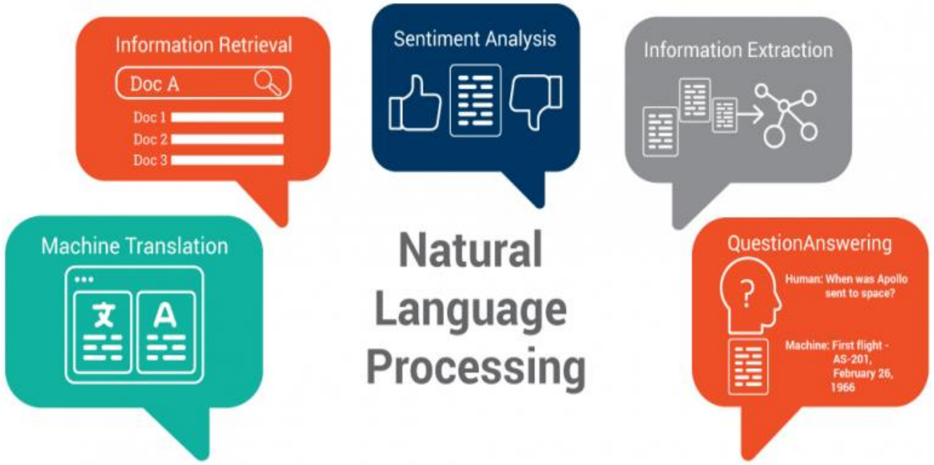


Recent trends in Natural Language Processing

Balayogi G

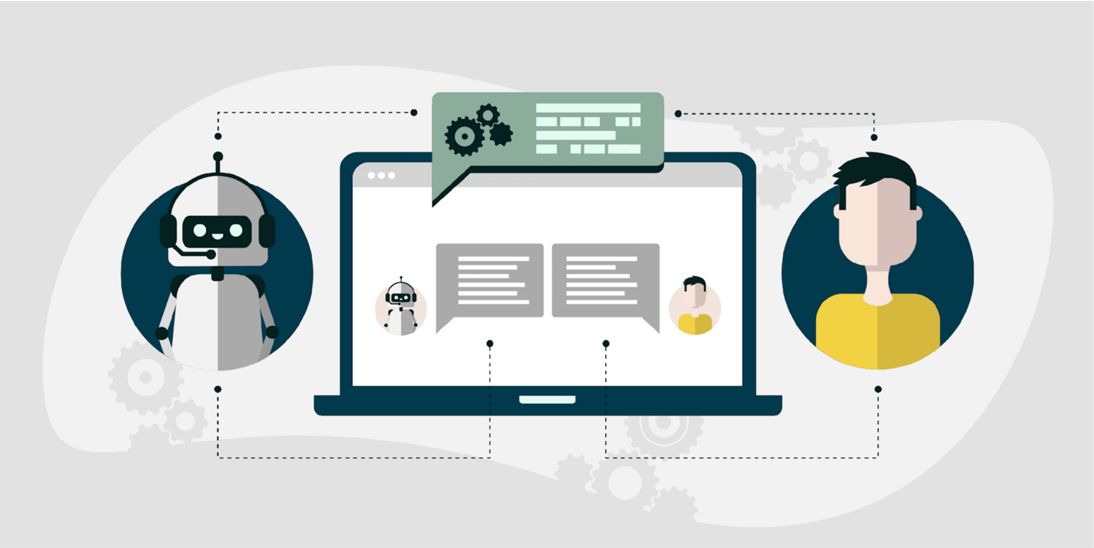
Balayogi Gnanasekar

Agenda

- What is NLP
- Motivation for NLP
  - Methods
    - Rule based
    - Statistical based
- Types of Neural Networks
- RNN Improvements
- State of the art
- Demonstration

What is NLP

- To analyze, understand, generate Human languages with the help of computers.
- An Interface between the humans and machines.
- E.g. Spell check, chat bots, Grammarly.



History

1950’s	Alan Turing Second paper “Computing machinery and intelligence”
1954	Georgetown experiment on fully automated Machine translation from Russian to English.
1966	ELIZA , First computer therapist Bot
1970’s	Conceptual dependency theory for NLP.
1980’s	Statistical Machine translation was developed.
1990’s	NLP Models to increase capabilities.
2006	IBM Watson
2010’s	NLP on homes – Siri, Alexa
2020*	AI powered Chatbots and more

Alan
Turing

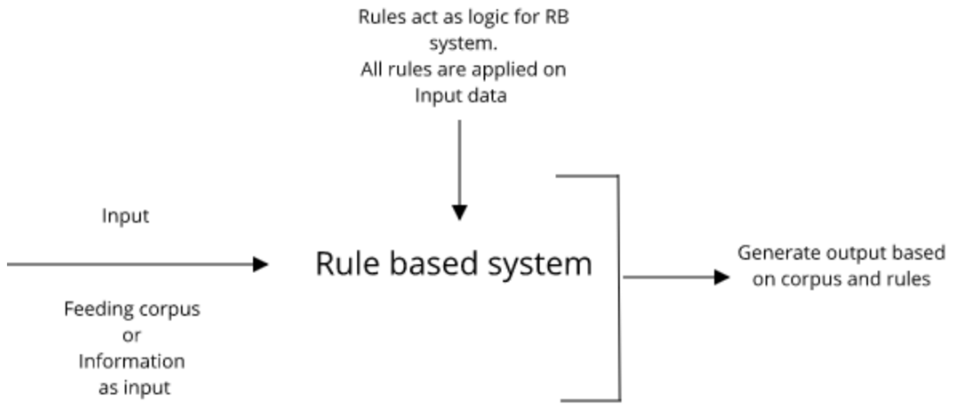
https://www.csee.umbc.edu/courses/471/papers/turing.pdf
http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf

Methods in NLP

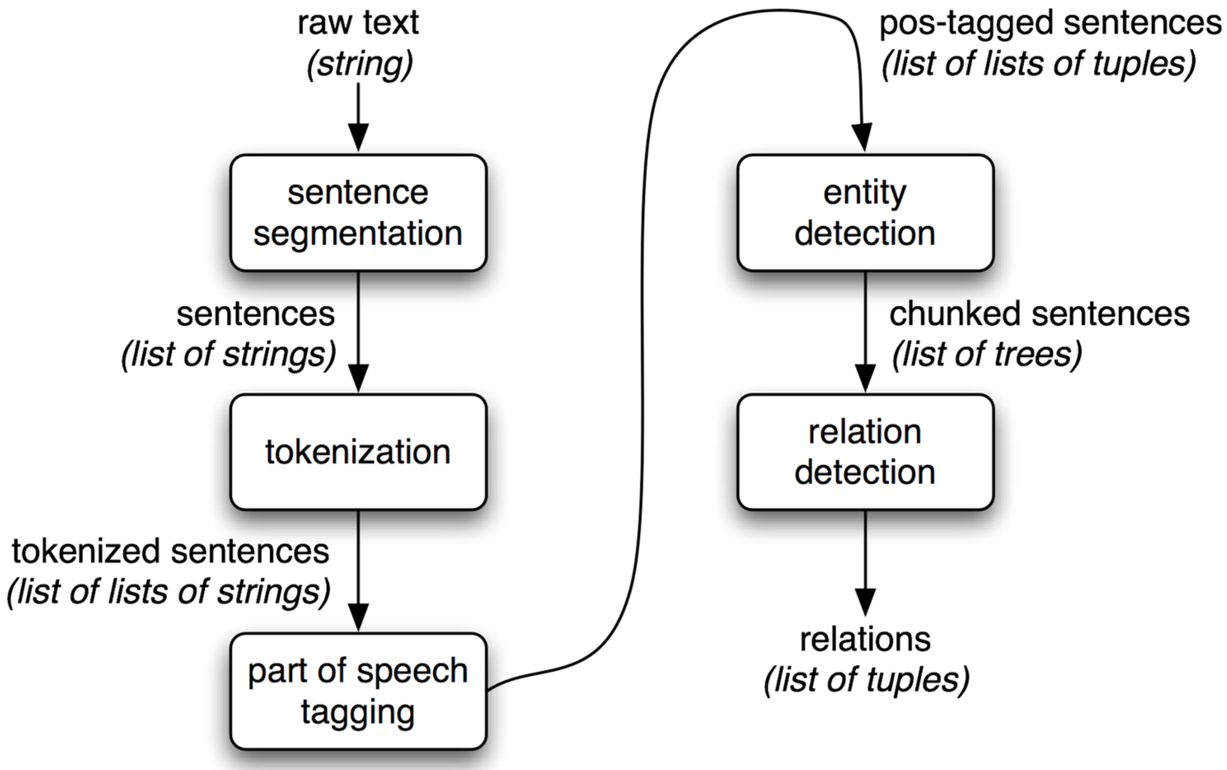
- Rule based NLP
  - Hard code (Grammars, Patterns, Heuristics, etc)
- Statistical based NLP
  - Learning rules from a large set of data(corpus) using statistical techniques like ML and DL.

Rule based NLP

- Perform well in simple , specific task. But can’t be generalize well.
- Some examples,
    - Preprocessing text
    - Searching for a specific pattern in a huge dataset.
    - Analysis of grammar rules.


Rules act as logic for RB system.
All rules are applied on Input data
Input
Feeding corpus or Information as input
Rule based system
Generate output based on corpus and rules

Rule based pipeline for NLP



Lemmatization

- Converting words into base form is called lemmatization.
- Python packages – Spacy, NLTK.
- e.g. Let’s take word, **Help**

Word	Lemma
Help	Help
Helping	Help
Helped	Help
Helps	Help

Stemming

- Converting words into stem word is called stemming.
- Python packages – Spacy, NLTK.
- e.g. Let's take few words,

Word	Lemma
Consign	Consign
Adjustable	Adjust
Studying	Study
Formality	Formaliti

Morphological Segmentation

- Converting word into Morphemes.
- Python packages, Spacy, NLTK, Polyglot.
- e.g. Let's take few words

Word	Morphemes
Governments	Govern – ment – s
Processing	Process – ing
Processor	Process – or
Invaluable	In – valuable

Word / Sentence Segmentation

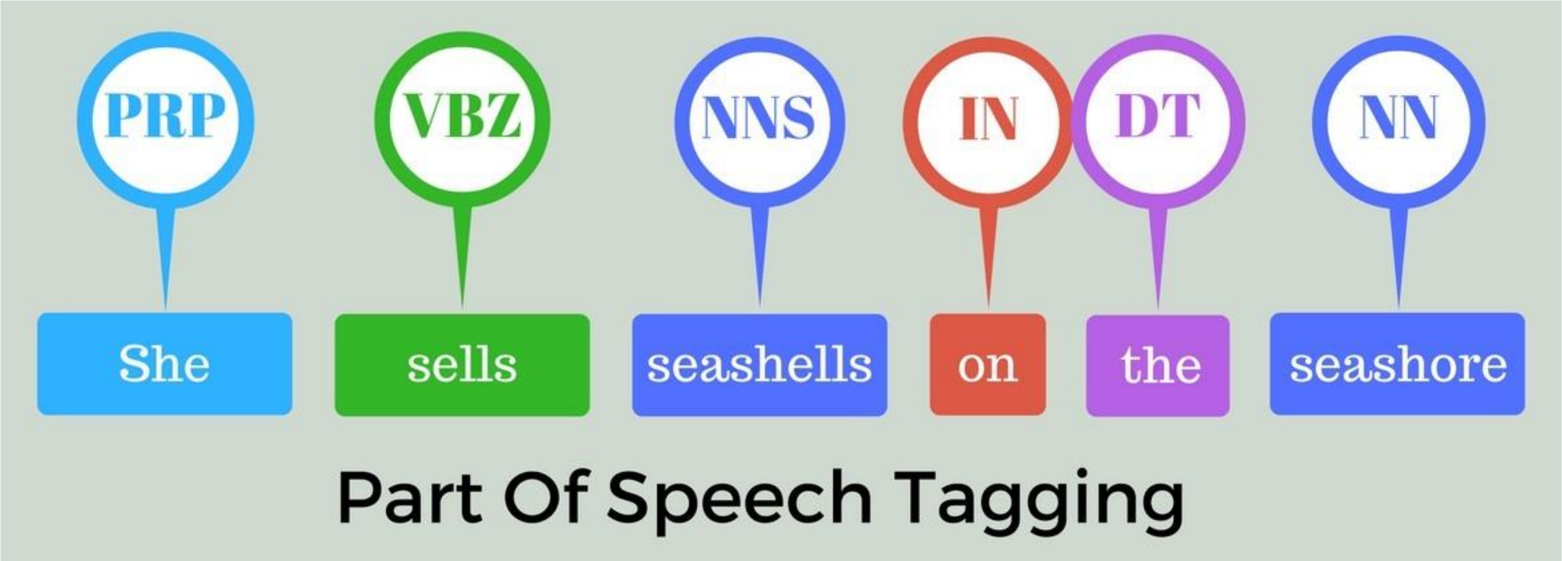
- Rule based approach: Morphological analysis based on lexical and grammatical knowledge.
- Corpus based approach: Learn words from corpus.

Continuous set of word	Word Segmentation
Whatdoesthisreferto	What does this refer to

Paragraph	Sentence Segmentation
He is a programmer. He always think about his work.	He is a programmer. He always think about his work.

POS-Tagging

- Rule based approach: Specific tag given to a token, parts of speech(POS).
- For example, verb, noun, pronoun, etc.



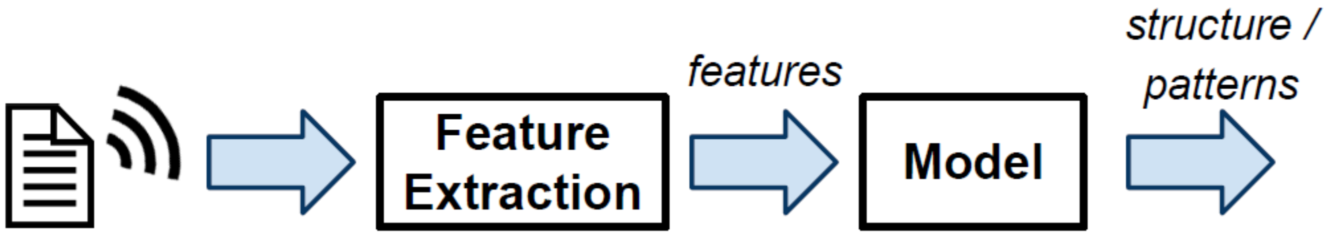
Rule based NLP - Demo

Basic NLP processing	https://colab.research.google.com/drive/1MVBJKFAAJyHjCOmuINC7IrHArNnwC8wz
POS tagging	https://colab.research.google.com/drive/1jQ7kp_RUE0Os0b4HLB7kLEQ6S1nzWVKJ

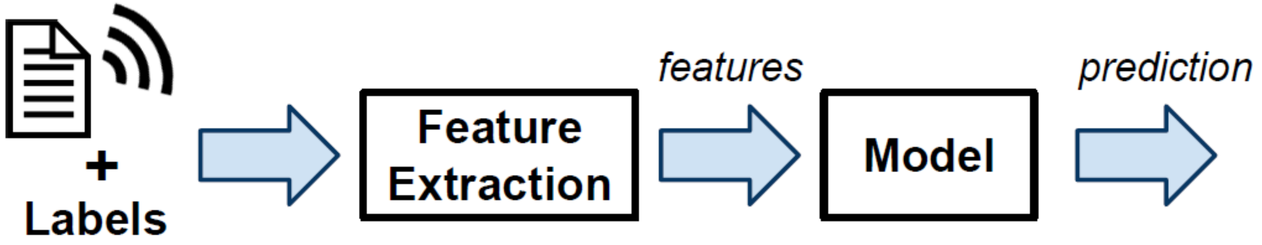
Statistical based Learning

Before Deep Learning

Unsupervised Learning
Learning structure from unlabeled data
(e.g. clustering, topic modelling)

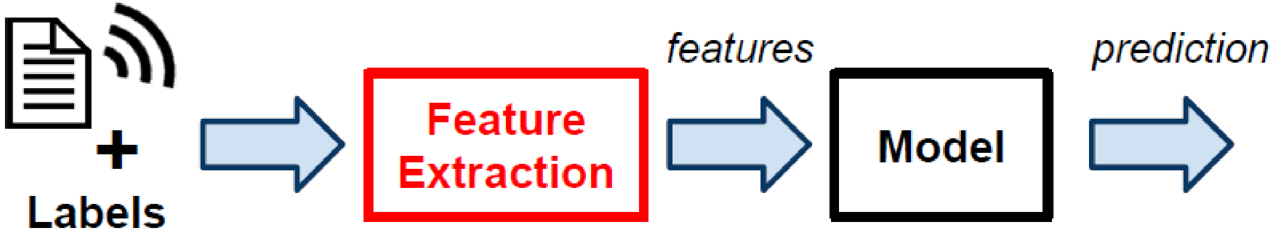


Supervised Learning
Learning to predict from labelled data
(e.g. regression,
classification, generation, etc.)

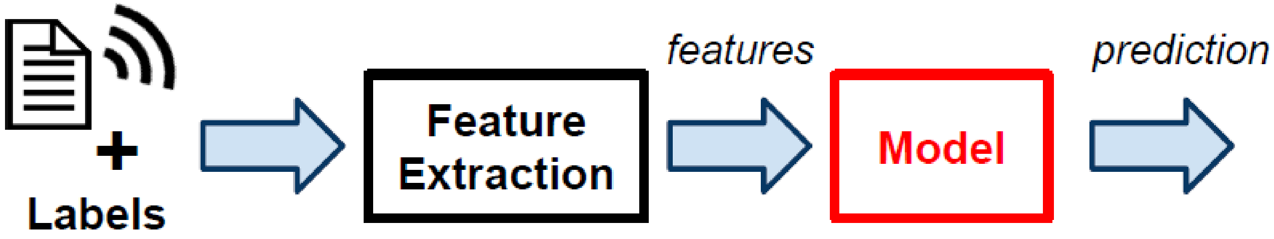


Feature extraction

- Manual features
- Bag of words features
- Tf-idf
- N-grams
- Word embedding



Types of Model



Unsupervised Learning

- Clustering(k-means)
- Topic modeling(LDA – Latent Dirichlet Allocation ,LSA – Latent semantic Analysis)
- Word Embeddings

Supervised Learning

- Decision trees
- Bayesian algorithms (Naive Bayes)
- Regression algorithms(Linear, Logistic)
- Instance based algorithms(KNN, SVM)
- Neural Networks

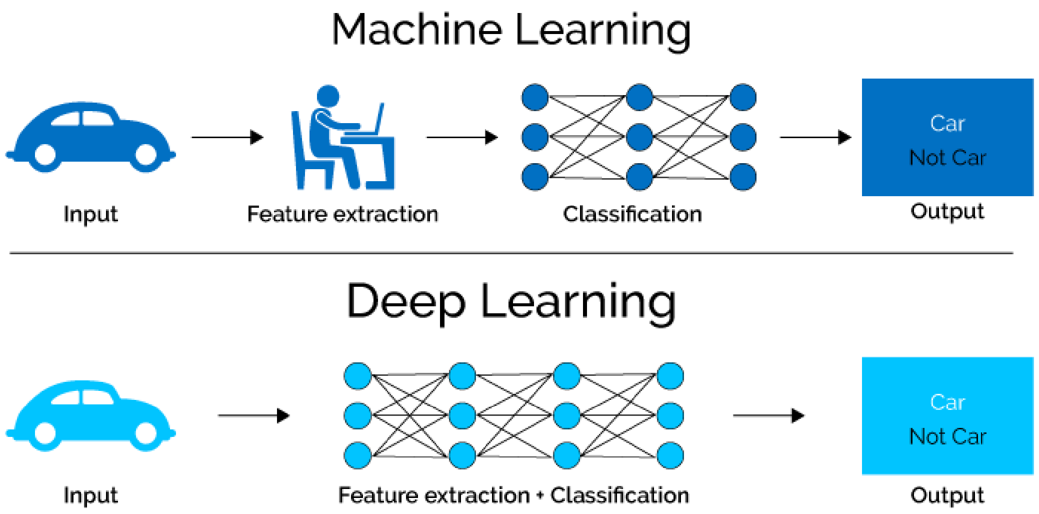
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Statistical based Learning

After Deep Learning

Issues with traditional ML techniques:

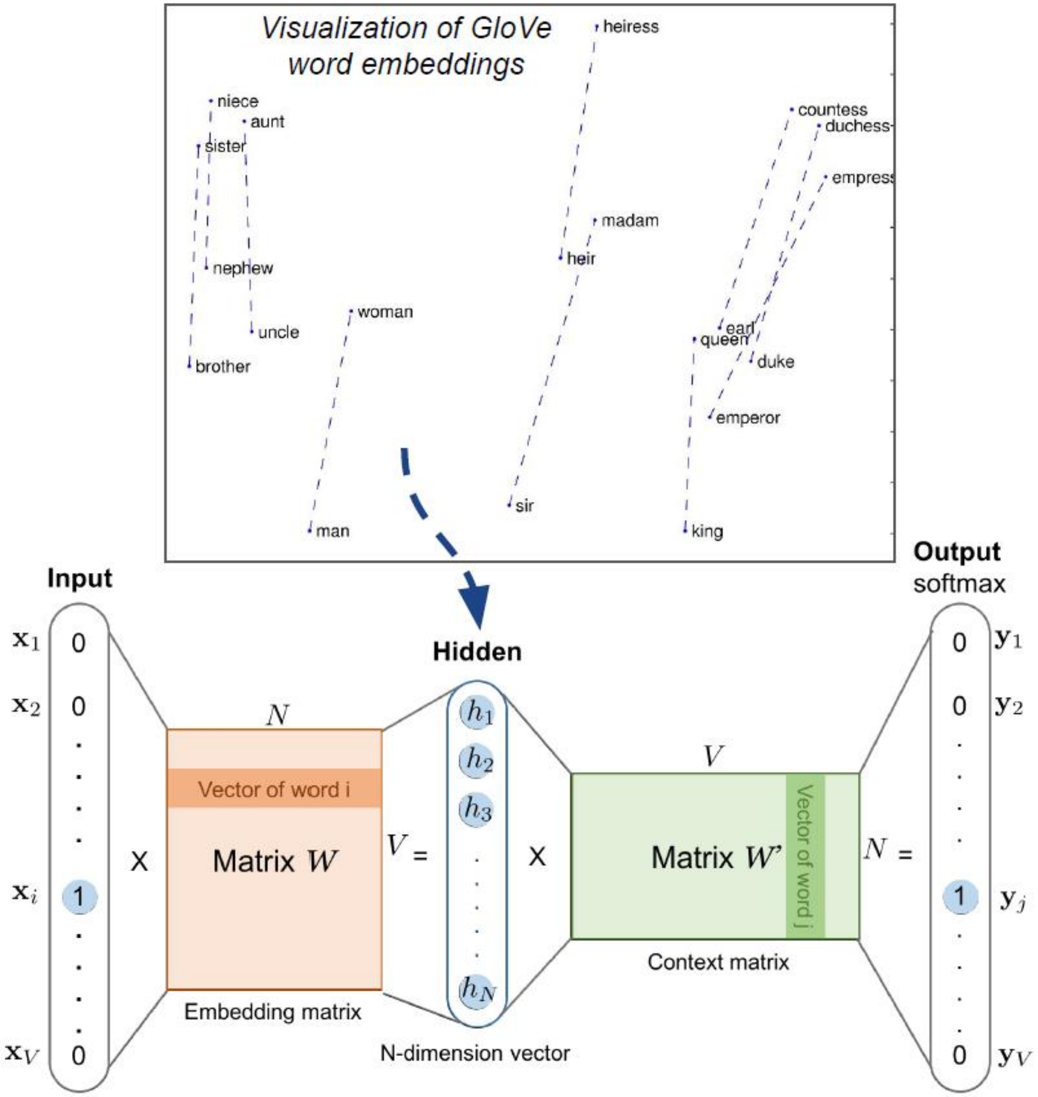
- Hand-engineered features are inefficient.
- Representational capacity of model are limited.
- Long or/ and variable length sequences are challenging.



Learning better features

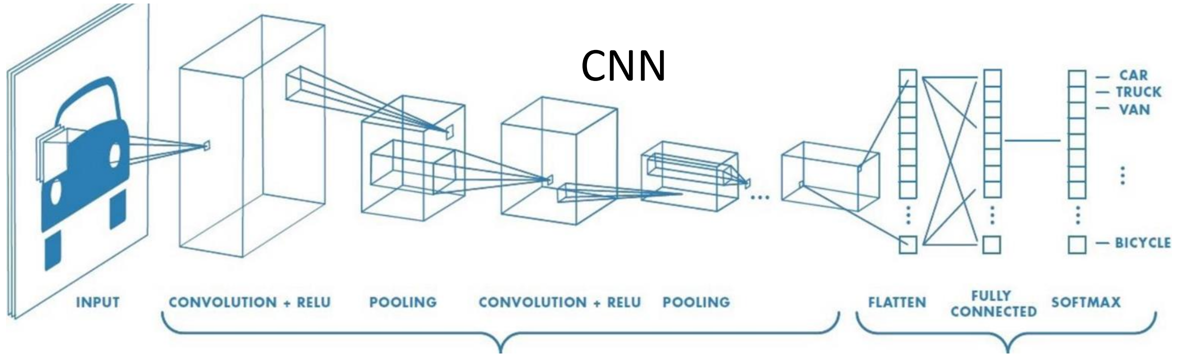
- To solve we have **word embeddings**
- Some of the models and algorithms:
  - Context independent (words):
    - Word2Vec
    - GloVe
    - FastText
  - Context dependent (sentence):
    - ELMo – Embedding from Language Model
    - InferSent

https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html

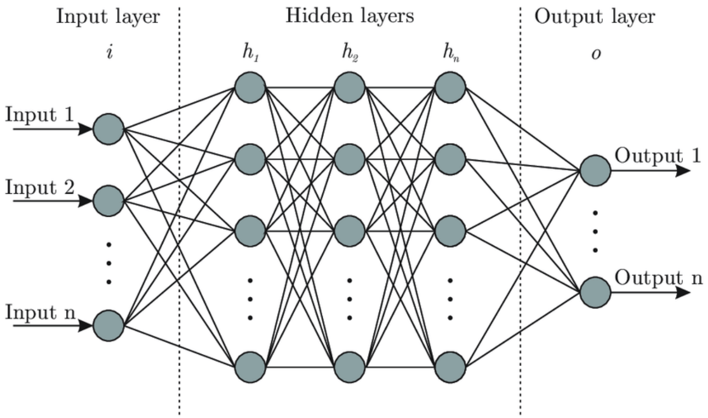


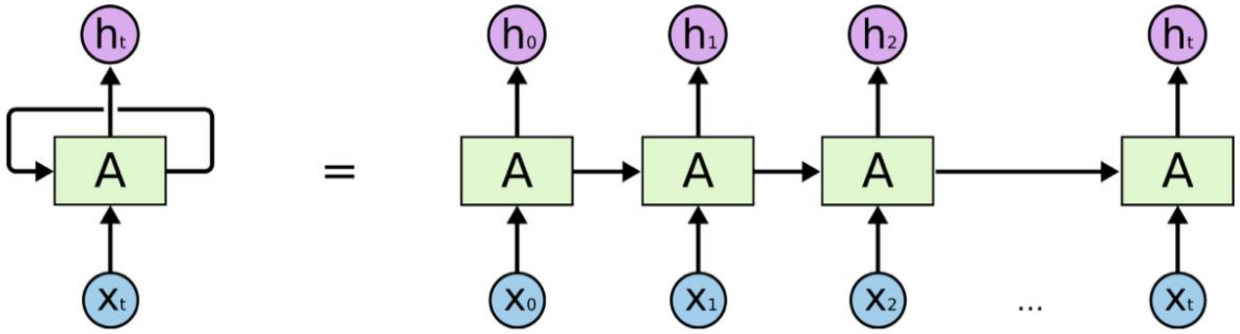
Types of Neural Networks

CNN

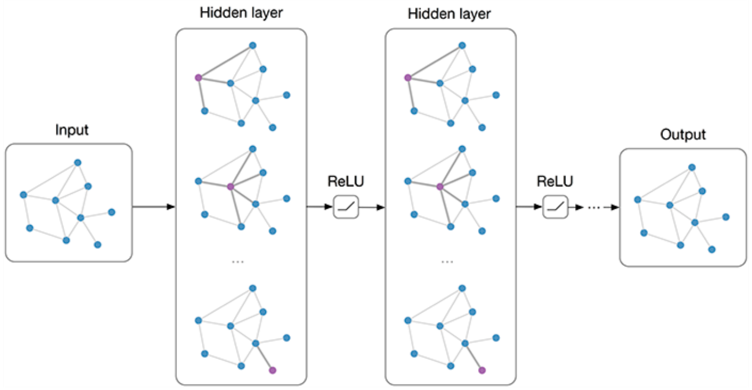


Fully connected NN





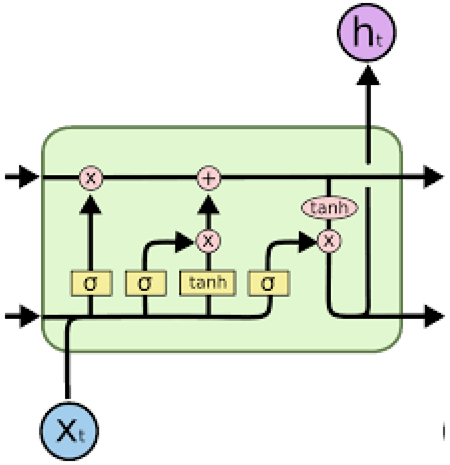
RNN



GNN

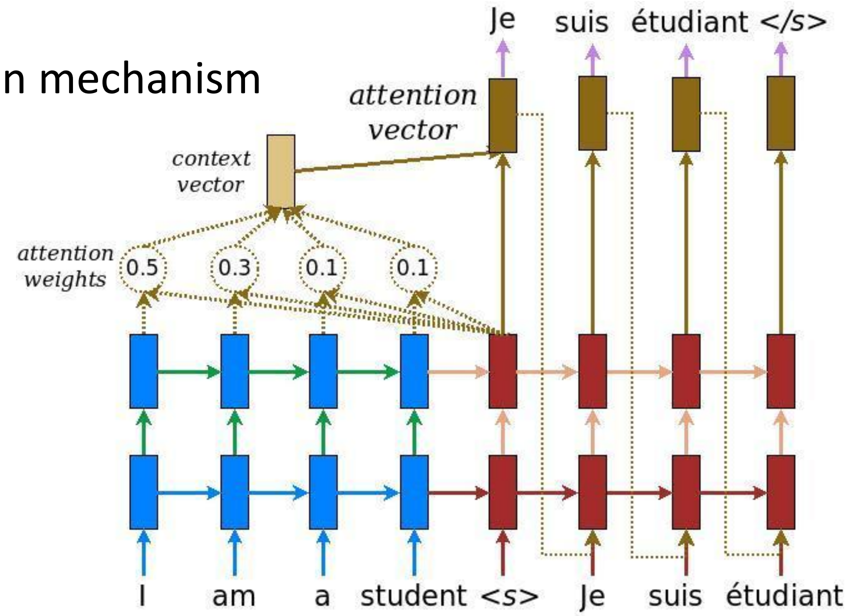
RNN Improvements

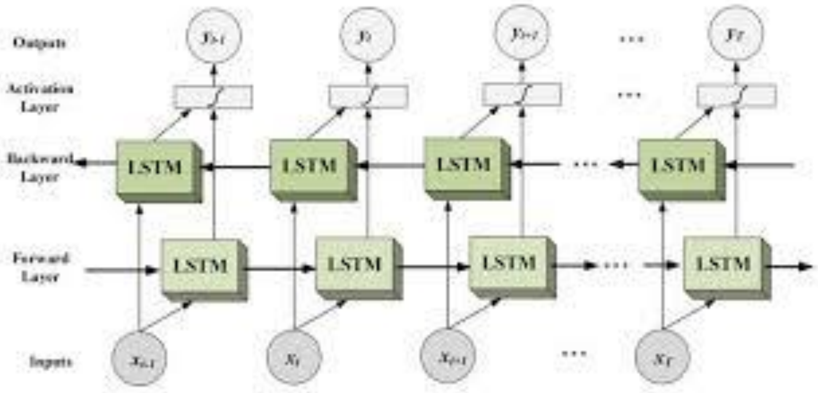
- LSTM – Long Short Term Memory units
- Bidirectional RNN
- Attention mechanism



LSTM unit

Attention mechanism





Bi-directional RNN

Current State of the Art NLP Model

- LSTMs are difficult to parallelize and have challenges for longer sequences.

- **Transformer networks** use only attention mechanisms, with some positional encoding information.

- however, they are often huge models with lots of weights.

http://jalammar.github.io/illustrated-transformer/

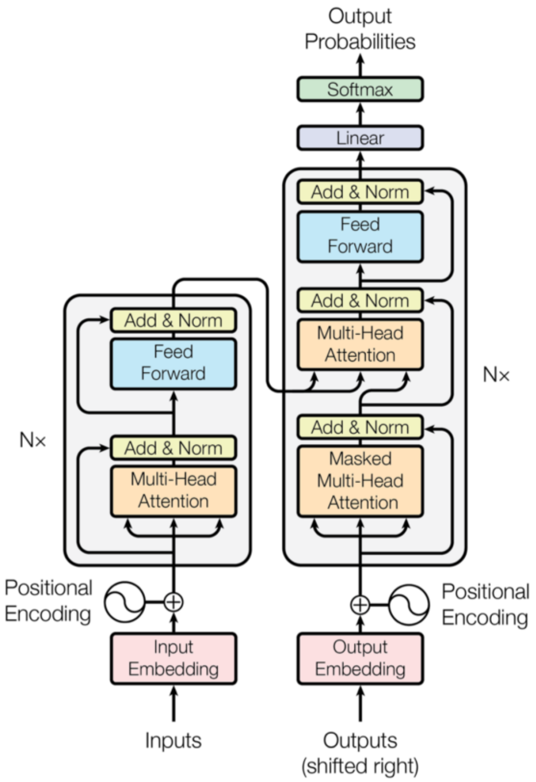


Figure 1: The Transformer - model architecture.

https://arxiv.org/abs/1706.03762

Statistical based NLP - Demo

Manual features	https://colab.research.google.com/drive/1LDqNmRgaiPRyUL-E82Wz31R9tRP35-BV
Transformers Networks	https://colab.research.google.com/drive/1RFJaAa0AWVLg8cLj23J5bigKjv-O9mBI#scrollTo=AqwORzIUX_OI
Text classification using Bag of words model	https://colab.research.google.com/drive/16uf3VAuzUIcbhpSmgINjx7wL8YdvbdNn

Applications

Group 1

- Tokenization
- Stemming
- Lemmatization
- POS tagging
- Query Expansion
- Parsing
- Topic segmentation/recognition
- Morphological Segmentation(word / sentence)

Group 2

- Information retrieval/ extraction
- Relationship extraction
- Named Entity recognition
- Sentiment analysis / sentence boundary disambiguation
- Word sense / disambiguation
- Text similarity
- Coreference resolution
- Discourse analysis

Group 3

- Machine translation
- Automatic summarization and paraphrasing
- Natural language generation
- Reasoning over knowledge based
- Question answer system
- Dialog system
- Image captioning and other multimodal tasks

https://www.datasciencecentral.com/profiles/blogs/overview-of-artificial-intelligence-and-role-of-natural-language

Deep learning algorithms & NLP

Deep Learning Algorithms	NLP Usage
Neural Network	Parts of speech tagging Tokenization Named entity recognition Intent extraction
Recurrent Neural Network	Machine translation Question answering system Image captioning
Recursive Neural Network	Parsing sentences Sentiment analysis Relation classification
Convolutional Neural Network	Spam detection Relation extraction and classification Categorization of search queries

NLP tools and resources

NLP tools:

- NLTK
- Spacy
- Pattern
- Stanza
- Gensim
- TextBlob

Machine learning packages:

- Pytorch
- Scikit-Learn
- Tensorflow

Machine learning packages for NLP:

- Transformers
- Allen NLP

Links:

https://medium.com/@phylypo/a-survey-of-the-state-of-the-art-language-models-up-to-early-2020-aba824302c6

https://github.com/sebastianruder/NLP-progress

https://paperswithcode.com/area/natural-language-processing

https://github.com/sea-bass/intro-nlp
(Programs taken from this repository)

Thanks for watching😊