Natural Language Processing (NLP)

22 October 2022

Krithiga Ramadass

Introduction

Krithiga Ramadass, Chennai
Overall experience of 12 years.
8 years in ML and DS.

TOYOTA
CONNECTED INDIA

We are a team of data scientists, engineers, and designers who share the vision of **transforming Toyota from an automotive giant to a mobility company** with cutting-edge technology.
Toyota Connected is enabling improved safety and convenience with a cloud-based **digital connected mobility intelligence platform**. We are leveraging **vehicle data** and **artificial intelligence** to change the way people interact with vehicles.

Lead ML Engineer
Natural Language Processing
Conversational AI

What's in it?

What's for the lecture

Introduction to Natural Language Processing
- Beginner friendly
- Basic NLP Tools & Techniques
- NLP Applications overview

What are we seeing today

Core Concept
Basic NLP tasks
NLP Tools

What's for tomorrow

How do we approach NLP problem
NLP Applications
What do we do in Toyota Connected India (TCIN)

Natural Language Processing (NLP)

- Computer's ability to understand text and spoken words in much the same way human beings can

- Enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment

Source: IBM

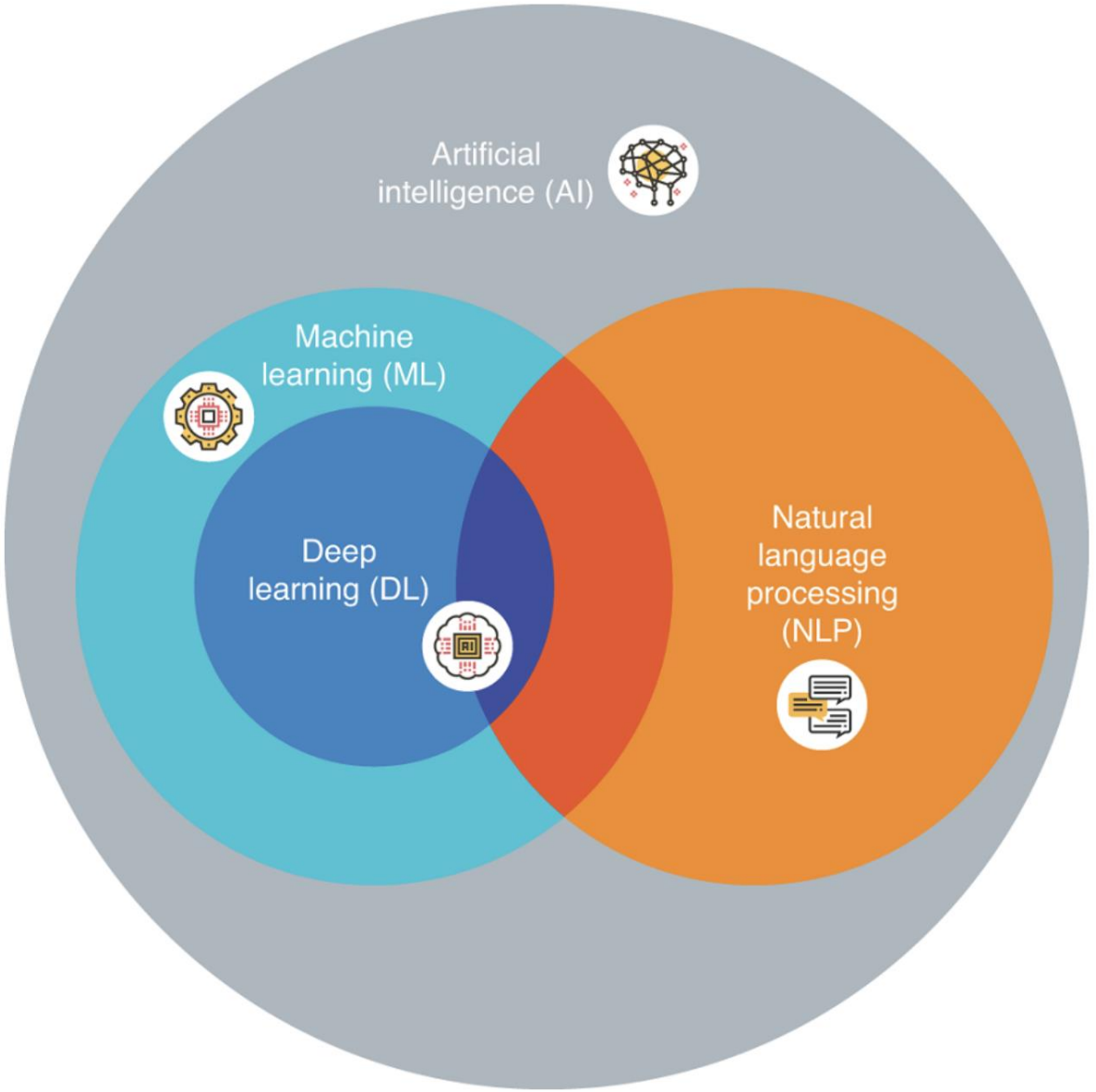
Natural Language Processing (NLP)

Natural language processing (**NLP**) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Source: Wikipedia

What is not NLP?

Natural Language Processing (NLP)

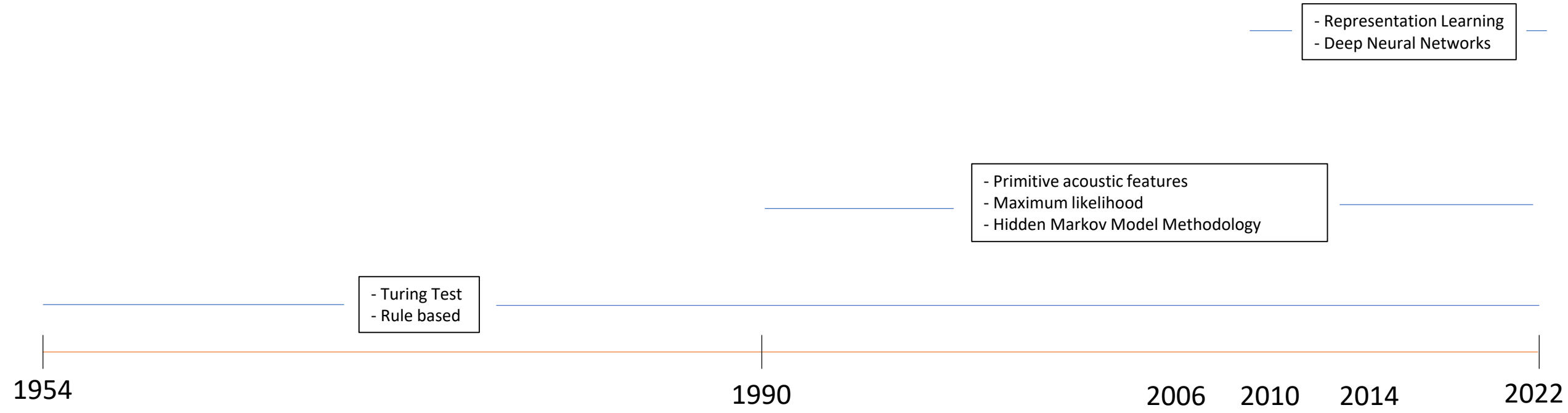


Source: O'Reilly

List down all the applications that you could think of in NLP?

https://www.menti.com/alj8bo747g2t

NLP Evolution



Why NLP?

*You should read this **book**; it's a great novel!*

*You should **book** the flights as soon as possible.*

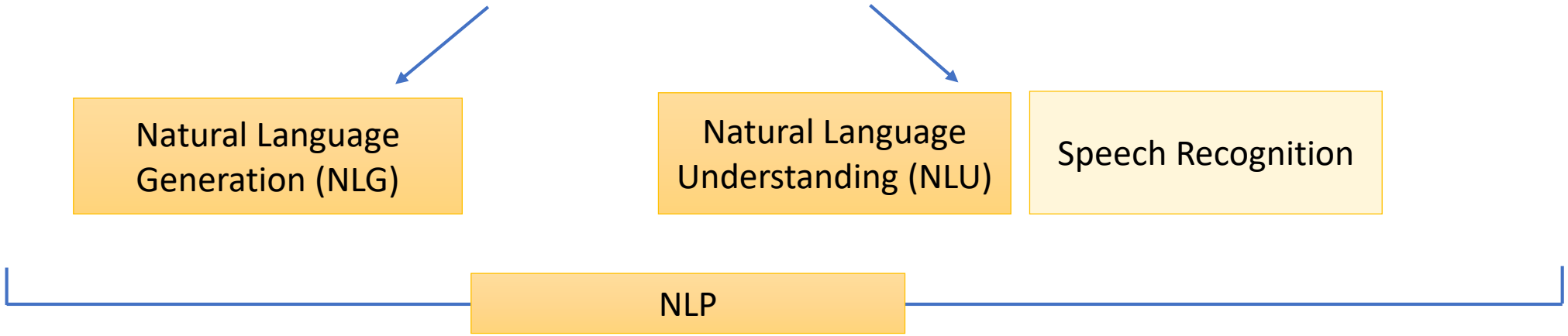
*You should close the **books** by the end of the year.*

*You should do everything by the **book** to avoid potential complications.*

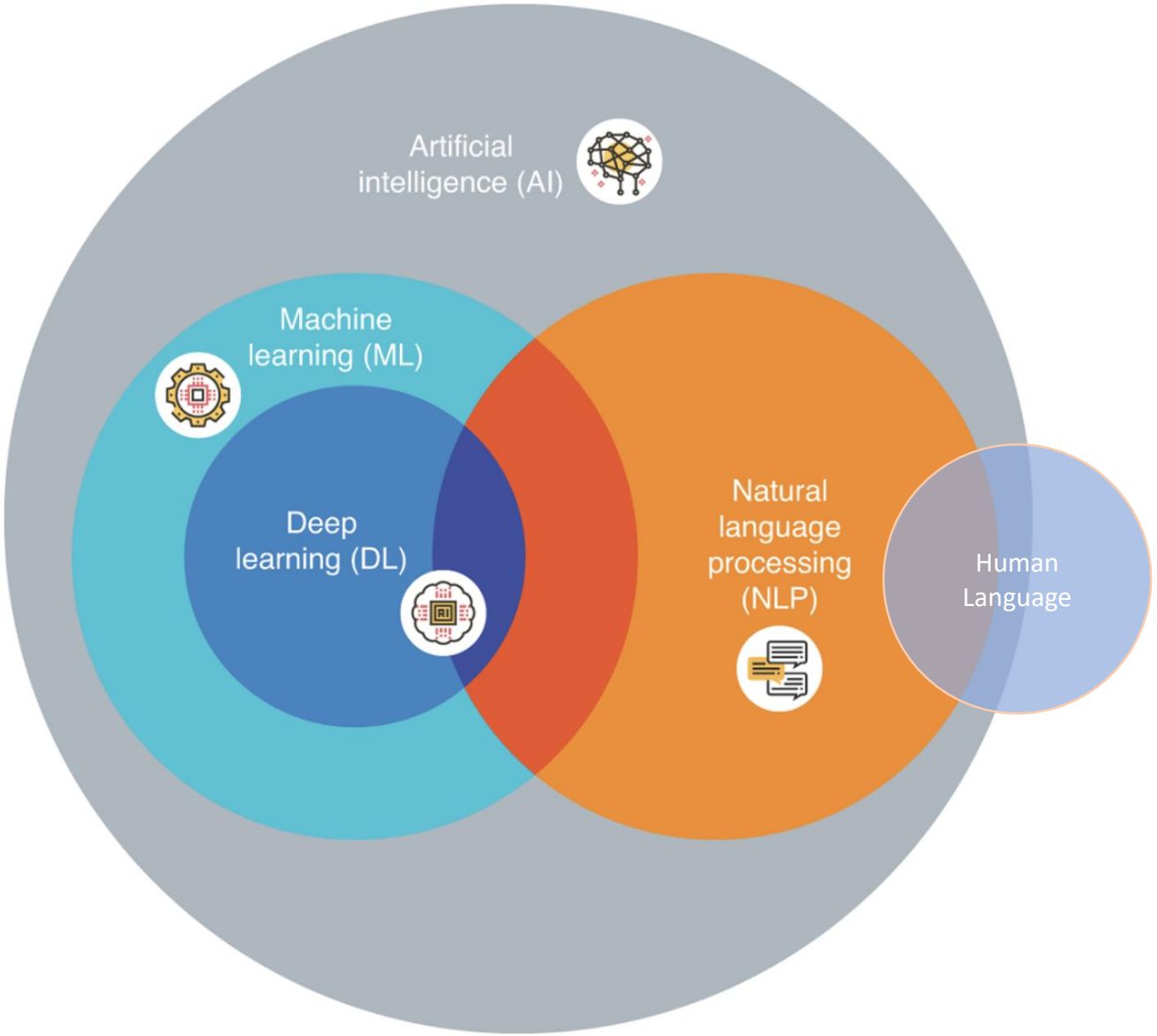
Human Language

- When we study human language, we are approaching what some might call the “human essence” the distinctive qualities of mind that are so far, unique to humans
- A communication tool
- What does it mean to know a language?

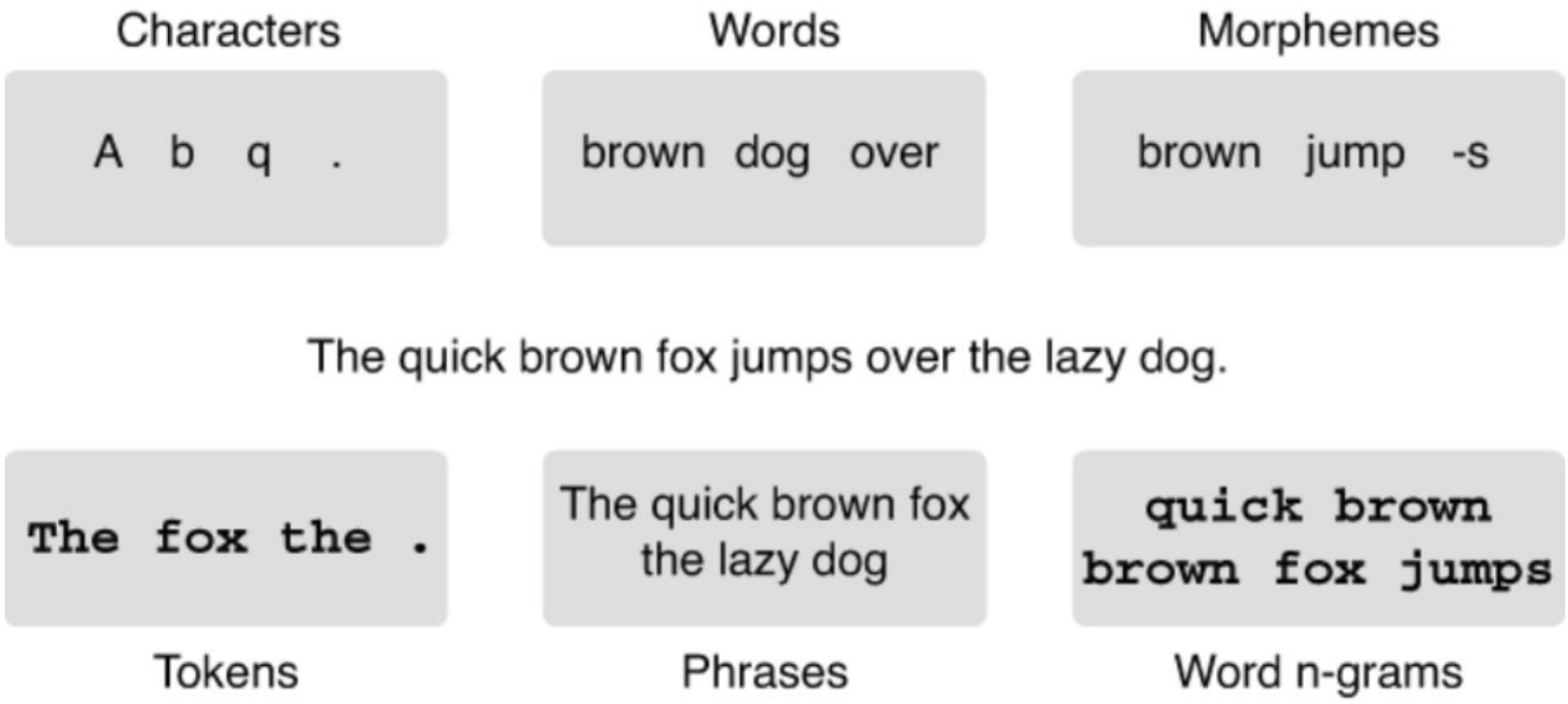
  - To be able to speak and be understood



Natural Language Processing (NLP)

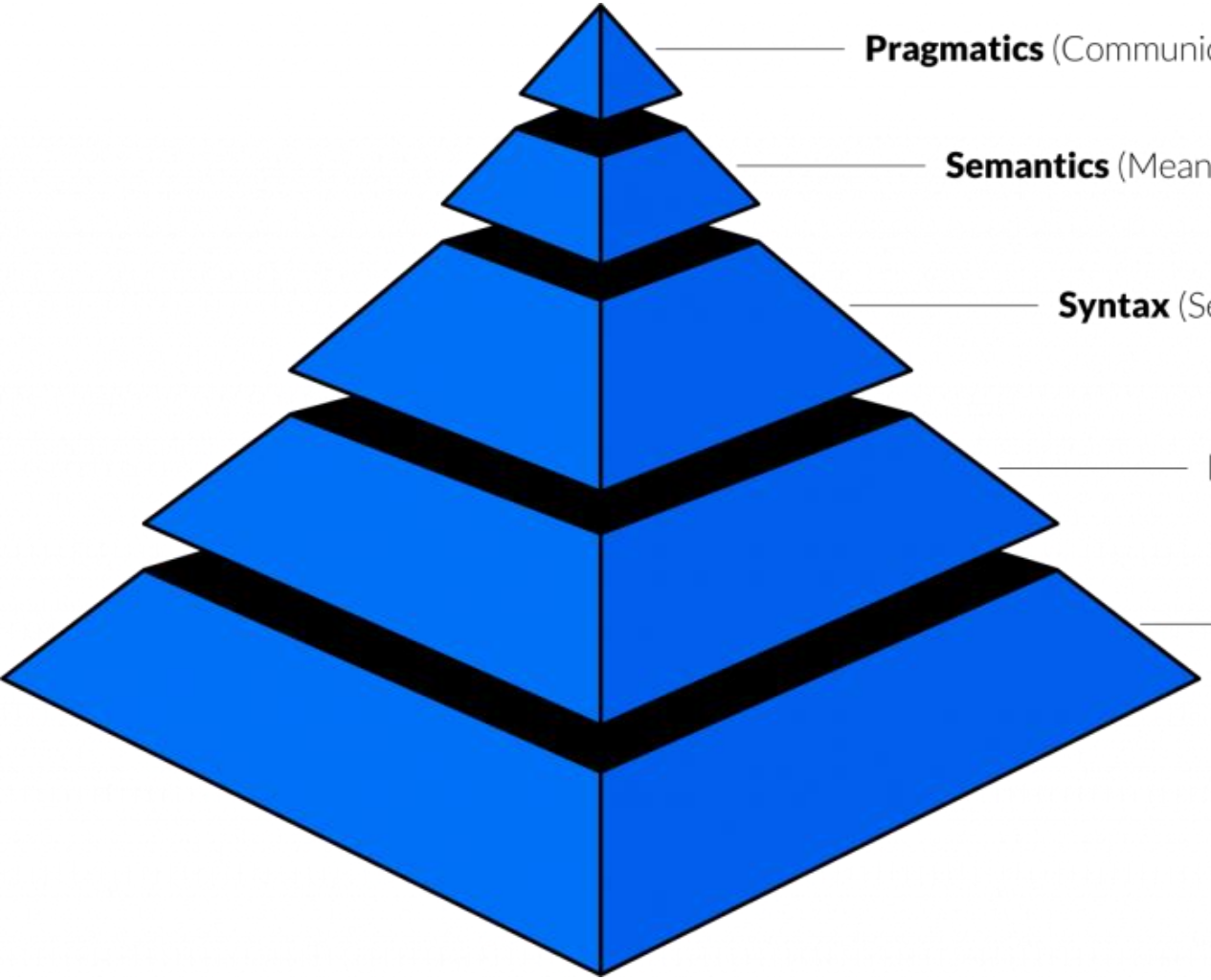


Building blocks of language & NLP



Source: O'Reilly

Natural Language Understanding Pyramid



NLP Applications

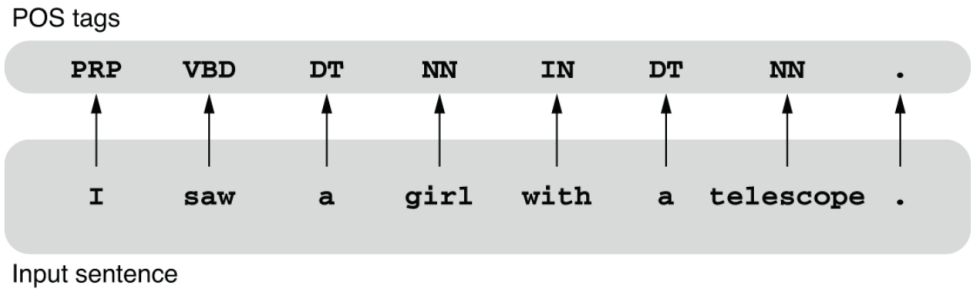
Word Sense Disambiguation

Parsing

Singular, plural
Gender
Prefix suffix

Source: Nextiva

PoS Tagging



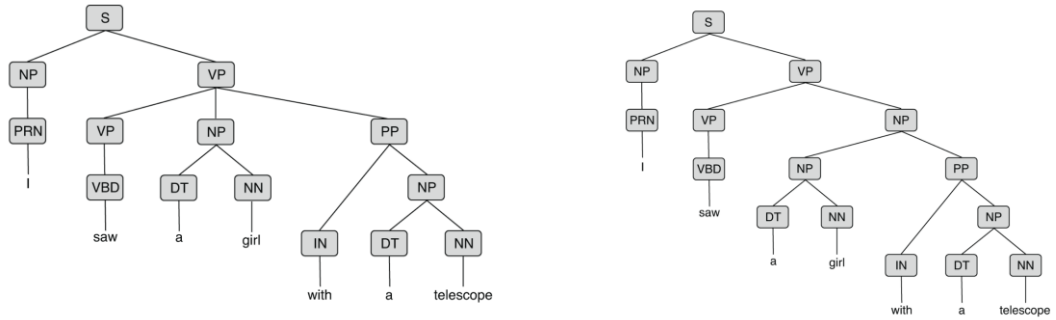
POS tag	Description
DT	Determiner
IN	Preposition
NN	Noun (singular or mass)
PRP	Pronoun
VBD	Verb (past tense)

sentence:	The	oboist	Heinz	Holliger	has	taken	a	hard	line	about	the	problems	.
original:	DT	NN	NNP	NNP	VBZ	VBN	DT	JJ	NN	IN	DT	NNS	.
universal:	DET	NOUN	NOUN	NOUN	VERB	VERB	DET	ADJ	NOUN	ADP	DET	NOUN	.

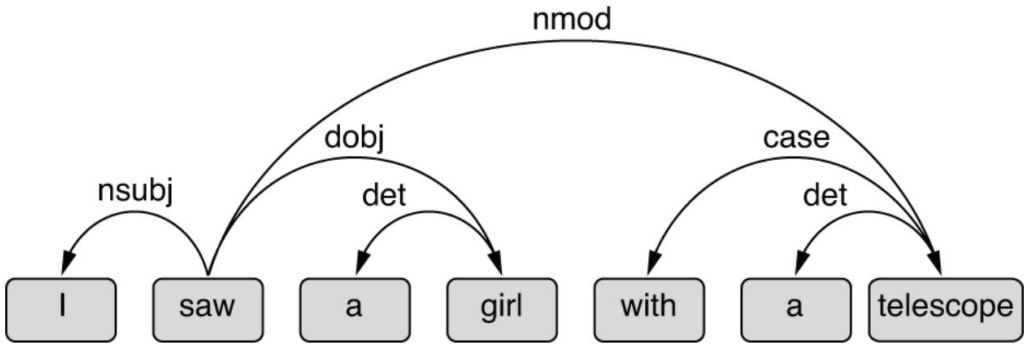
Source: O'Reilly

Parsing

- Dependency Parsing



- Constituency Parsing



Production Rule	Meaning
$S \rightarrow NP \quad VP$	A sentence is a noun phrase followed by a verb phrase
$NP \rightarrow DT \quad NN$	A noun phrase is a determiner followed by a noun
$NP \rightarrow NP \quad PP$	A noun phrase can also be another noun phrase followed by a prepositional phrase
$VP \rightarrow Vi$	A verb phrase is simply an intransitive verb
$VP \rightarrow Vt \quad NP$	A verb phrase is a transitive verb followed by a noun phrase
$VP \rightarrow VP \quad PP$	A verb phrase can also be another verb phrase followed by a prepositional phrase
$PP \rightarrow IN \quad NP$	A prepositional phrase is a preposition followed by a noun phrase
$DT \rightarrow$ the	*The* is a determiner
$NN \rightarrow$ dog | cat | tree	*Dog*, *cat*, and *tree* are nouns
$IN \rightarrow$ on	*On* is a preposition
$Vt \rightarrow$ chased | perched	Both *chased* and *perched* are transitive verbs
$Vi \rightarrow$ sneeze	*Sneeze* is an intransitive verb

Source: O'Reilly

Tokenization

- Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* , perhaps at the same time throwing away certain characters, such as punctuation.

Input: Friends, Romans, Countrymen, lend me your ears;
Output: Friends | Romans | Countrymen | lend | me | your | ears

  - Word Tokenization
  - Sentence Tokenization
  - White space
  - Punctuation based tokenization
  - Treebank tokenizer
  - Tweet tokenizer
  - Multi-word tokenizer
- Limitations:
  - Doesn’t support all the languages

For *O'Neill*, which of the following is the desired tokenization?

neill
oneill
o’neill
o’ | neill
o | neill ?

And for *aren't*, is it:

aren’t
arent
are | n’t
aren | t ?

Source: Stanford NLP

Stemming and Lemmatization

Stemming just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling.

Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.

Sometimes, the same word can have multiple different Lemmas. We should identify the Part of Speech (POS) tag for the word in that specific context.

1. lemmatize(‘**walking**’) -> ‘**walk**’.
   stem(‘**walking**’) -> ‘**walk**’.
2. Verb lemmatize(‘**Stripes**’) -> ‘**Strip**’.
   Noun lemmatize(‘**Stripes**’) -> ‘**Stripe**’.
3. Stem('**Caring**’) -> '**Car**’
   lemmatize('**Caring**’) -> '**Care**’.

Lemmatization is computationally expensive since it involves look-up tables and what not. If you have large dataset and performance is an issue, go with Stemming. Remember you can also add your own rules to Stemming. If accuracy is paramount and dataset isn't humongous, go with Lemmatization.

Source: SO

Stop word removal

The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.
These words have no significance in some of the NLP tasks like information retrieval and classification, which means these words are not very discriminative.

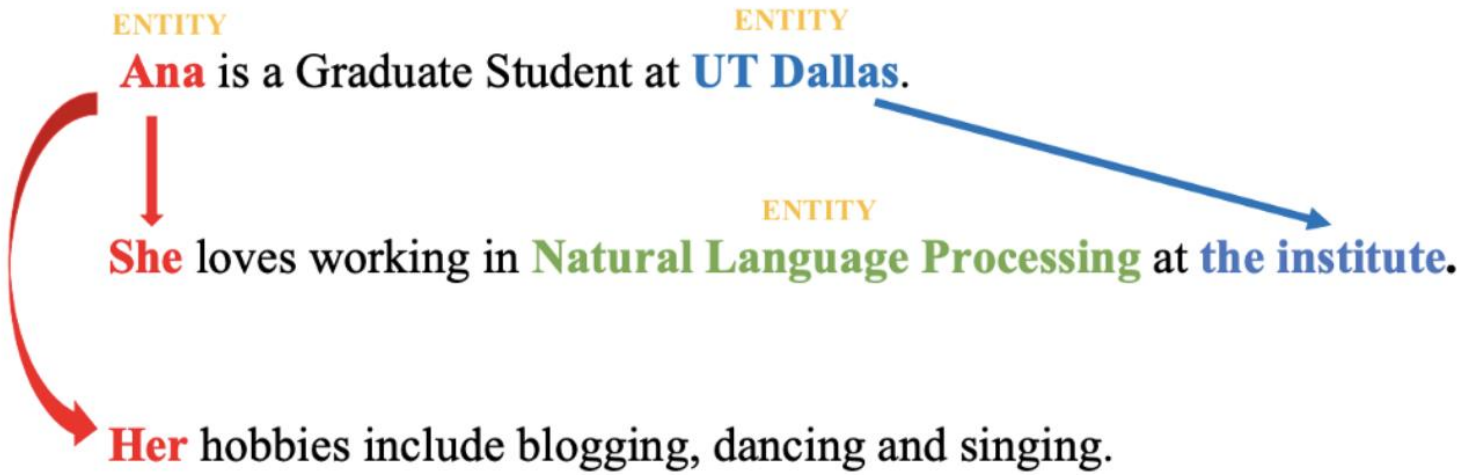
a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

On the contrary, in some NLP applications stop word removal will have very little impact. Most of the time, the stop word list for the given language is a well hand-curated list of words that occur most commonly across corpuses.

Source: O'Reilly

Coreference Resolution

Ana is a Graduate Student at UT Dallas. She loves working in Natural Language Processing at the institute. Her hobbies include blogging, dancing and singing.



Source: https://towardsai.net/p/nlp/c-r

Entity Recognition

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Source:SpotDraft

Vector Representation

Bag of words

- Representation of words
- One-hot Encoding

Sentence 1: I have a dog

Sentence 2: You have a cat

	I	have	a	dog	you	cat
Sentence 1	1	1	1	1	0	0
Sentence 2	0	1	1	0	1	1

Source:PyImageSearch

Document-Term **Matrix**

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

- bird → [5,1,1]
- the → [2,1,2]
- word → [0,0,1]
- ...

Source:PyImageSearch

Disadvantages

- We end up counting the word occurrences. Some words appears in a document more than the other
- Not normalized

TF-IDF

A **tf-idf score** is a decimal number that measures the importance of a word in any document. It gives small values to frequent words in all the documents and more weight to those more scarce across the corpus.

TF – Term Frequency - the number of times the word appears in each document.

IDF – Inverse Document Frequency - an inverse count of the number of documents a word appears in. Idf measures how significant a word is in the whole corpus.

$$tf(t,d) = |\text{Number of times term } t \text{ appears in document } d|$$

$$idf(t,D) = \frac{|\text{Number of documents}|}{|\text{ number of documents that contain term } t|}$$

$$tfidf(t,d,D) = tf(t,d).\, idf(t,D)$$

Whereby:

- t is the word or token.
- d is the document.
- D is the set of documents in the corpus.

https://github.com/Shubha23/Text-processing-NLP/blob/master/NLP%20-%20Text%20processing%20pipeline.ipynb

Source: OpenClassRooms

Disadvantages

- Indirectly depends on the word occurrences – Relative to corpus
- Score varies from document to document
- Matrix becomes large and sparse
- Inability to learn:
  - Grammar
  - Semantics

Embeddings

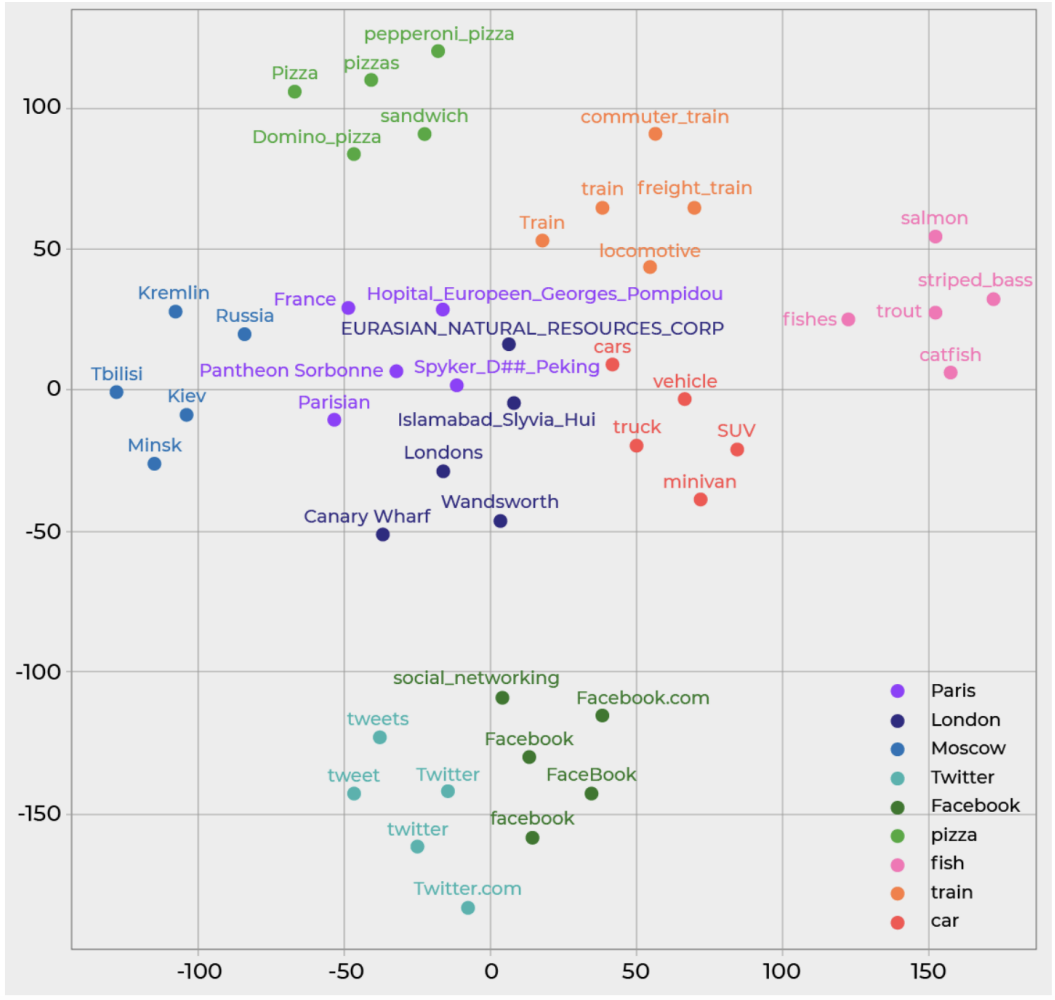
$$\overrightarrow{queen} - \overrightarrow{woman} = \overrightarrow{king} - \overrightarrow{man}$$

$$\overrightarrow{France} - \overrightarrow{Paris} = \overrightarrow{Germany} - \overrightarrow{Berlin}$$

- Word2Vec – Words appearing in similar context
- Glove – Words Cooccurrences in the corpus

- Captures Analogies
- Distance between words
- Dense vectors compared to CV/ TF-IDF
- Constant vector size
- Universal vector Representation
- Can be extended to sentences, paragraphs, documents

http://projector.tensorflow.org/



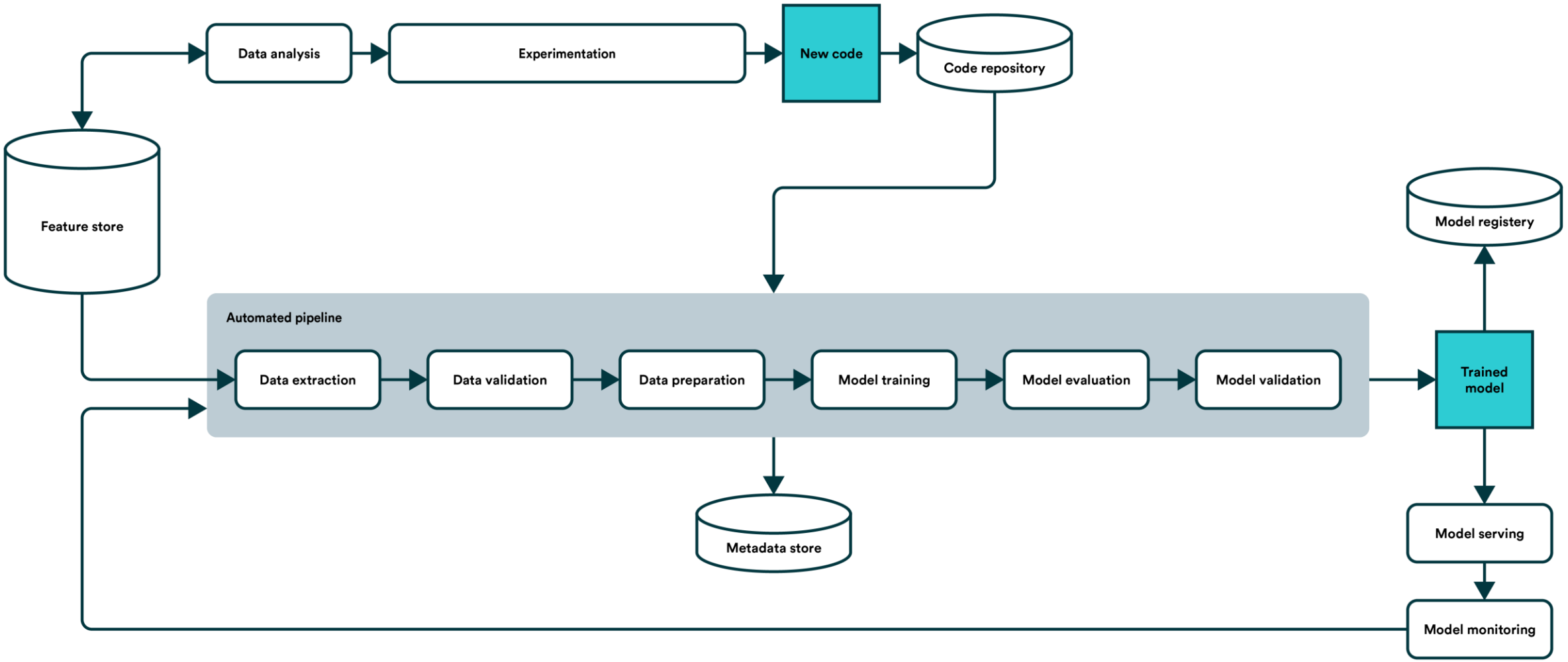
Disadvantages

- Cultural Bias
- Out of Vocabulary words

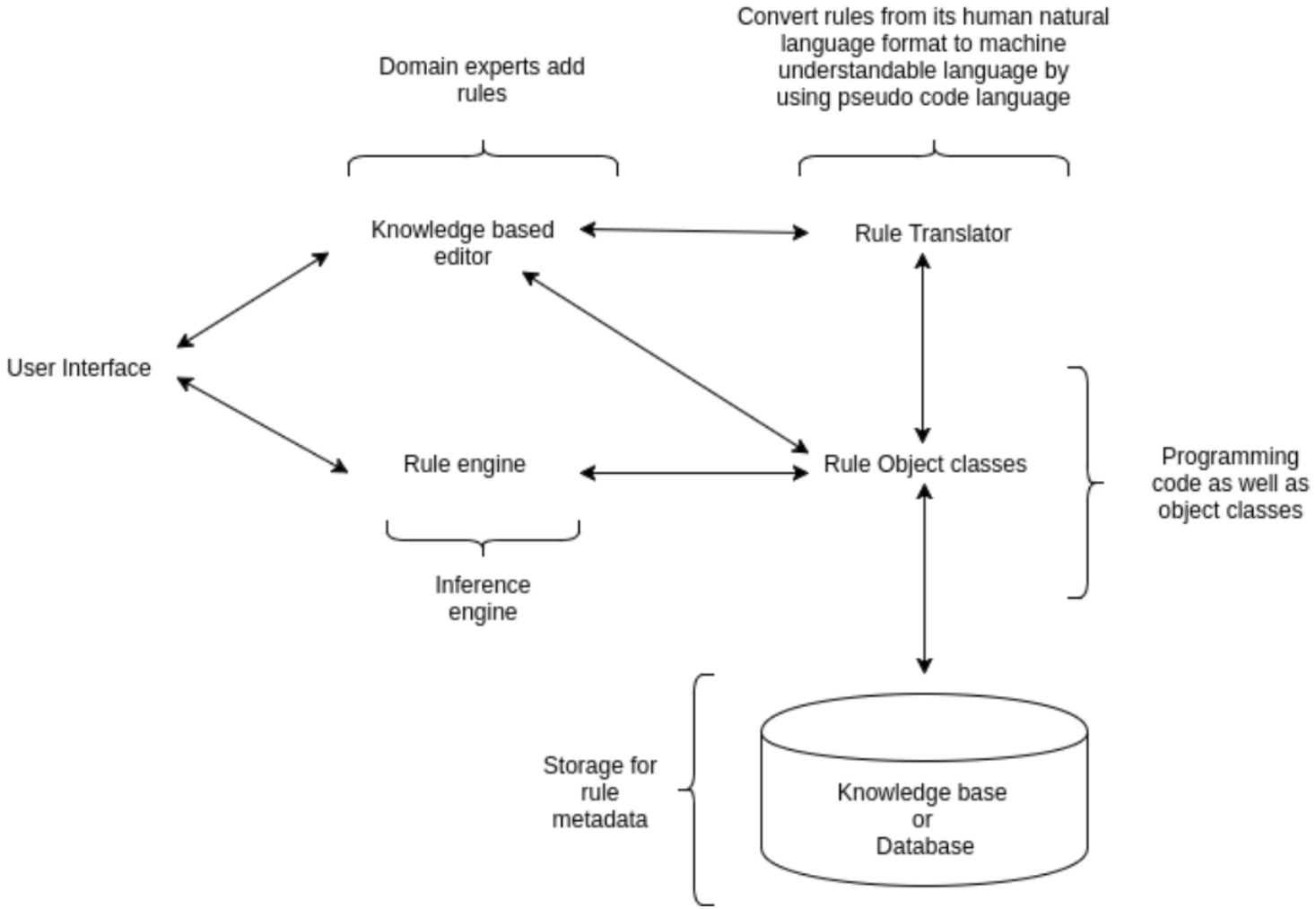
How do we approach an NLP problem

Machine Learning Workflow

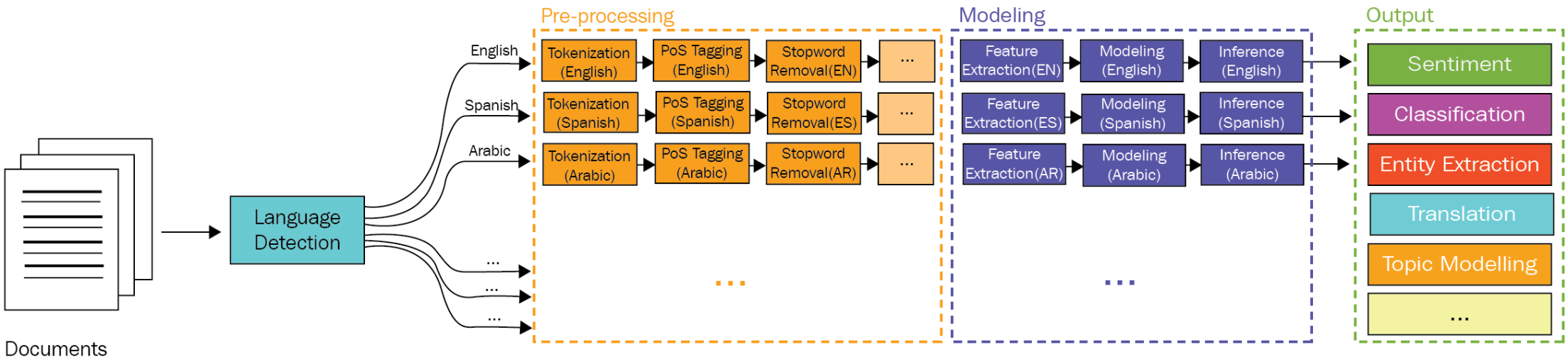
Machine Learning Pipeline Automated



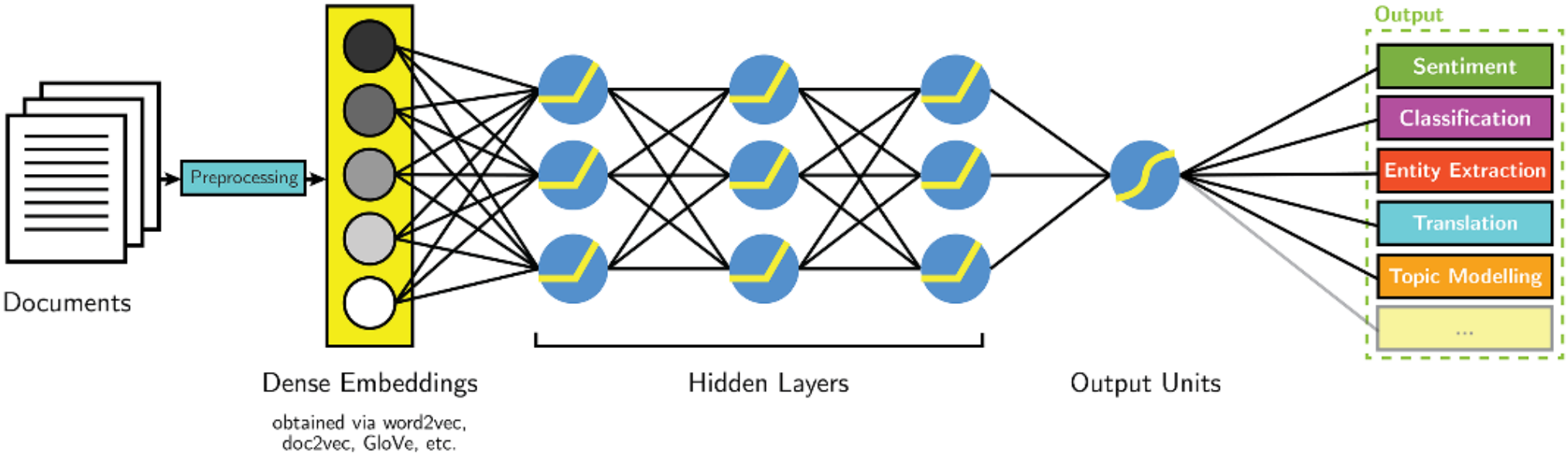
Rule based NLP



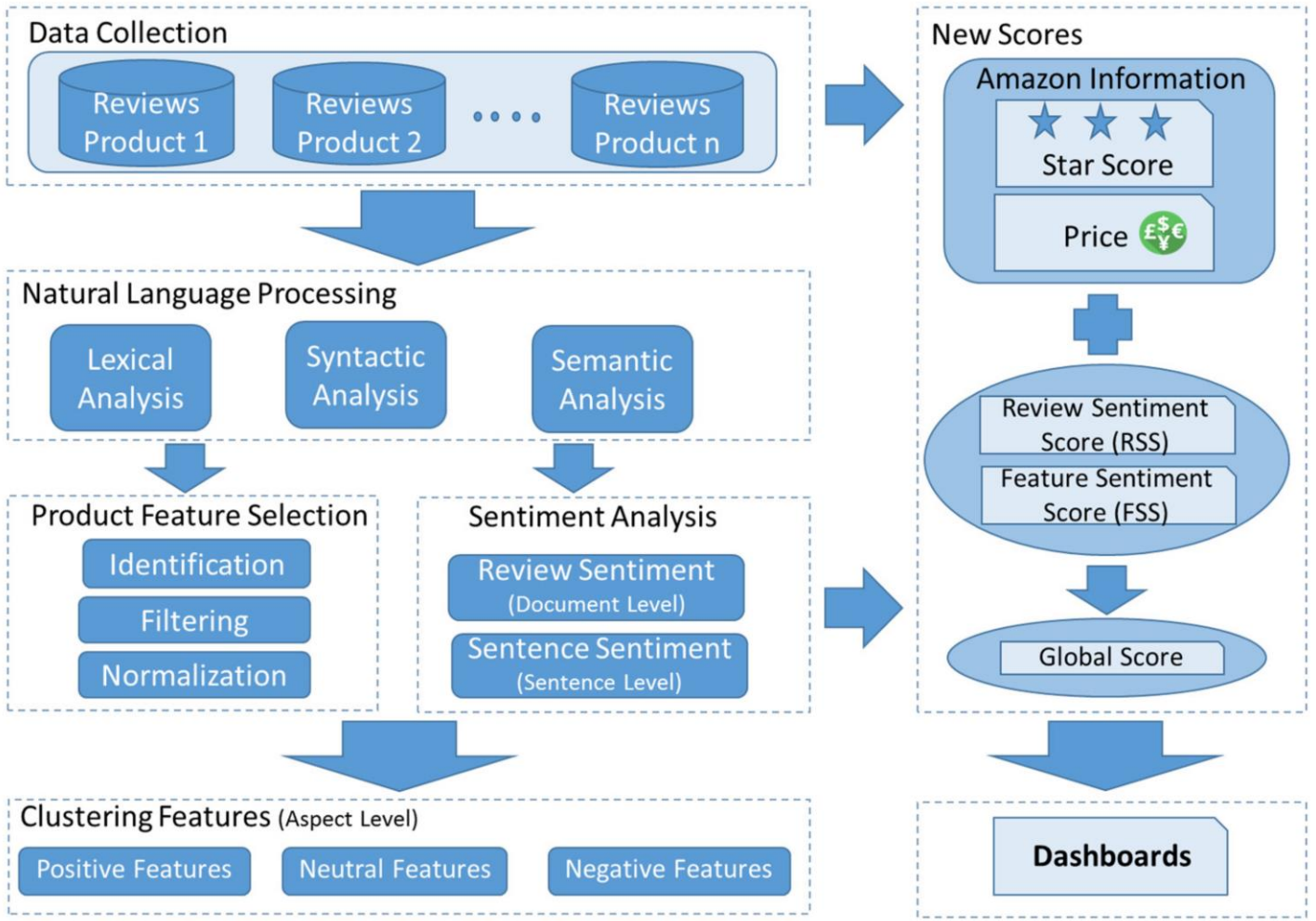
Classical NLP



Deep Learning-based NLP

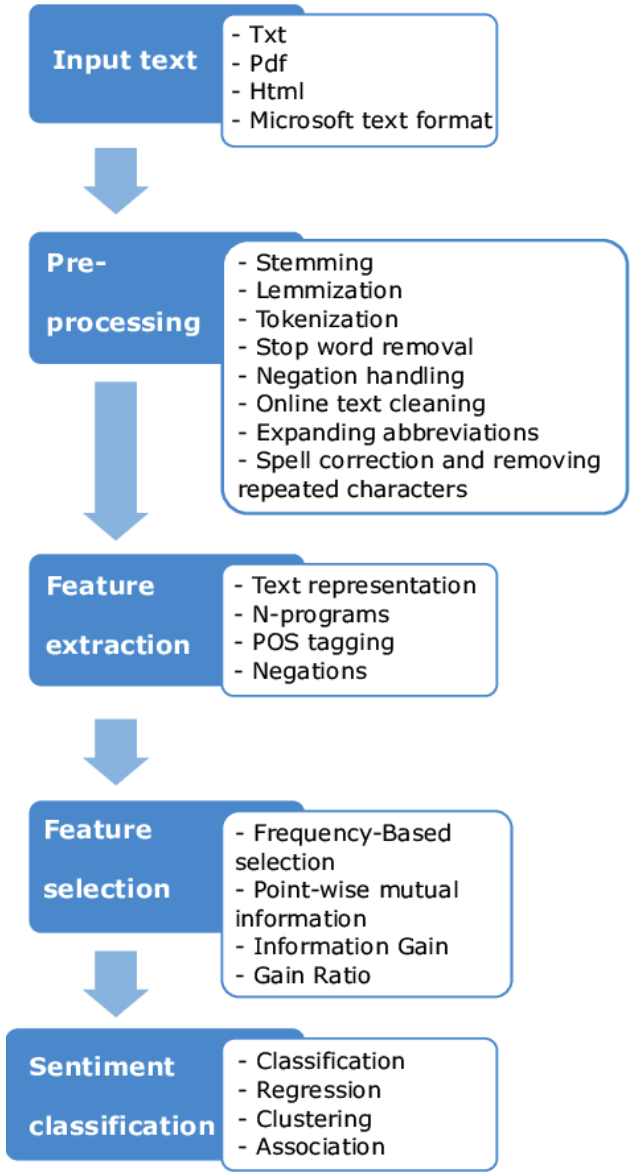


Sentiment Analysis

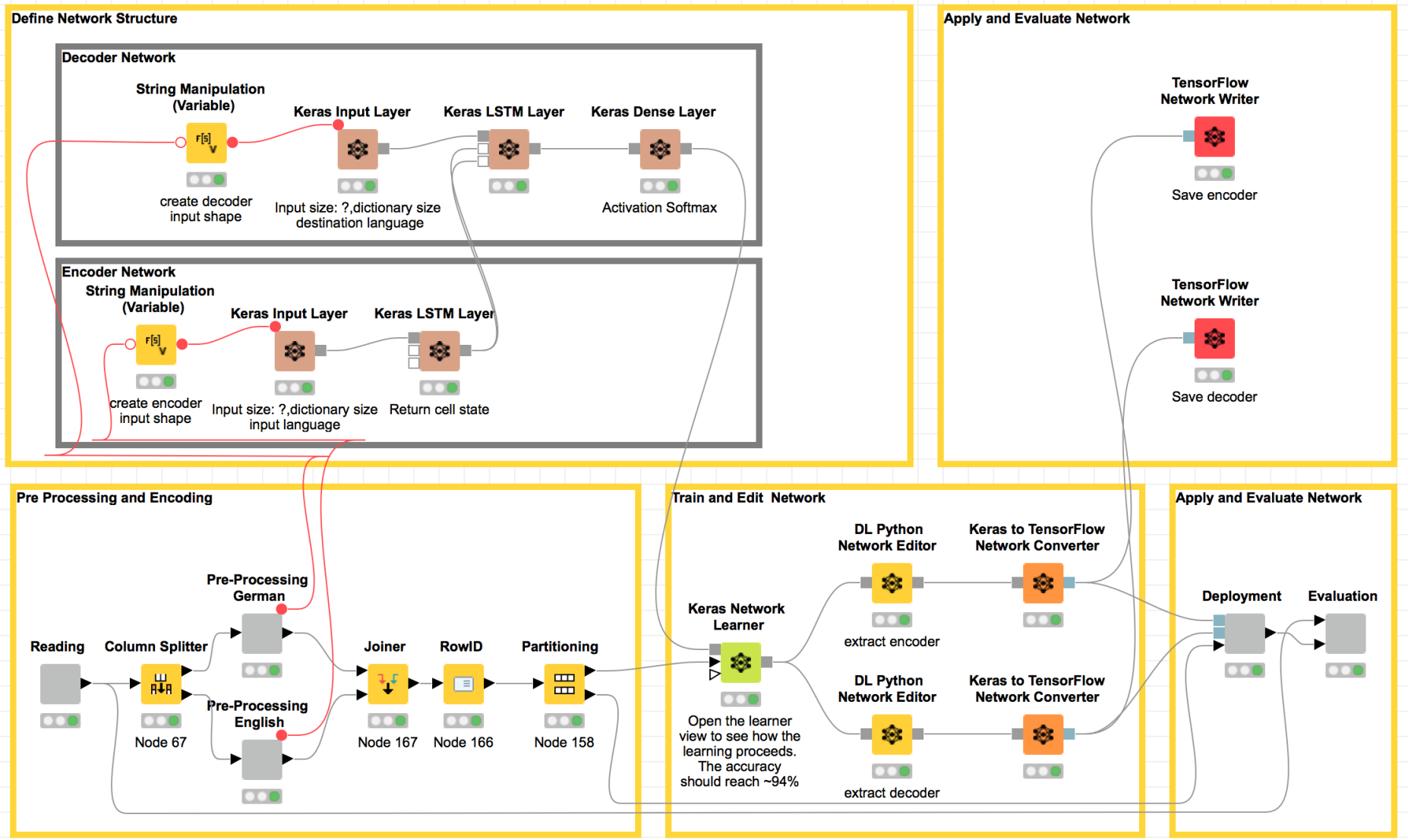


https://pair-code.github.io/lit/tutorials/sentiment/

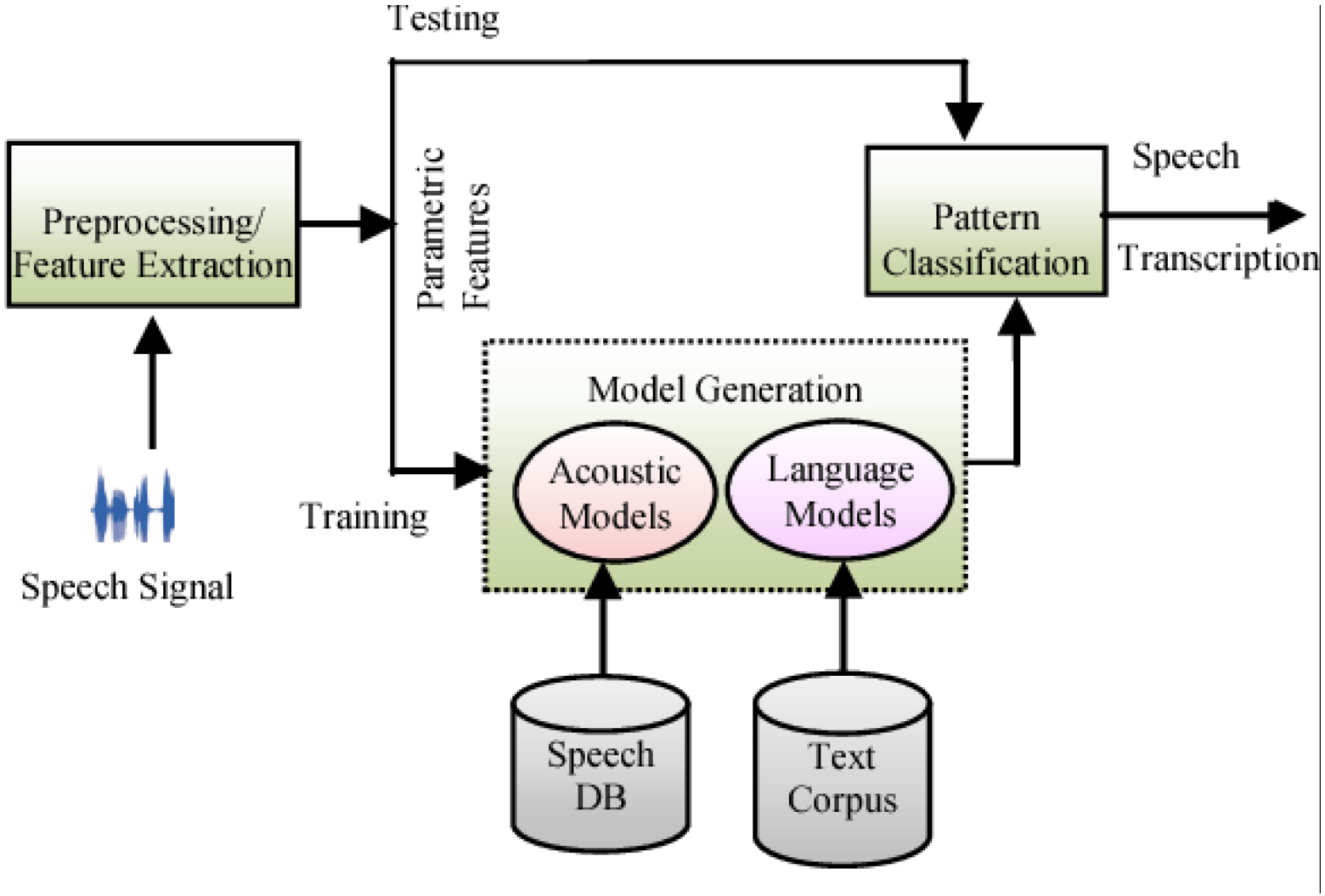
Sentiment Analysis - NLP



Machine Translation

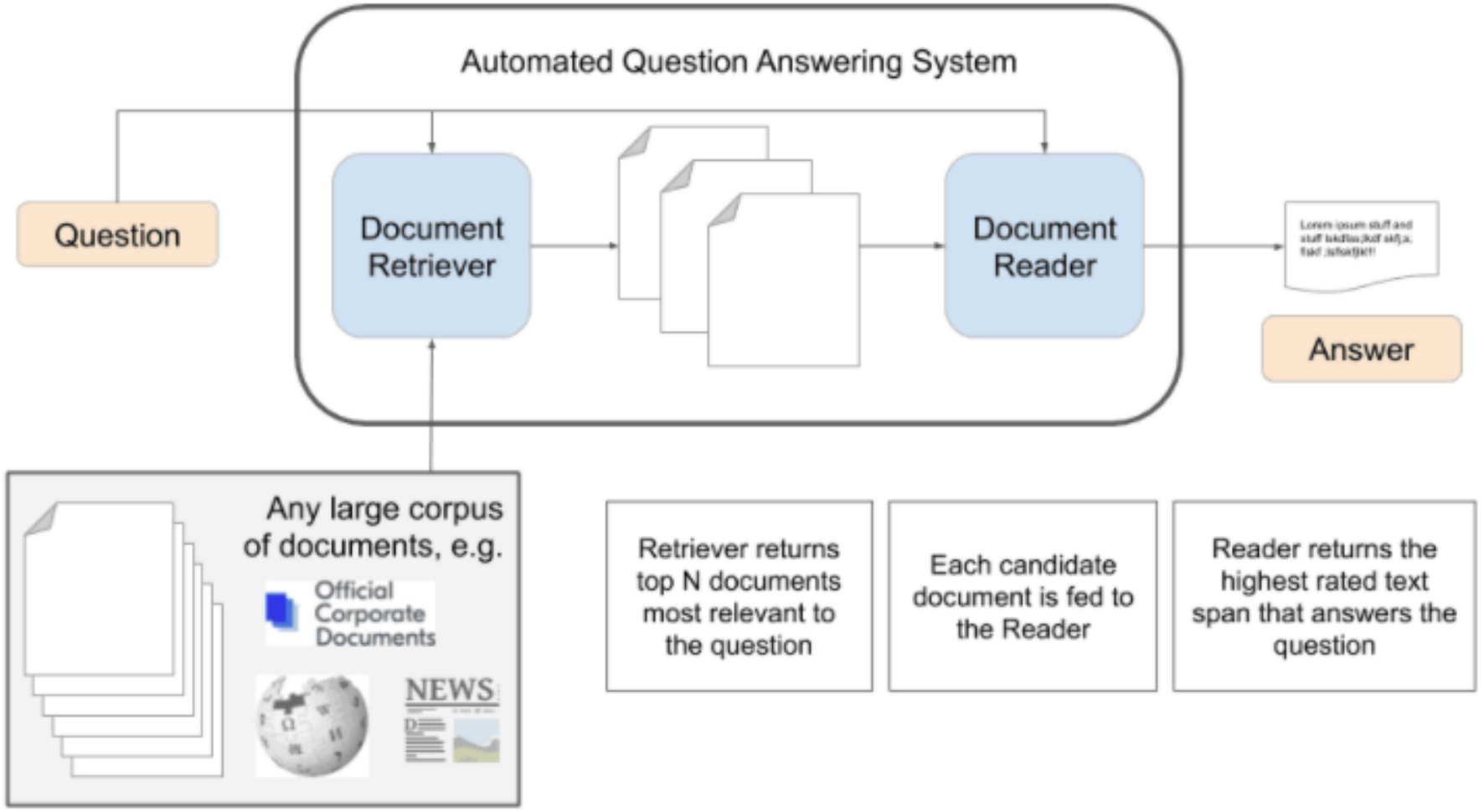


Speech Recognition

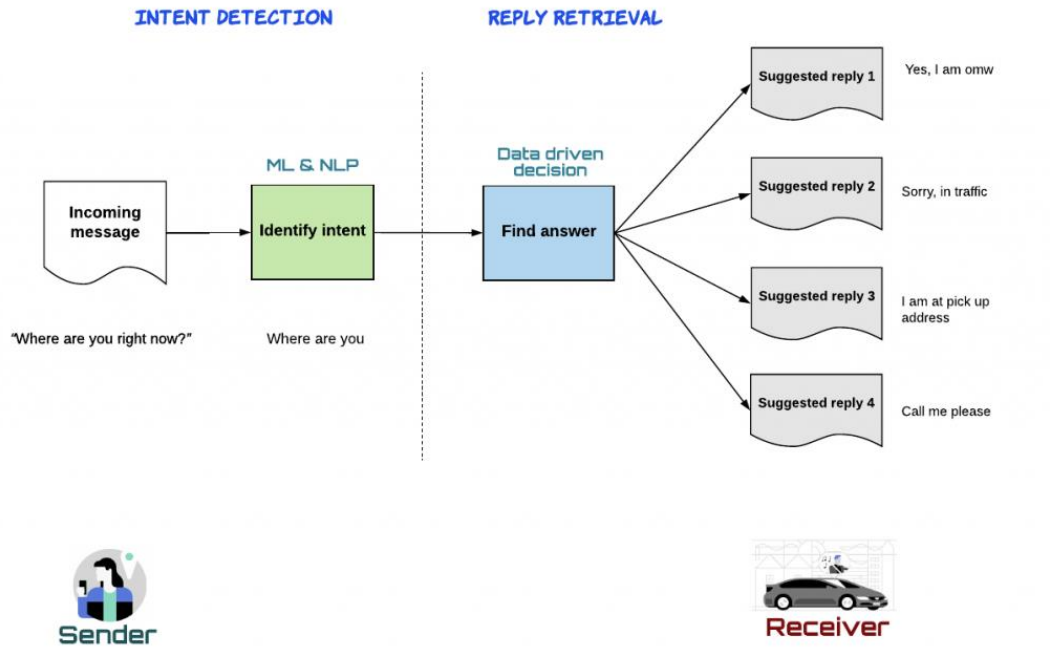


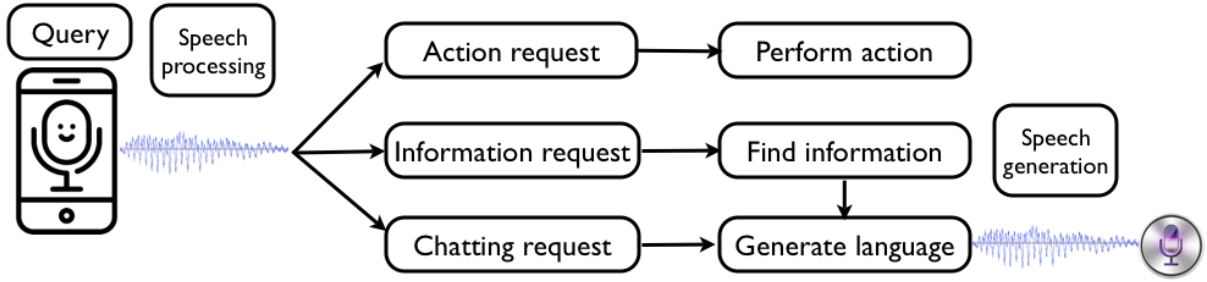
Natural Language Generation

Question & Answering



Virtual Assistants





NLP Tools

- NLTK
- Spacy
- Gensim
- Scikit-learn

Toyota Connected NLP Application

- Intelligent Assistant

Reading Materials

- Linguistics:
  https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf
- NLP:
  CS224n – NLP with Deep Learning - Christopher Manning
  Real-World Natural Language Processing by Masato Hagiwara
- Deep Learning:
  https://www.deeplearningbook.org/ - Ian Goodfellow and Yoshua Bengio and Aaron Courville

  http://projector.tensorflow.org/
  https://pair-code.github.io/lit/tutorials/sentiment/
  https://www.youtube.com/watch?v=kiPysxvkmoU&t=63s