NATURAL LANGUAGE PROCESSING

NLP – Master 2 - AIBD - 2024/2025

NLP: Subject matter

- Coefficient : 2 Credit : 3
- Evaluation :
  - Attendance/2 pts,
  - Test/10 pts,
  - Oral report/8 pts,

NLP

- Links:
  - Blog: http://nlp-dz.blogspot.com
  - E-mail: mistudents14@gmail.com
  - Course: elearning.univ-km.dz

References

- J. Eisenstein, Introduction to Natural Language Processing. MIT Press. 2019

- D. Jurafsky, J. H. Martin, Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3rd ed. 2023

DeepLearning.AI https://www.deeplearning.ai/

towards data science https://towardsdatascience.com/

Top NLP & Data science leaders to follow

PLAN

- Chapter 1 : NLP: An Overview

  Definition, NLU vs NLG, Processing levels, Apps,..

- Chapter 2 : Text preprocessing & Data representation

  Preprocessing: Segmentation, Text cleaning, Normalization, Tokenization,

  Text encoding: BOW, One-Hot, N-Ggrams, TF-IDF, Word2Vec,..

  Corpora and Datasets

- Chapter 3 : Morpholexical analysis

  Stemming, Lemmatization, Word formatting, Flexion vs Derivation

- Chapter 4 : Syntactical analysis

  Word categorization, Relations & Structures, Grammar, Syntactical tree

- Chapter 5 : Semantic & Pragmatic

  Word sense, Semantic similarity, Topic modeling, Language model

- Chapter 6 : Recent advances (Open-Content)

  Transformer, Attention, Pretrained Language Models

- Chapter 7: NLP System Evaluation

CHAPTER I

NLP: An Overview

INTRO - PLAN

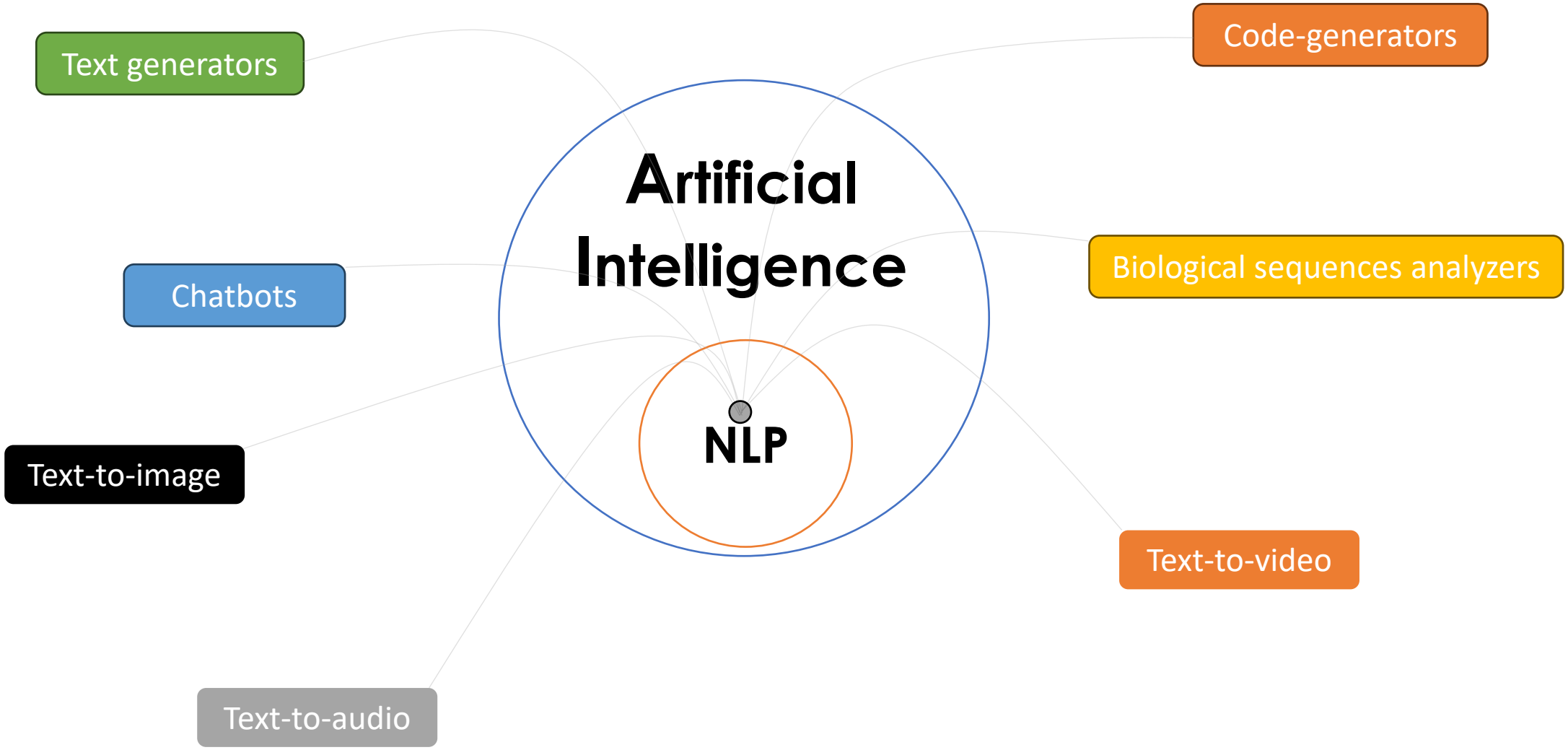
- What is NLP?
- NLP Applications
- How does NLP work?
- NLP techniques
- Libraries and Frameworks for NLP

The Top 50 Gen AI Web Products, by Unique Monthly Visits

1. ChatGPT	11. IIElevenLabs	21. PhotoRoom	31. PIXAI	41. MaxAI.me
2. Gemini	12. Hugging Face	22. YODAYO	32. ideogram	42. Craiyon
3. character.ai	13. Leonardo.Ai	23. Clipchamp	33. invideo AI	43. OpusClip
4. liner	14. Midjourney	24. runway	34. Replicate	44. BLACKBOX AI
5. QuillBot	15. SpicyChat	25. YOU	35. Playground	45. CHATPDF
6. Poe	16. Gamma	26. DeepAI	36. Suno	46. PIXELCUT
7. perplexity	17. Crushon AI	27. Eightify	37. Chub.ai	47. Vectorizer.AI
8. JanitorAI	18. cutout.pro	28. candy.ai	38. Speechify	48. DREAMGF
9. CIVITAI	19. PIXLR	29. NightCafe	39. phind	49. Photomyne
10. Claude	20. VEED.IO	30. VocalRemover	40. NovelAI	50. Otter.ai

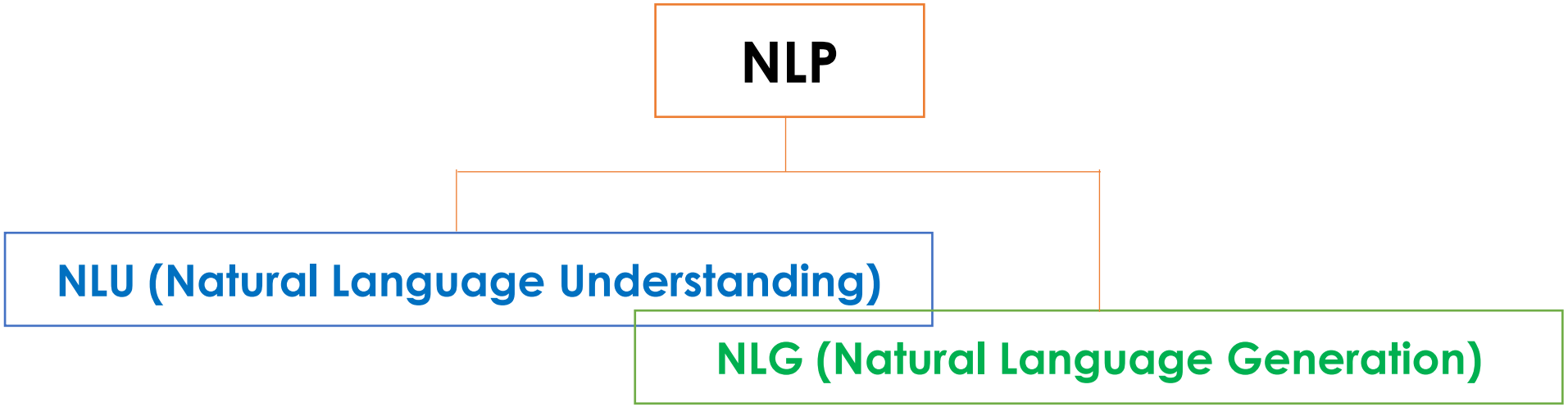
https://www.popwebtools.com/blog/top-generative-ai-web-products

NLP: A fast-growing research field in AI

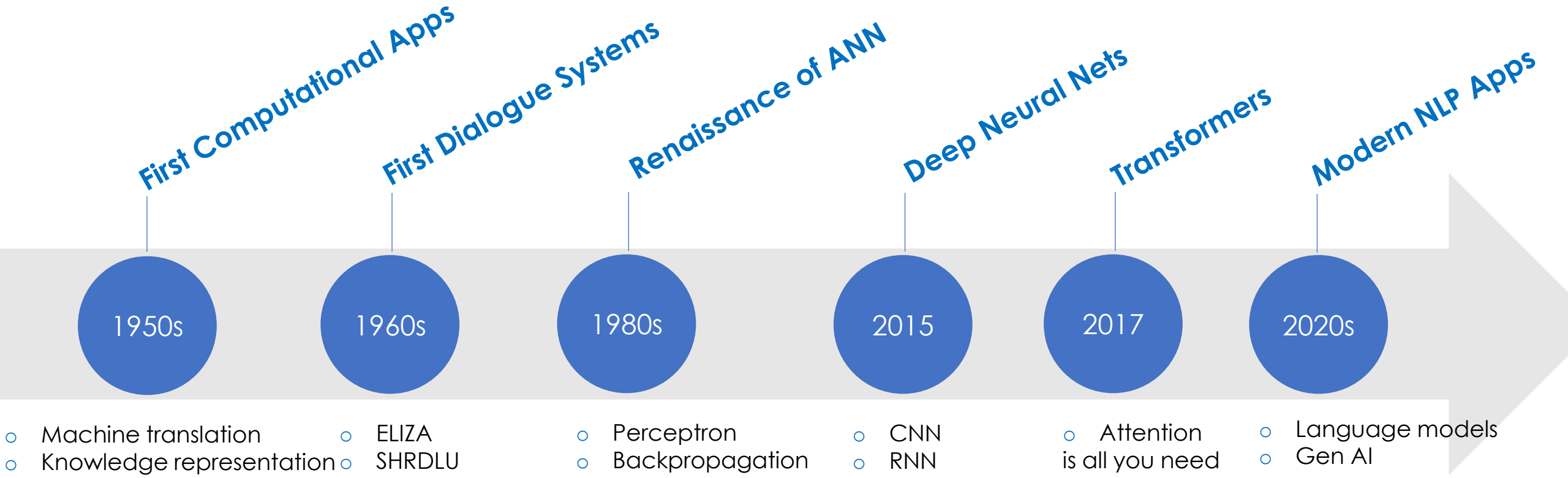


What is NLP ?

How to program computers to **analyze** the meanings of input text and **generate** meaningful, expressive output.



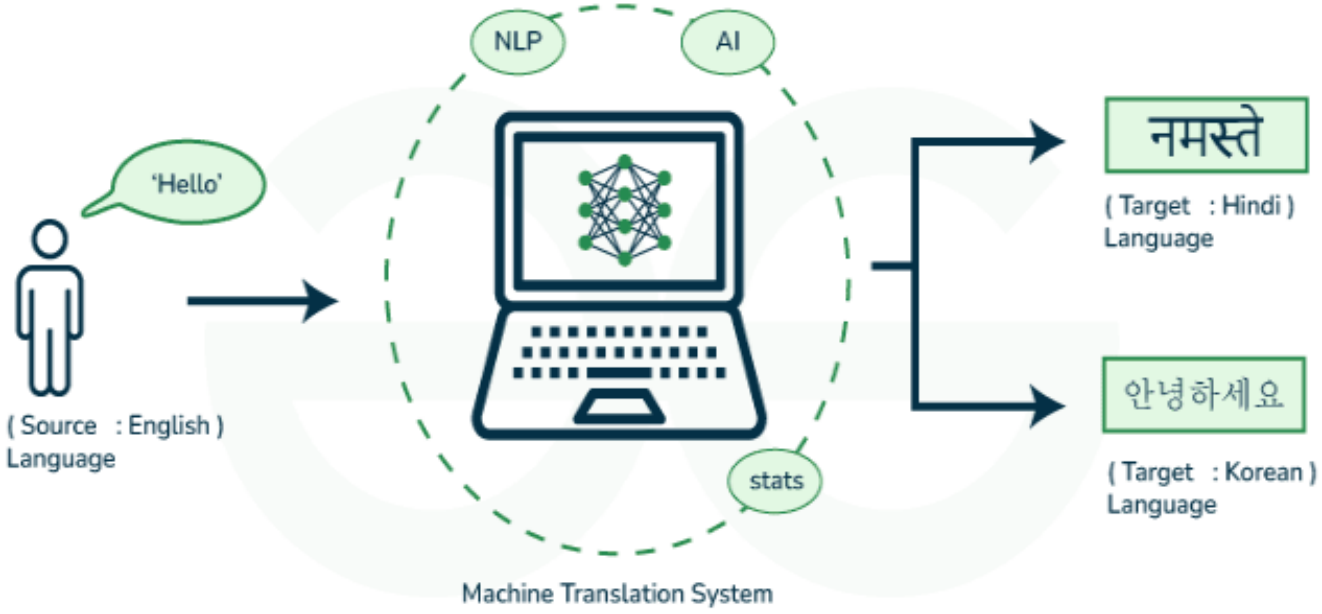
The early days of NLP

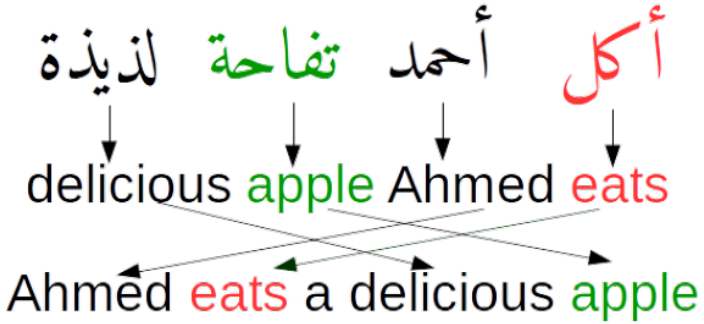


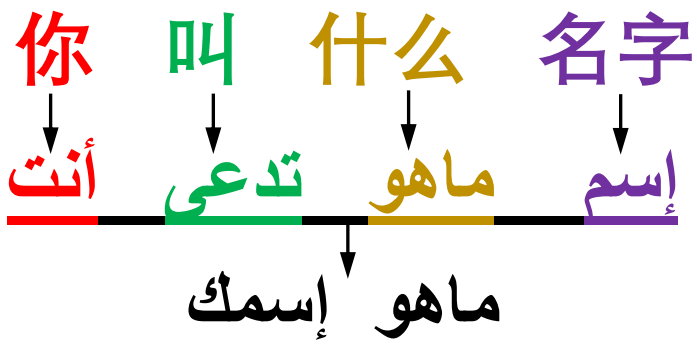
Main NLP Applications

1. Machine Translation

- **Rule-based**
  (Grammar rules and dictionaries)
- **Statistical**
  (Examine extensive human translations)
- **Neural**
  (Training on Source-Target language dataset)
- **Hybrid**
  (Use of multiple machine translation models)



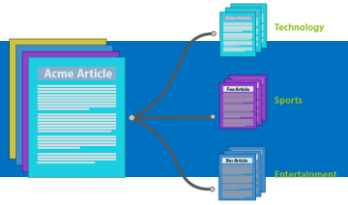




(nǐ jiào shénme míngzì)

Main NLP Applications

2. Text Classification



- **Document classification**

  (Document categorization: Techno, Sport, Art,..)

- **Sentiment analysis**

  (Classifying emotional quality)

- **Toxicity detection**

  (Detecting threats, insults, hatred towards entities)

- **Spam detection**

  (Classify emails as either spam or not)

- **Hadith authentication**

  (Verify originality of Prophetic Hadiths)

- **Misinformation and Fake news detection,…**

Main NLP Applications

3. Named Entity Recognition

Extract entities in a piece of text into predefined categories such as personal **names**, **organizations**, **locations**, and **quantities**.

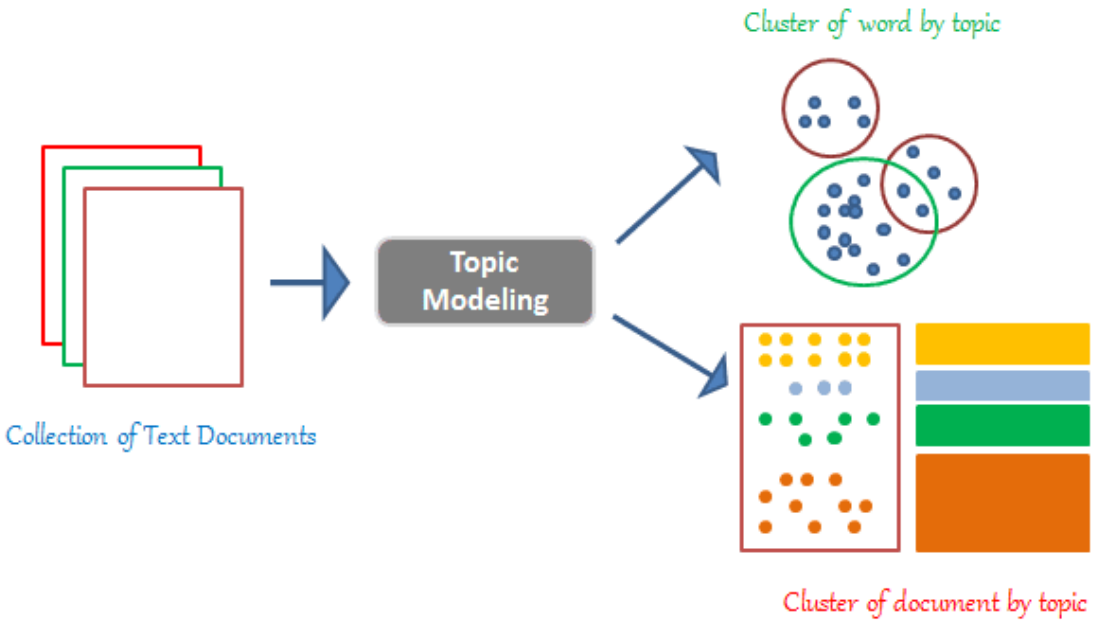
Main NLP Applications

4. Topic Modeling

Unsupervised text mining task that takes a corpus of documents and discovers abstract topics within that corpus.

Techniques:

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)
- LDA2Vec
- BERTopic



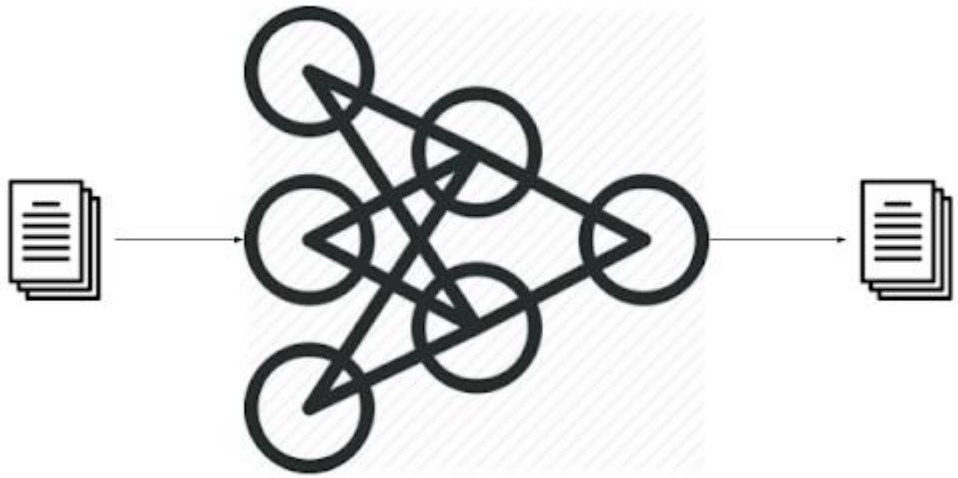
Main NLP Applications

5. Text Generation

Automatically produces text that is similar to human-written text (such as: Tweets, Blogs, Essays, Computer code,..): LSTM-RNN, BERT, BARD, ChatGPT,...

Variations:

- **Autocomplete:** predicts what word comes next
- **Chatbots:** automate one side of a conversation
  - Questions & Answers database
  - Conversation generation



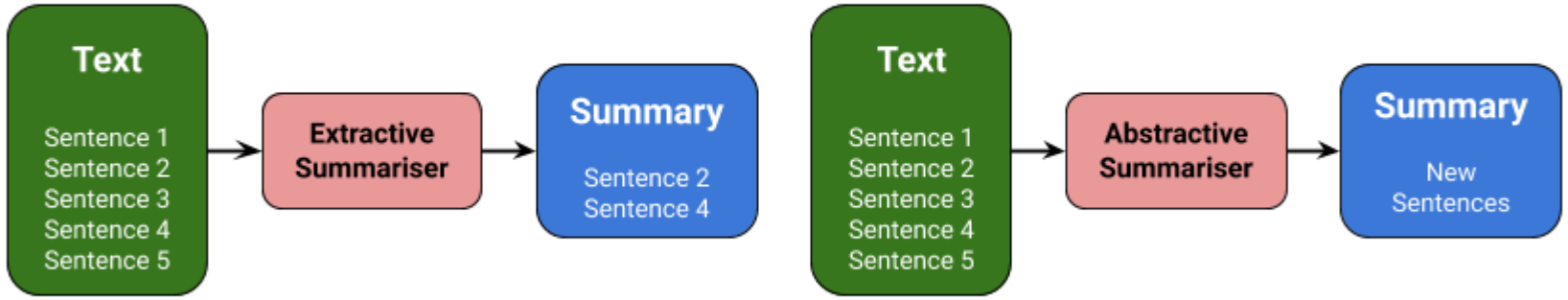
Main NLP Applications

6. Text Summarization

Shortening text to highlight the most relevant information

Variations:

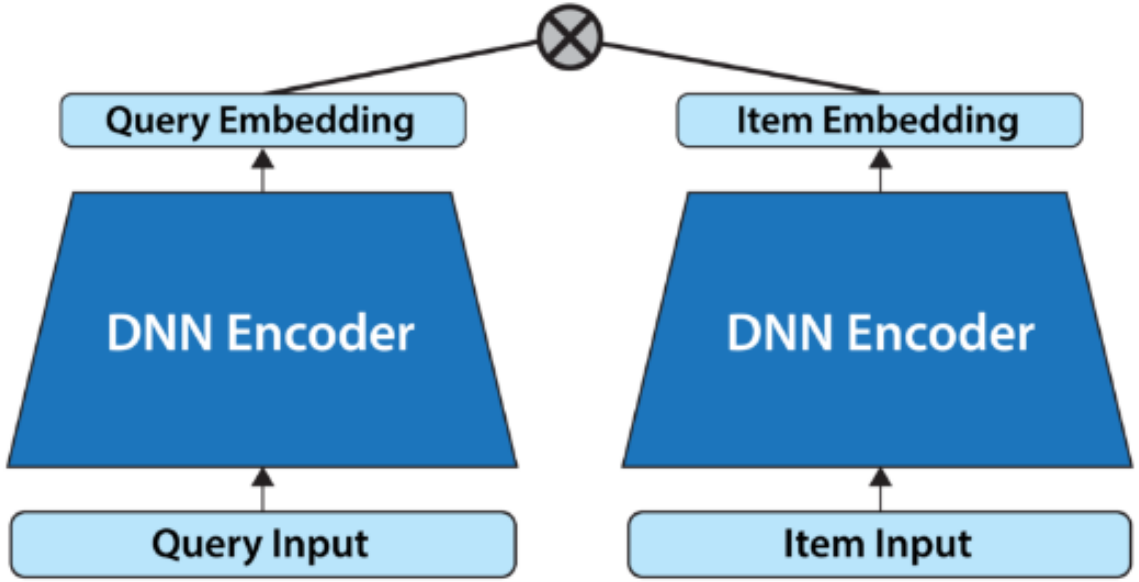
- **Extraction:** extracting the most important sentences from a long text and combining these to form a summary
- **Abstraction:** writing the abstract that includes words and sentences that are not present in the original text



Main NLP Applications

7. Information Retrieval

Finds (indexing and matching) the documents that are most relevant to a query.



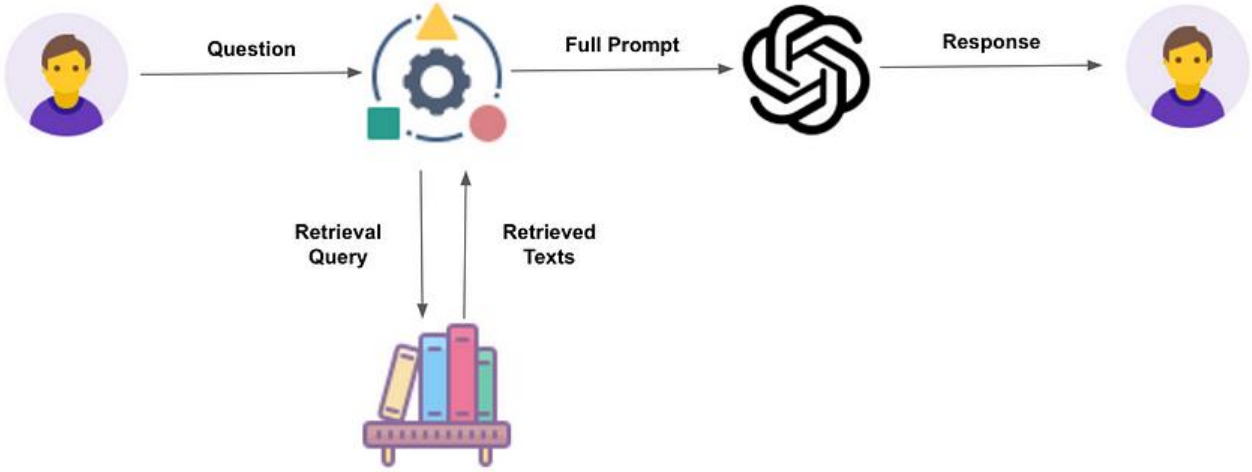
- **Indexing:** using a vector space
- **Matching:** using similarity score

Main NLP Applications

8. Question/Answering

Answering questions asked by humans in a natural language

- **Multiple choice:** question problem is composed of a question and a set of possible answers
- **Open-domain:** the model provides answers to questions in natural language without any options provided

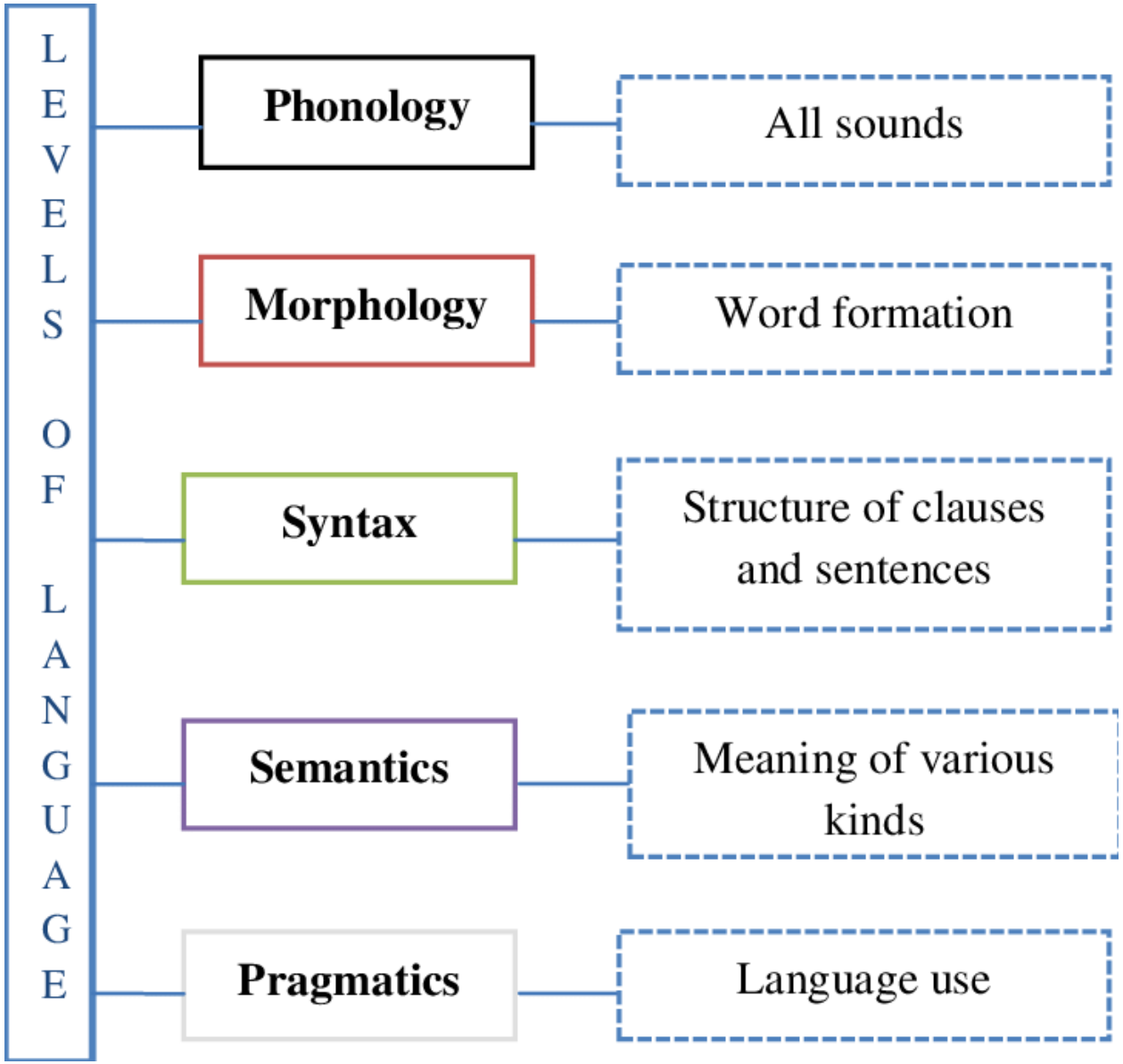


Main NLP Applications

9. Other NLP apps

- **Grammatical error correction:** encode grammatical rules to correct the grammar within text.
- **Part-of-Speech Tagging:** classifying words in a text according to their grammatical categories (such as noun, verb, and adjective).
- **Language modeling:** building models that predict the probability of a sequence of words.
- **Speech recognition:** transform spoken language into a machine-readable format.

NLP Processing levels



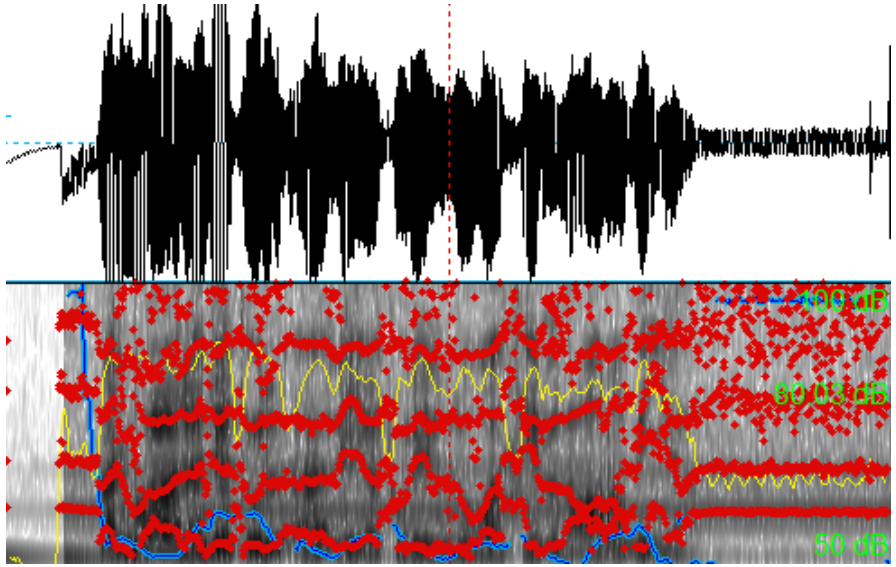
NLP Processing levels

LEVELS OF LANGUAGE

Phonology	All sounds
Morphology	Word formation
Syntax	Structure of clauses and sentences
Semantics	Meaning of various kinds
Pragmatics	Language use

- Phoneme detection
- Prosody identification
- Transitions

العربية | اللغة | الآلية | المعالجة



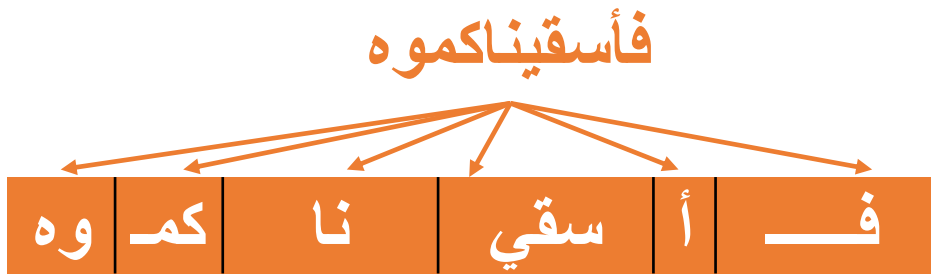
al-muʿālağatu al-ʾālīatu liluġati al-ʿārabīa

NLP Processing levels

LEVELS OF LANGUAGE

- **Phonology** — All sounds
- **Morphology** — Word formation
- **Syntax** — Structure of clauses and sentences
- **Semantics** — Meaning of various kinds
- **Pragmatics** — Language use

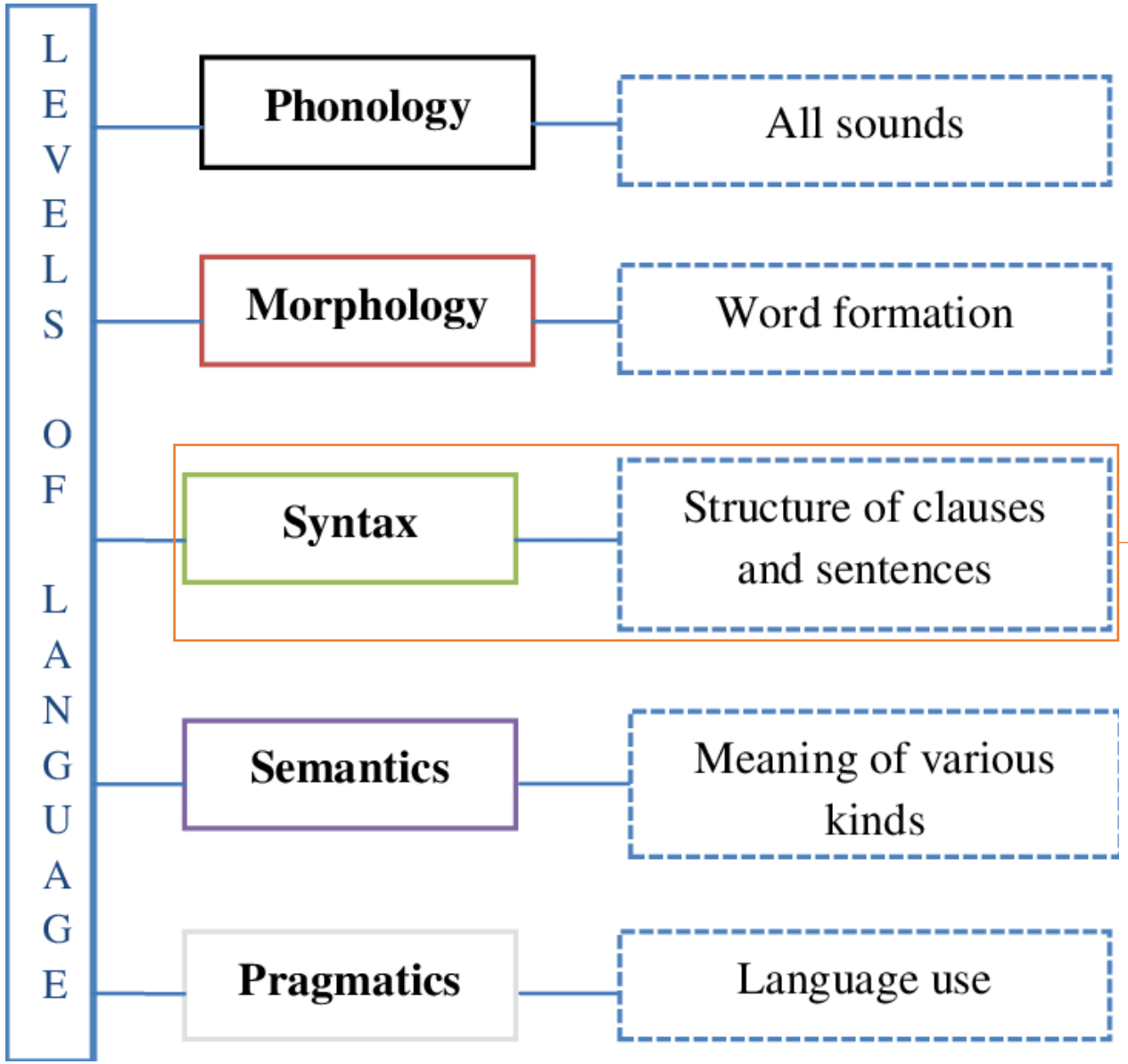
- Morphemes
- Tokens

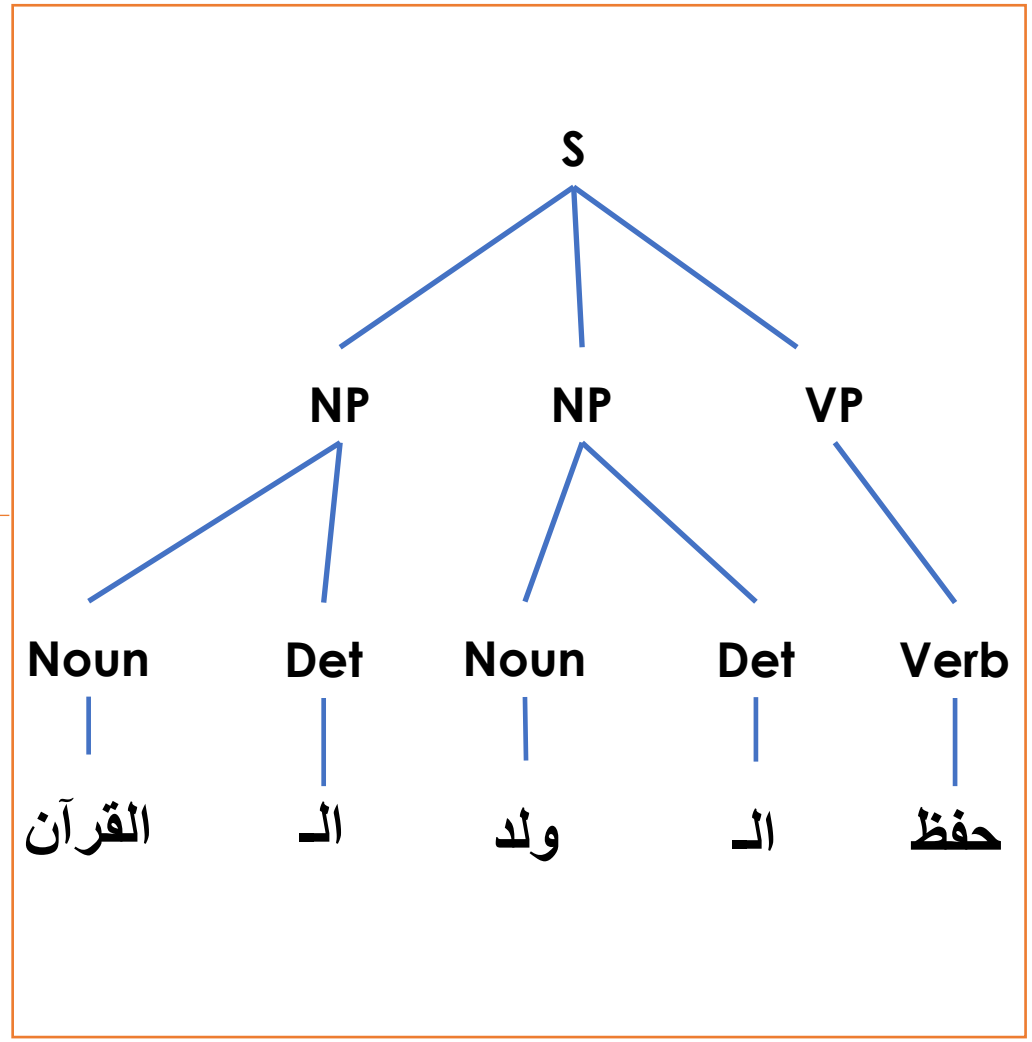


يذهب محمد إلى المسجد كل يوم

يذهب، محمد، إلى، المسجد، كل، يوم

NLP Processing levels

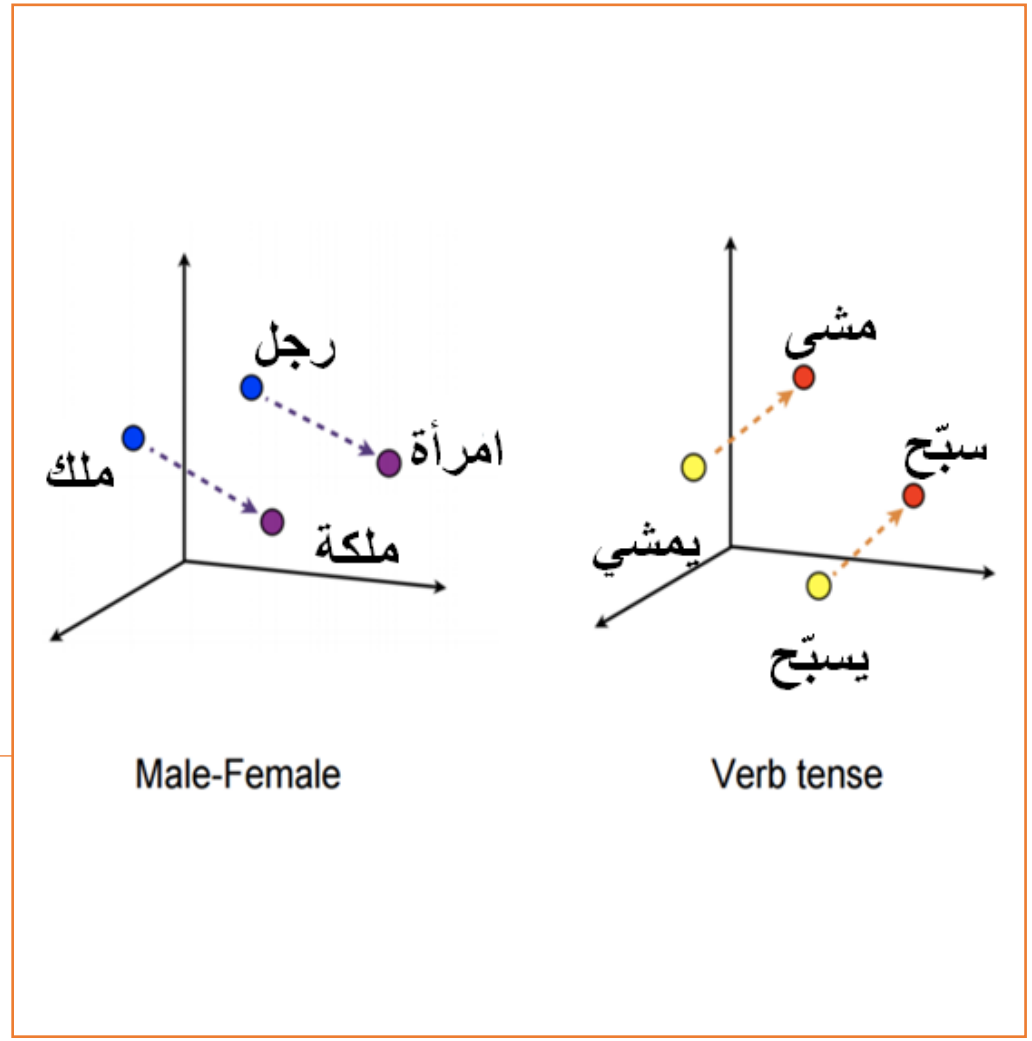




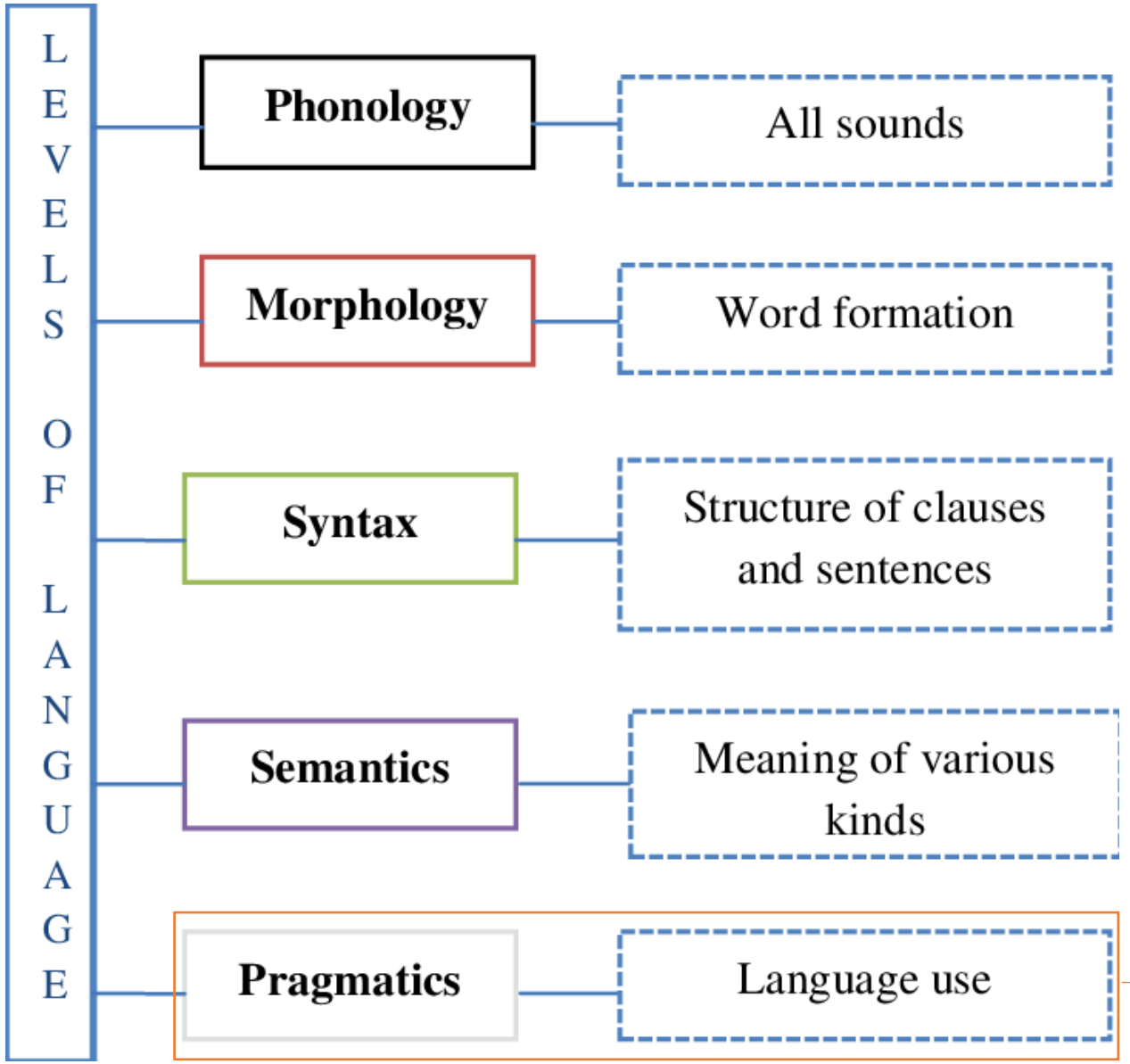
NLP Processing levels

LEVELS OF LANGUAGE

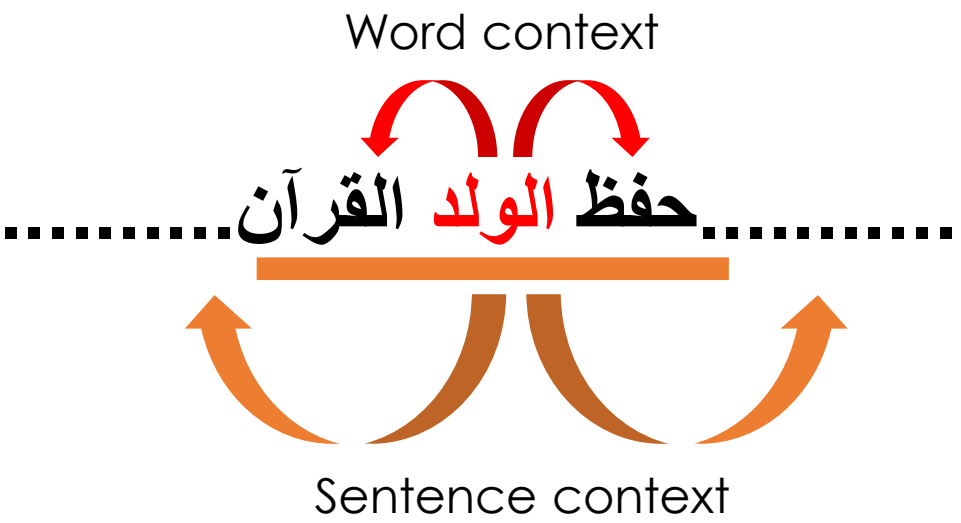
Phonology	All sounds
Morphology	Word formation
Syntax	Structure of clauses and sentences
Semantics	Meaning of various kinds
Pragmatics	Language use



NLP Processing levels



- Meaning in the context



How does NLP work?

Data preprocessing

↓

Feature extraction

↓

Fed into NLP architecture

Text preprocessing

Prepare the text data for the model building. It is the very first step of NLP projects. It improves data quality, reduces noise, and facilitates effective analysis and modeling.

- **Steps**
  - Removing punctuations like . , ! $( ) * % @
  - Removing URLs
  - Removing Stop words
  - Lower casing
  - Tokenization
  - Stemming
  - Lemmatization

Feature extraction

Extracting features from text, such as word frequencies, n-grams, or word embeddings, which are essential for building machine learning models.

- **Techniques**
  - Bag-of-Words
  - One-Hot-Encoding
  - N-Grams
  - TF-IDF
  - Word Embeddings
    - Word2Vec (CBoW, Skip-Gram)
    - GLoVE

NLP techniques

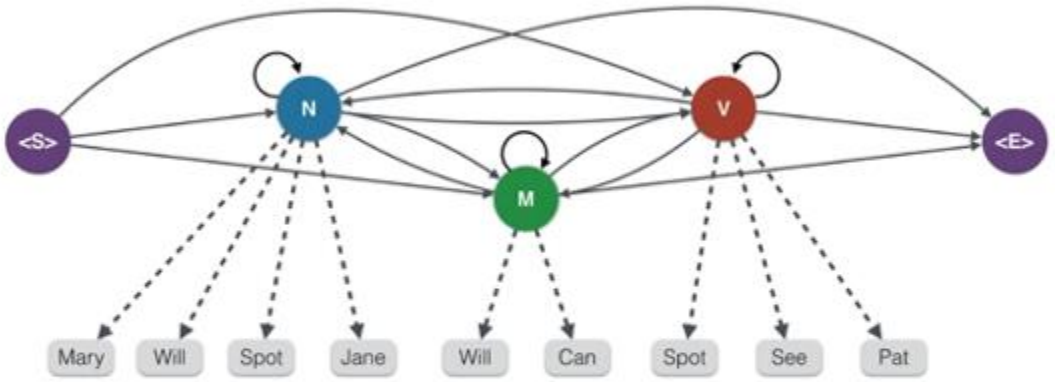
- **Traditional machine learning techniques**
  - Logistic regression
  - Naive Bayes
  - Decision trees
  - LDA/LSA
  - Hidden Markov Models (HMM)
- **Deep learning techniques**
  - Convolutional Neural Networks (CNN)
  - Recurrent Neural Networks (RNN)
  - Autoencoders
  - Seq2Seq models
  - Transformers

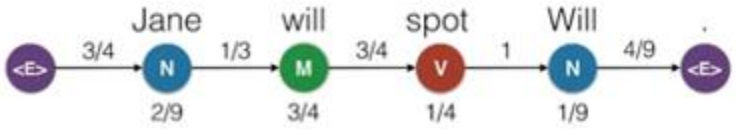
NLP techniques

POS-Tagging with HMM

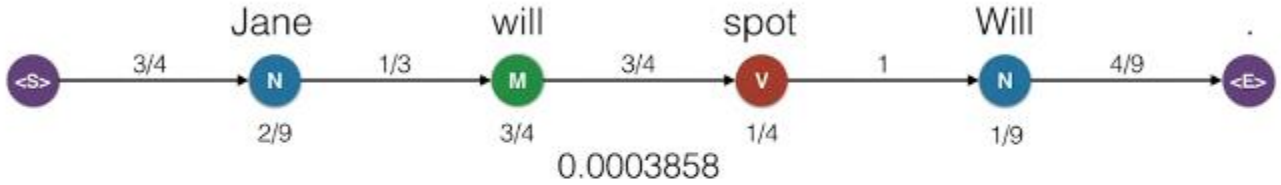
S = Jane will spot Will

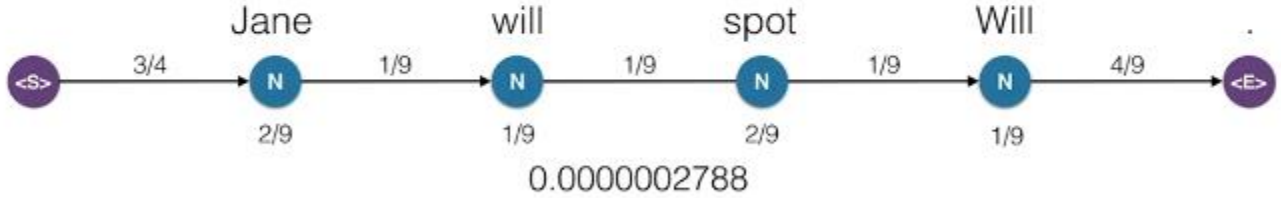
What will be the most likely assignment for each word**?**





High probability sequence

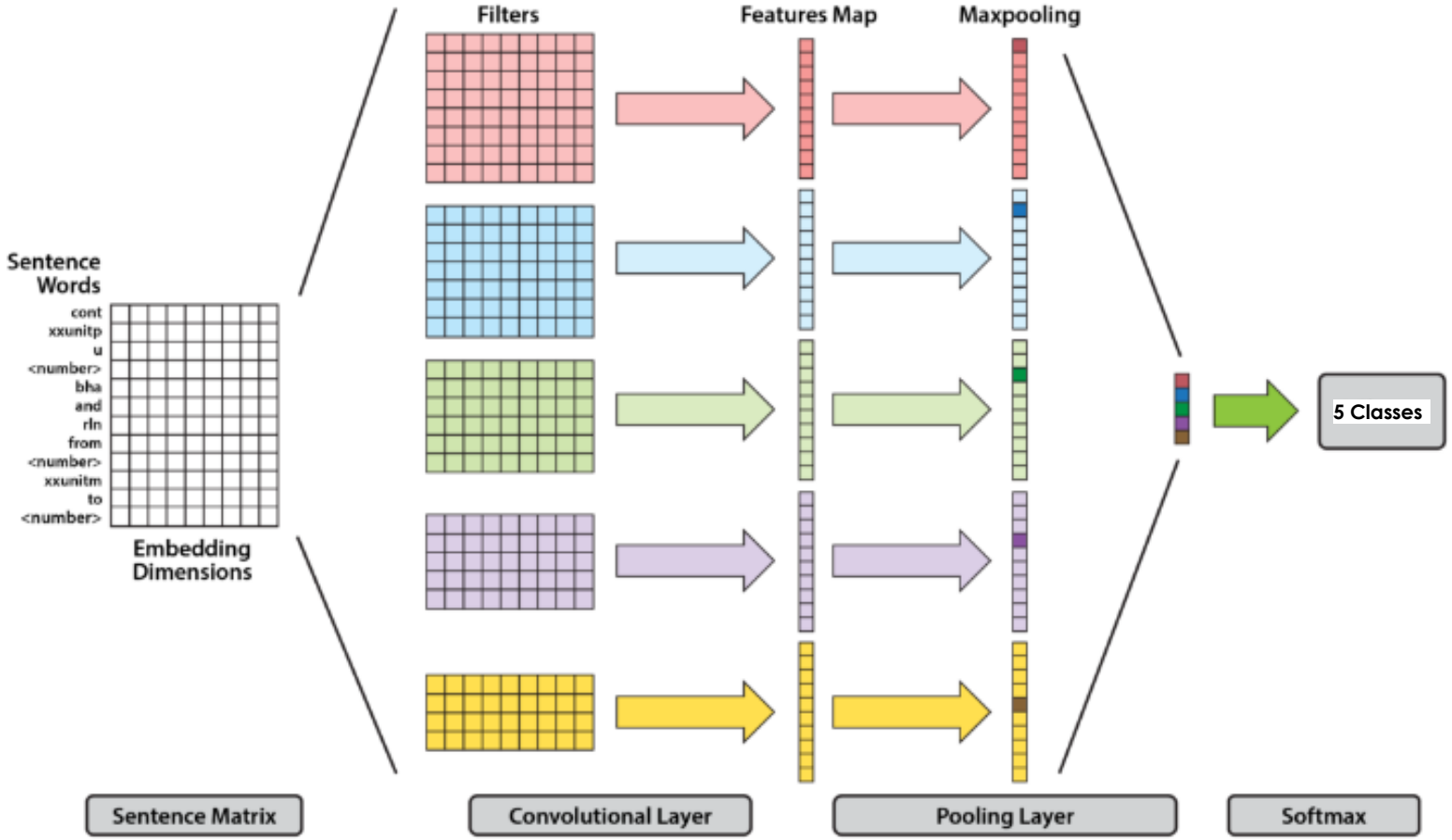




NLP techniques

CNN-Based text classfication

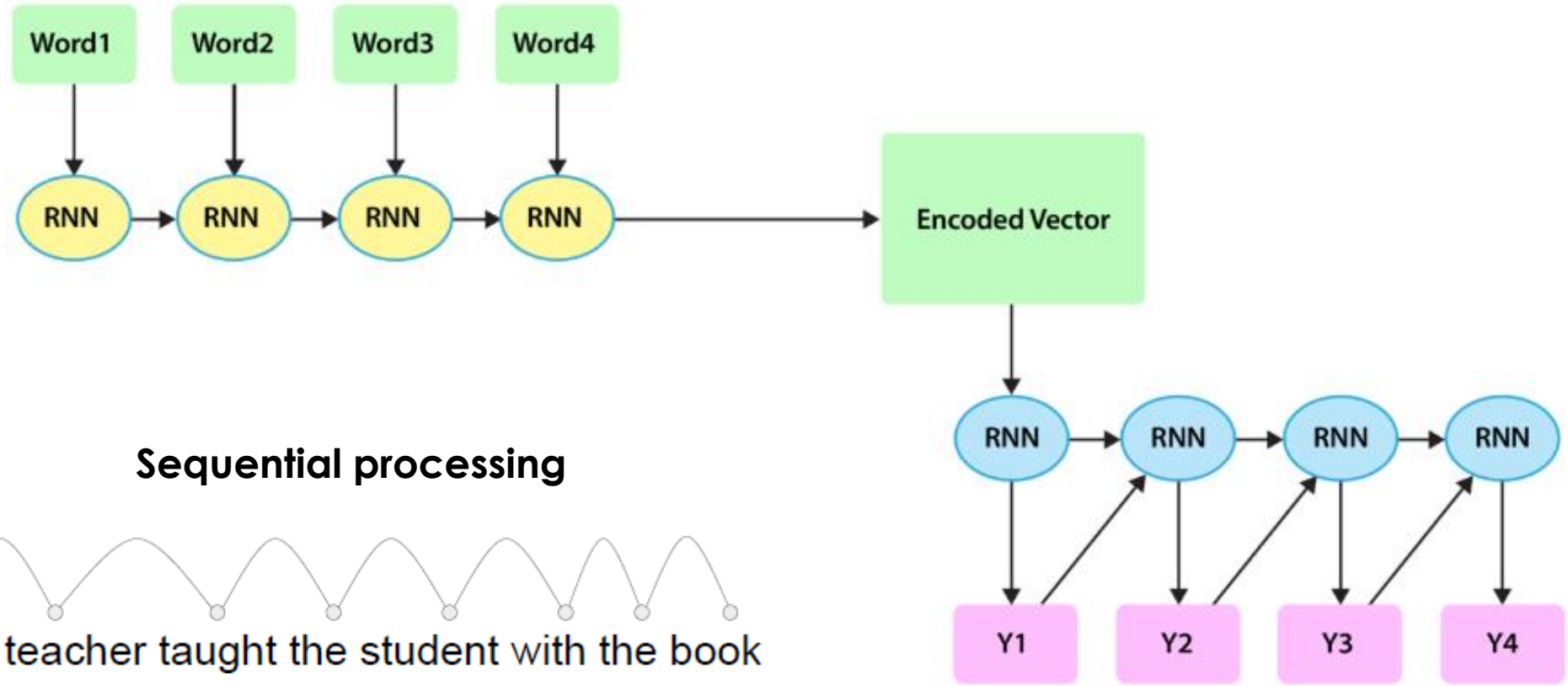
Given a sentence, a CNN uses convolutional layers to refine represenations of input words, before combining them to render a classification



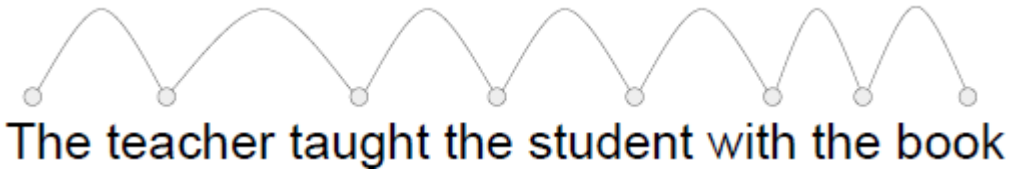
NLP techniques

RNN-Based Seq2Seq model for Machine translation

Given a sentence, a RNN encodes the sequence and then iteratively generates a translation



Sequential processing



NLP techniques

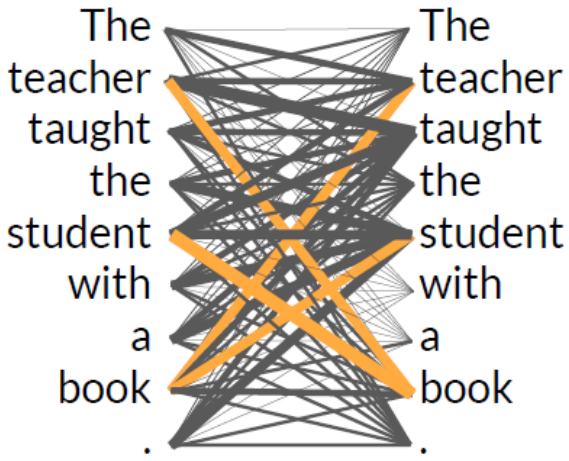
Transformer architecture

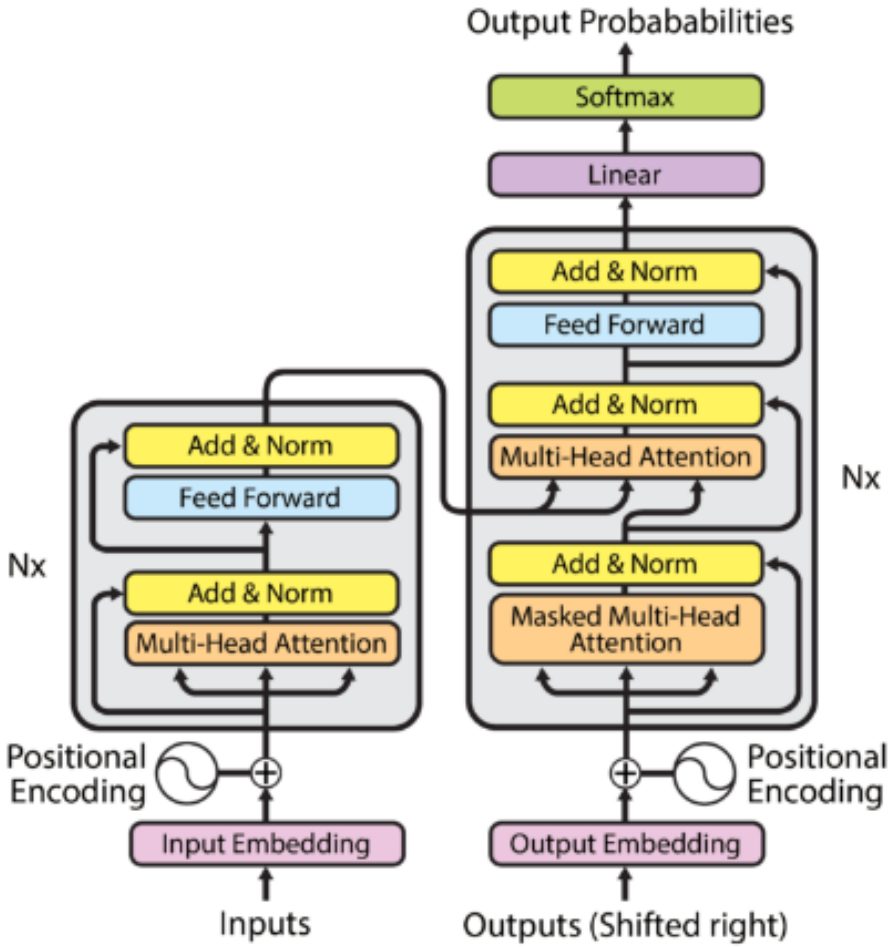
Relies entirely on a self-attention mechanism to draw global dependencies between input and output,

It is at the core of new language models:

- **Autoencoder (Encoder only): BERT, ROBERTA**
- **Autoregressive (Decoder only): GPT, BLOOM**
- **Seq2Seq (Encoder-Decoder): T5, BART**

Parallel processing





NLP: Programming languages, libraries and Frameworks

- **Python**
  - Natural Language Toolkit (NLTK)
  - scikit-learn (Traditional machine learning algorithms)
  - spaCy
  - Deep learning libraries (keras, Tensorflow, PyTorch)
  - Gensim
  - Hugging Face: open-source models and implementations
- **R**
  - TidyText
  - Weka
  - SpaCyR, Tensorflow, PyTorch
- **JavaScript, Java, Julia**
- **Cloud platforms:** Google colab, Kaggle, AWS,…

Presentations

1. Machine translation
2. Text summarization
3. Speech recognition
4. Text classification (News article categorization)
5. Sentiment analysis
6. Hadith authentication
7. Spam detection
8. Toxicity detection
9. Text To Speech (TTS)
10. Question/Answering
11. Fake news detection

- Presentations (.ppt) should include (but not limited to):
  - Theory behind
  - Algorithms
  - Demo (implementation)
- Presentations in English or Arabic
- First presentation : Oct. 18, 2024

LAB – Part 1

- **Text cleaning**
- **Word tokenization**
- **TF-IDF vectorization**
- **Word cloud**
- **Word embeddings**
  - CBOW, Skip-Gram
  - Semantic similarities
  - Operations on words