that overlap among the source text and the extracted text.

Text analysis APIs

- Python:
  - Large collection of libraries: Support numerical, scientific models is common for usage in identification and extraction process of text.
- Natural language toolkit:
  - Can be used for analysis of text.
  - Predefined methods and library functions: Lot of operations needed in text analysis.
- Scikit learn:
  - This library works with other modules like Numpy, Scipy and MatplotLib.
- Tensorflow:
  - Library tool can also be used along with Python.
  - Multiple library functionalities and ready-made models are also available in tensorflow.
- PyTorch:
  - Designing the neural network architecture.
  - Provide high performance code.
- Keras:
  - Rapid iteration process needed for deep learning neural networks.

Levels of NLP

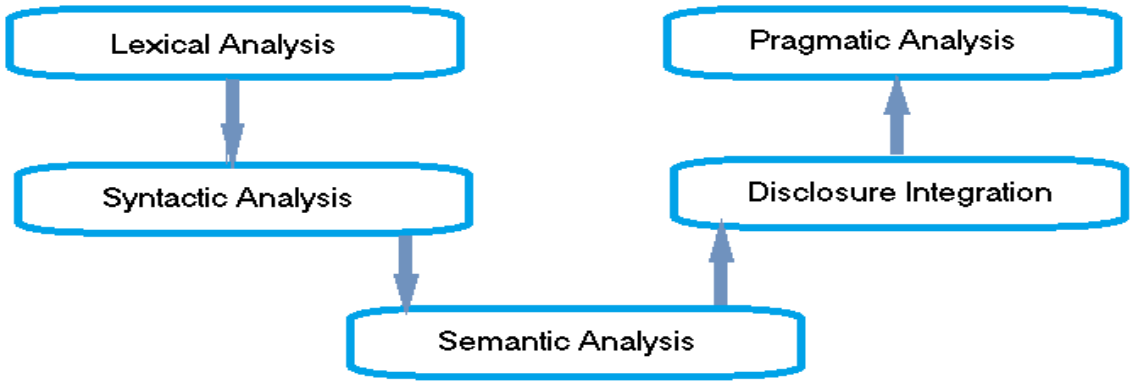


Figure: Levels of NLP

Self evaluation: Exercise 1

- To continue with the training, after learning the basics of Linguistic Structure and various steps in Natural Language Processing, it is time to start writing code to perform tasks at a basic level. It is instructed to utilize the concepts of reading data from files to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 1: Python code to read data from a CSV file, Process the data and Create Files based on the Data.

.

Self evaluation: Exercise 2

- To continue with the training, after learning the basics of Linguistic Structure and various steps in Natural Language Processing, it is time to start writing code to perform tasks at a basic level. It is instructed to utilize the concepts of reading data from files to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 2: Code to read data from a CSV file, Process, Subset, Merge and Clean the given Text Data.

Lexical analysis

- Identifying and analyzing the structure of words.
- Lexicon: Collection of words and phrases in a language.
- Lexical analysis: Divide the whole chunk of txt into paragraphs, sentences, and words

- Token: Token name and optional token value.

Token name	Sample token values
identifier	x, color, UP
keyword	**if**, **while**, **return**
separator	}, (, ;
operator	+, <, =
literal	true, 6.02e23, "music"
comment	*/* Retrieves user data */, // must be negative*

Pre-processing activity


```
texto = re.sub(r'''['|'|`]m''',' am', texto)
texto = re.sub(r'''['|'|`]re''',' are', texto)


sent_tokenizer=nltk.data.load('tokenizers/punkt/english.pickle')
sentences = sent_tokenizer.tokenize(texto)


pattern =r'''(?x)
        ([A-Z]\.)+
        | \w+['|']s
        | \w+(-\w+)*
        | \$?\d+(\.\d+)?%?
        | \.\.\.
        | [][\.,;"?():-_!]
    '''
token = []
for sentence in sentences:
    words = nltk.tokenize.regexp_tokenize(sentence, pattern)
    token.append(words)
```

POS tagging (1 of 2)

Abbreviation	Meaning
CC	coordinating conjunction
CD	cardinal digit
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subordinating conjunction
JJ	adjective (large)
JJR	adjective, comparative (larger)
JJS	adjective, superlative (largest)
LS	list market
MD	modal (could, will)
NN	noun, singular (cat, tree)
NNS	noun plural (desks)
NNP	proper noun, singular (sarah)
NNPS	proper noun, plural (indians or americans)
PDT	predeterminer (all, both, half)
POS	possessive ending (parent\ 's)
PRP	personal pronoun (hers, herself, him,himself)
PRP$	possessive pronoun (her, his, mine, my, our)
RB	adverb (occasionally, swiftly)
RBR	adverb, comparative (greater)
RBS	adverb, superlative (biggest)
RP	particle (about)
TO	infinite marker (to)
UH	interjection (goodbye)
VB	verb (ask)
VBG	verb gerund (judging)
VBD	verb past tense (pleaded)
VBN	verb past participle (reunified)
VBP	verb, present tense not 3rd person singular(wrap)
VBZ	verb, present tense with 3rd person singular (bases)
WDT	wh-determiner (that, what)
WP	wh- pronoun (who)
WRB	wh- adverb (how)

POS tagging (2 of 2)

- Lexical categories.
- Called as word classes.
- Tag set of that language.

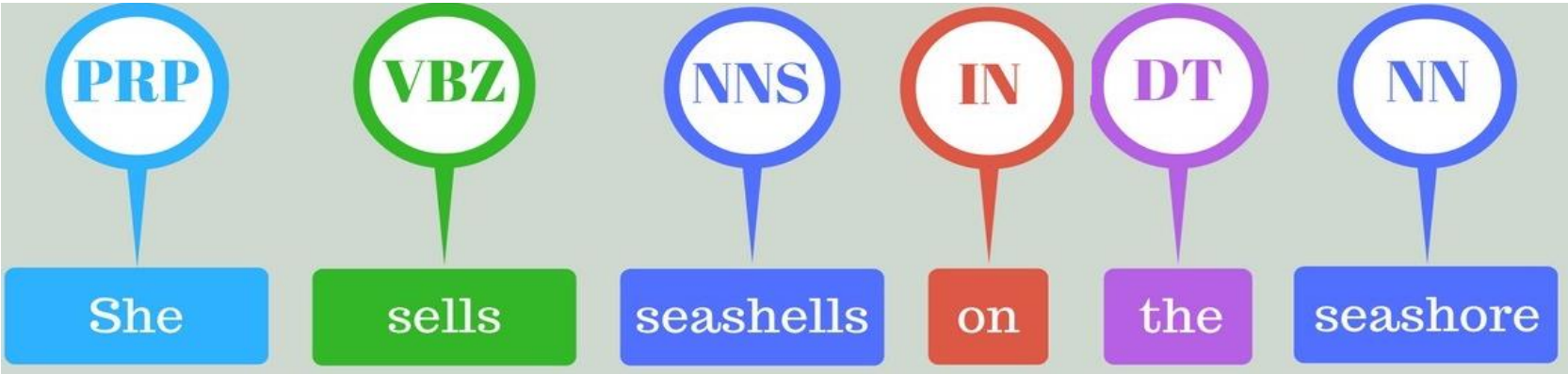


Figure: Tag set of language.

Syntactic parsing

- Third phase during text analysis process.
- Identification of the meaningfulness present in each text.
- Identifying the meaning of the text that belongs to a natural language.
- Within the boundary rules of the grammar of that language.

- Parser: Tokens from lexical analyzer: Meaningful representation.
- Uses symbol table.

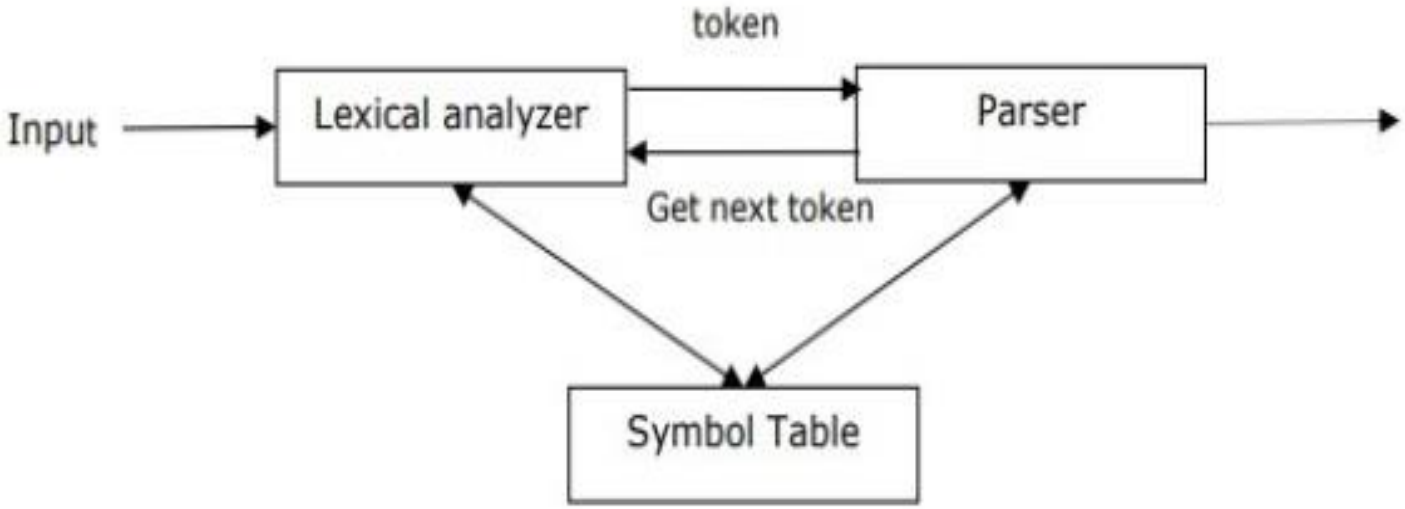


Figure: Syntactic parsing

Source: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_syntactic_analysis.htm

Types of parsing

- Top down parsing:
  - Starts at the start symbol and proceeds further.
  - Every input symbol is read and recursively proceeds to process every element in the input.
  - Due to the recursive nature backtracking occurs.

- Bottom up parsing:
  - Inverse of top down parsing.
  - Derivation is defined as the production rules that are used during parsing.

Derivation logic

- Leftmost derivation: The provided input string is scanned from left to right in sentential form.
- Rightmost derivation: The provided input string is scanned from right to left in sentential form.

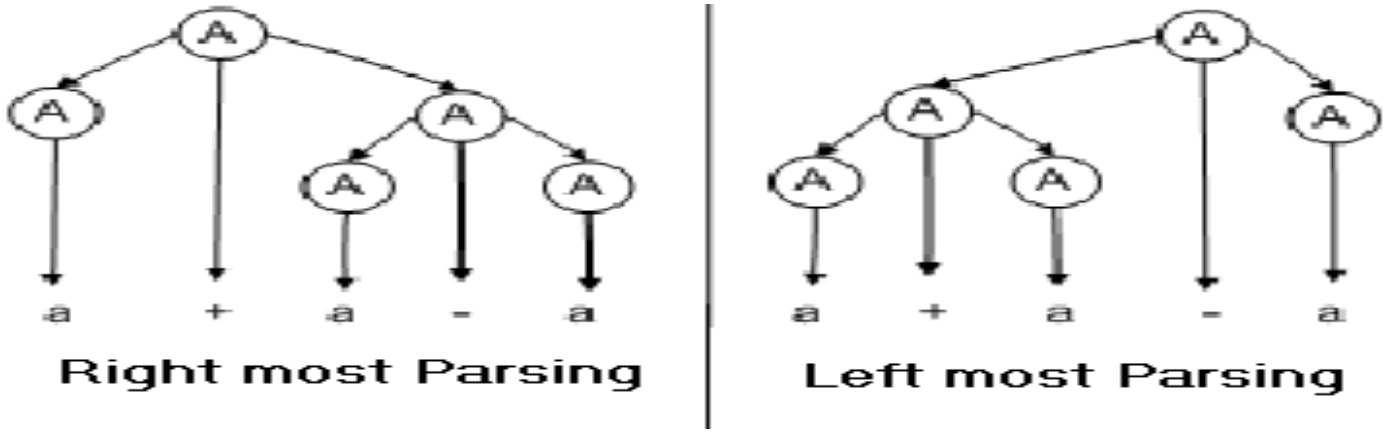


Figure: Parse trees

Source: SELF

Grammar

- Grammar defined by G for any language is written in 4 tuple format as (N, T, S, P).

Constituency grammar:

- Noam Chomsky grammar: Subject predicate division methodology.
- Major clauses: Noun phrase and verb phrase.

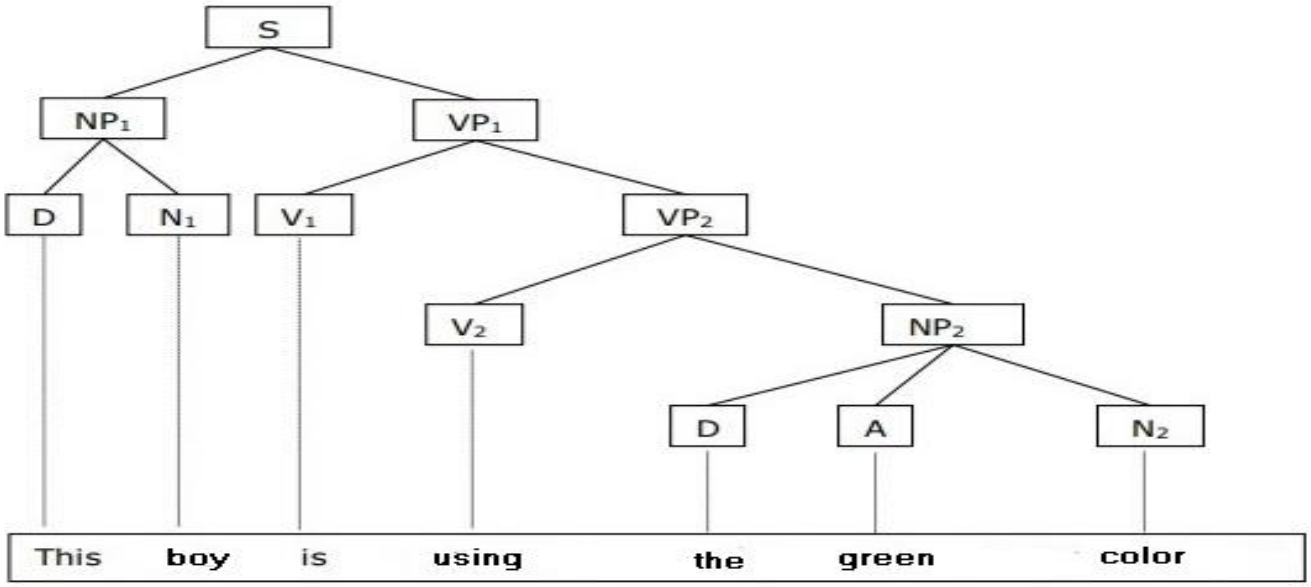


Figure: Noam Chomsky is based upon constituency relationship

Source: SELF

Semantic analysis

- Concept identify and extract the meaning specified according to a language dictionary.
- Semantic analyzer: Extraction of the meaningfulness.
- Identify the meaning of individual words.
- Identify the meaning of the sentence: Combining the meaning of the individual words.

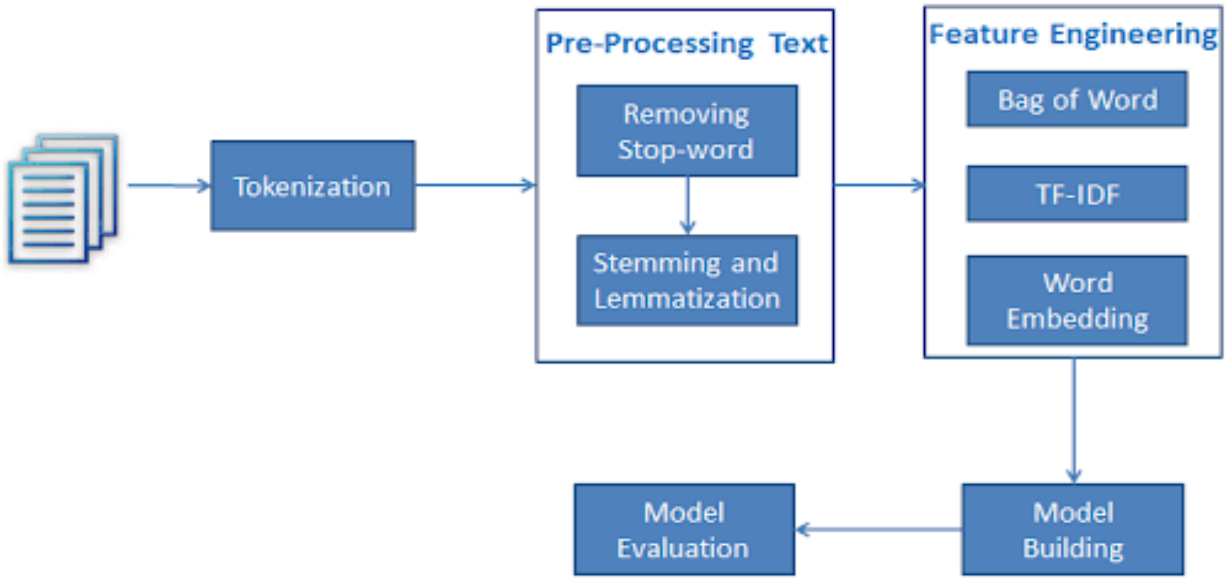


Figure: Semantic analysis

Source: https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Semantic analysis elements

Synonymy: Relationship between two elements: expressing similar meaning.

- Antonymy: Relationship between two lexical items: provide dissimilarity in their meaning.

- Hyponymy: Represents the relationship between generic term: Instances.
  - Example:
  - Hypernym - Color
  - Hyponym - Red, Green, Blue

- Homonymy: Relationship between two or more words: similar spelling: different meaning
  - Example: Tablet - Medicine or Computer Device

Representation in semantic analysis

- Semantic analysis core components
- Entities: Noun representation.
- Concepts: Generic noun representation.
- Relations: Relationship between entities and concepts.
- Predicates: Word representation.

Self evaluation: Exercise 3

- To continue with the training, after learning the various steps involved in Natural Language Processing, it is instructed to utilize the concepts of Lexical, Syntactic and Semantic Analysis to process text from files to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 3: Read a text and perform basic Pre processing activities on the text like Removal of Stop words, Tokenization using SPACY.

Self evaluation: Exercise 4

- To continue with the training, after learning the various steps involved in Natural Language Processing, it is instructed to utilize the concepts of Lexical, Syntactic and Semantic Analysis to process text from files to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 4: Read any available text information, perform pre-processing on the data, Remove Stop words, Perform Tokenization, Classify the document words based on specific topics with Modelling.

Natural language generation

- Natural language understanding: Identification of the natural language elements.
- Natural language processing: Understanding the nuances of language.
- Natural language generation: Generating simple natural language responses.

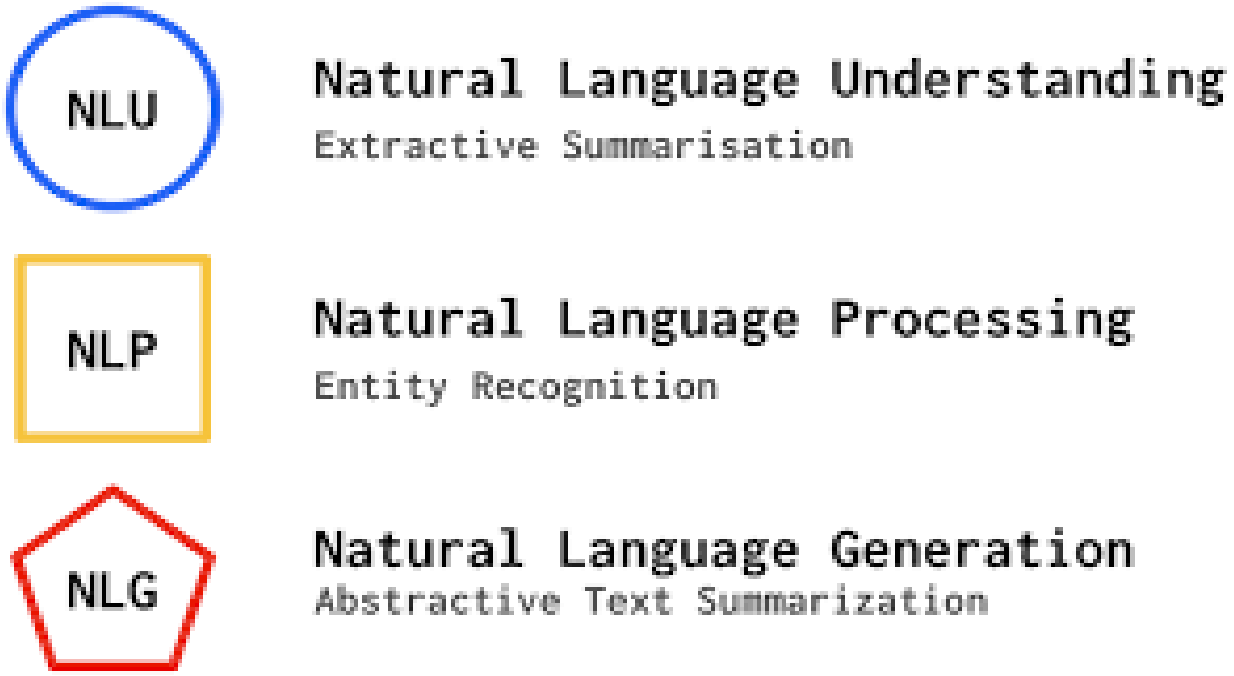


Figure: Natural language generation

Source: https://wordlift.io/blog/en/advanced-seo-natural-language-processing/

NLP vs NLG

- NLP: Describe a machine's ability to ingest what is said.
- NLU: Subset of NLP: Handle unstructured inputs: Structured form.
- NLG: Processes turn structured data into text.

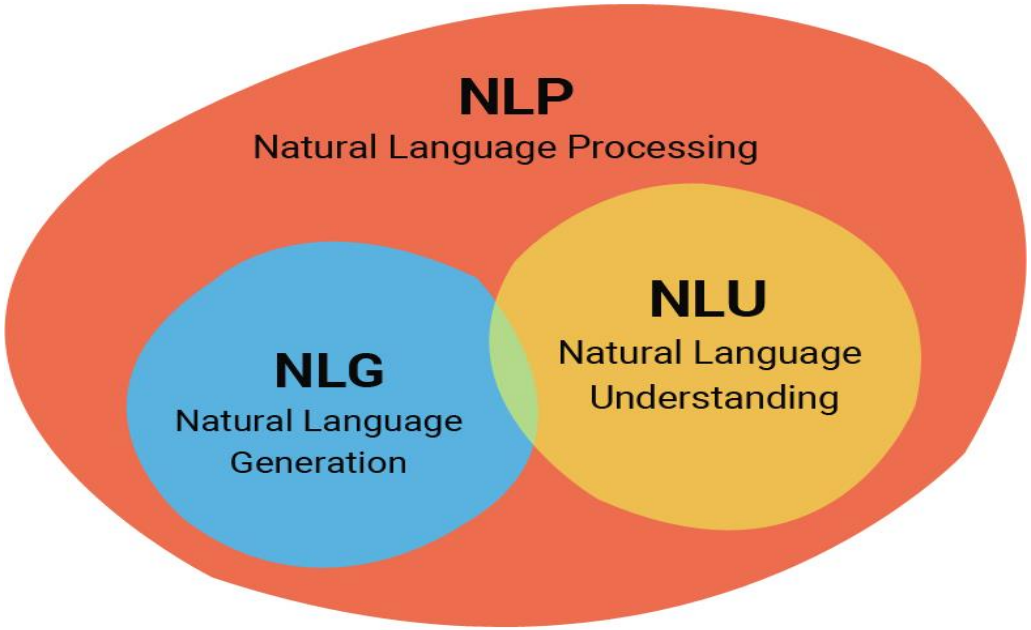


Figure: NLP vs NLG

Source: https://www.aismartz.com/blog/the-past-and-the-presence-of-natural-language-generation/

History of NLG

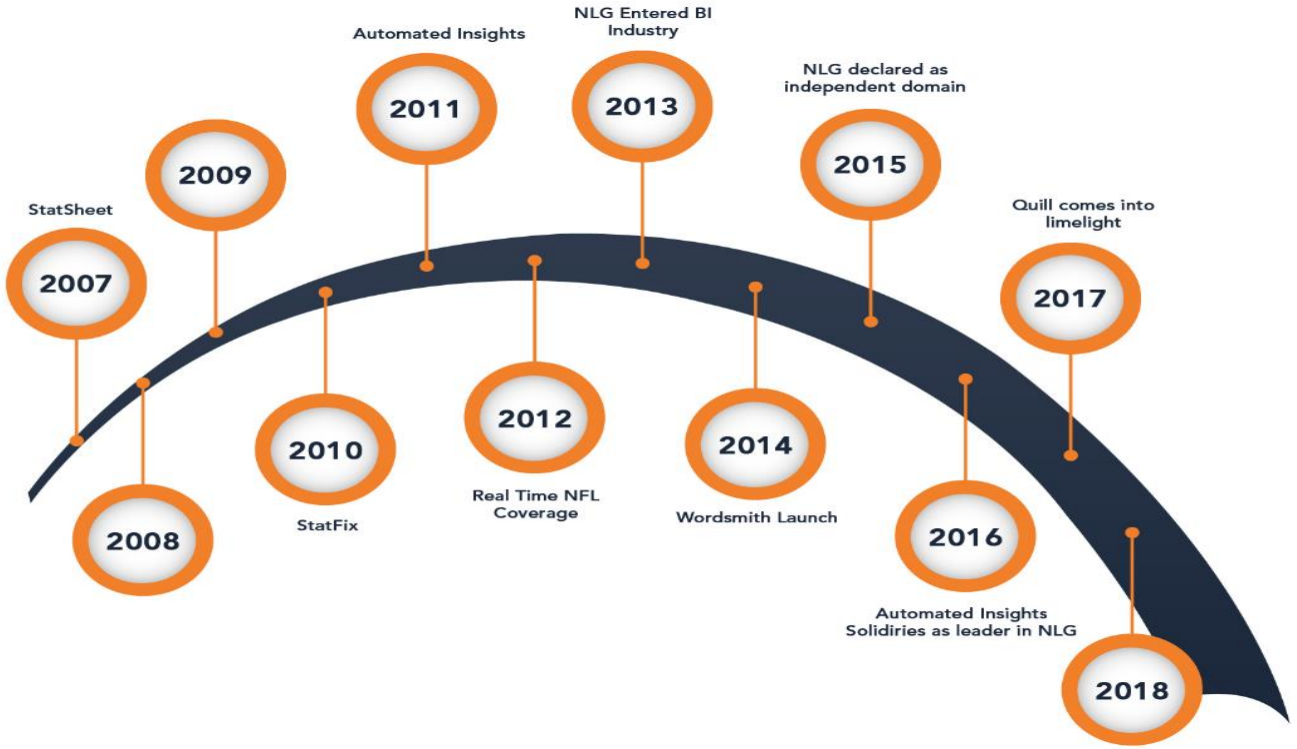


Figure: History of NLG

Source: https://www.aismartz.com/blog/the-past-and-the-presence-of-natural-language-generation/

Working principle of natural language generation

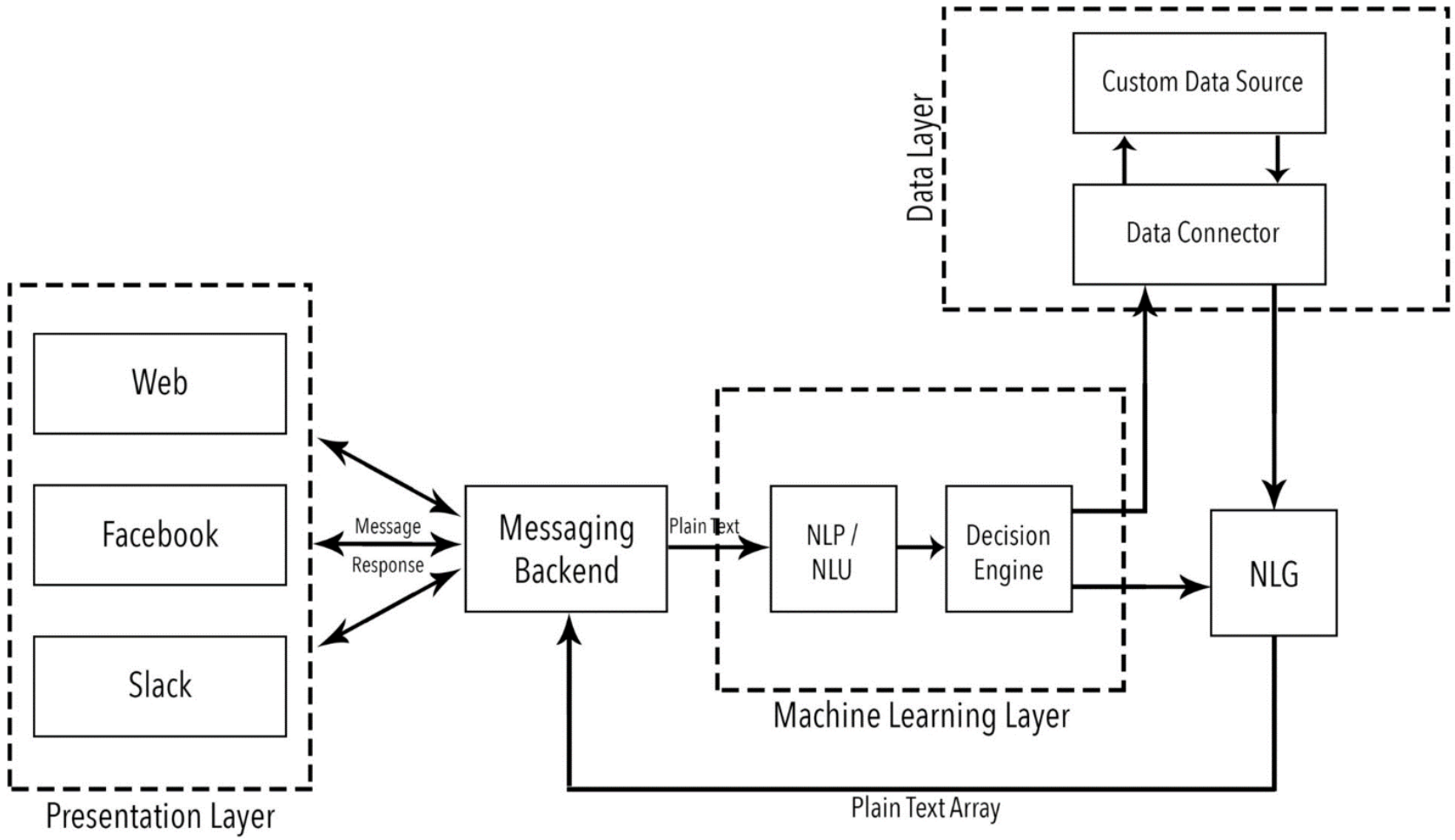


Figure: Working principle of natural language generation

Source: https://chatbotslife.com/nlp-nlu-nlg-and-how-chatbots-work-dd7861dfc9df

Limitations in natural language generation

- Difference in the stress level.
- Empathy not easily identified.
- Capture: Nuance in the communication.
- Response in natural language as per the context.