Hacking into the NLP and ML behind Chatbots

Shubhi Saxena
Product Manager,
Yellow messenger

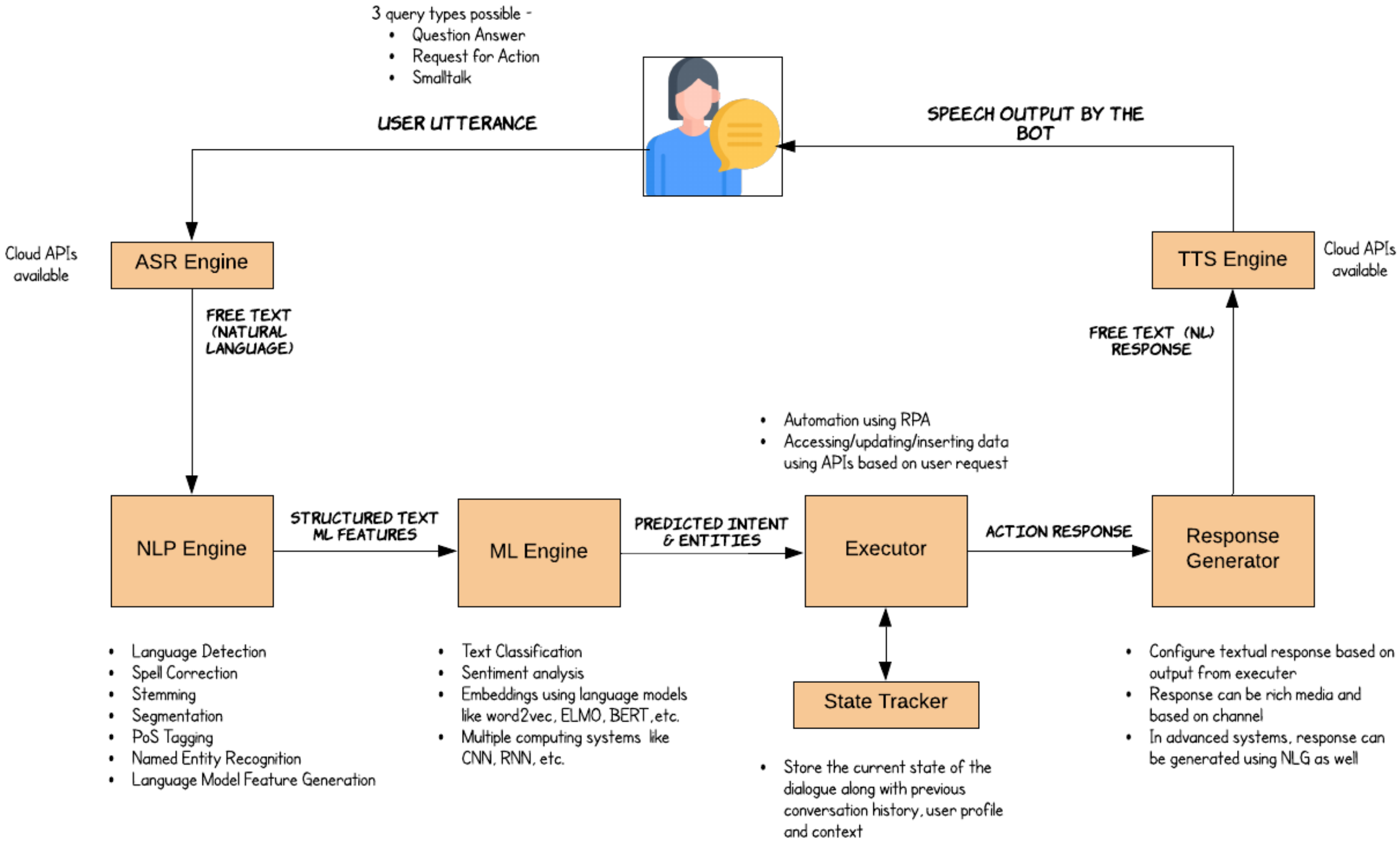
Why are enterprises talking about chatbots?

- No friction
- Instant answers
- Always available
- Automated Actions
- Natural conversations
- Personalised experiences
- Bots don’t forget or judge!

Let’s meet some real bots!

(Live Showcase)

How do chatbots work?



Dan Jurafsky

Ambiguity makes NLP hard: "Crash blossoms"

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Dan Jurafsky

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

But that's what makes it fun!

Present State of Language Technology

```
import nltk
sentence = "Awesome to be at Pyladies!"
token = nltk.word_tokenize(sentence)
nltk.pos_tag(token)
```

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Basic Text Processing

- Tokenisation - language issues, proper noun issues, abbreviations, periods, symbols, OOV words, etc.

- Normalisation & stemming (e.g. U.S., US, U.S.A. —> usa ; case folding)

- Lemmatisation (the boy's cars are different colors → the boy car be different color)

- Stemming (e.g. automate(s), automatic, automaton - all reduced to automat.

- Sentence segmentation (difficult in speech-to-text processing)

Intro to n-Grams

Dan Jurafsky

Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5...w_n)$

- Related task: probability of an upcoming word:

  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

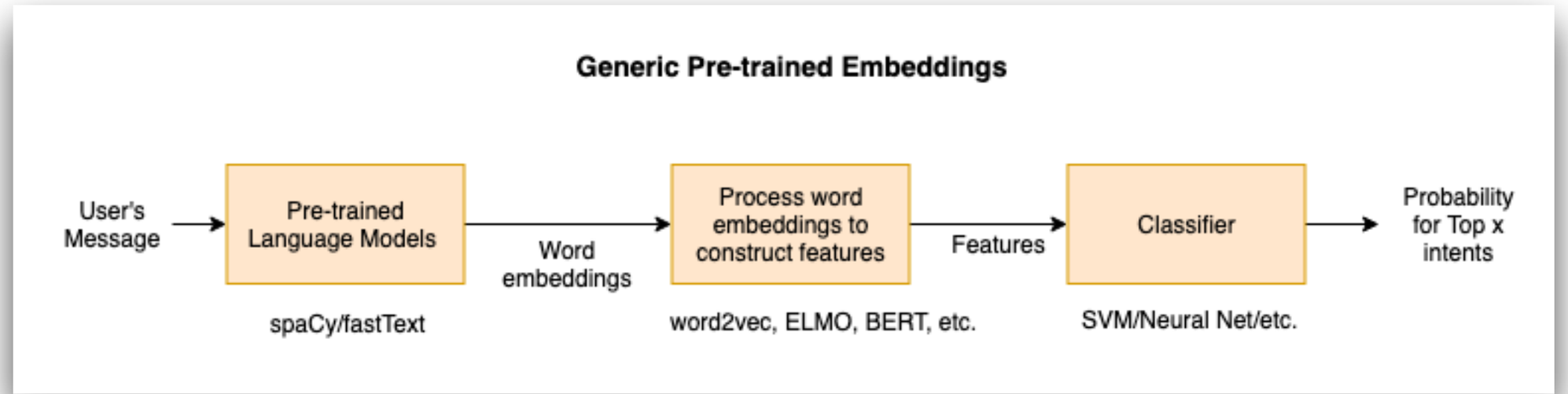
  $P(W)$ or $P(w_n|w_1,w_2...w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

Word embeddings

- Word embeddings are distributed representations of text in an n-dimensional space (to bridge the gap between human understanding and machines).
- One-hot encoding : vector the size of label array - not efficient
- Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions
- Each unique word in the corpus is assigned a corresponding vector in the space.
- Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.
- Other models : Glove (co-occurence) , fastText (character level representation)

NLU in chatbots : Intent Classification



- What is an intent
- What are word embeddings
- What is a classifier
- What are classification features
- Drawbacks of this approach
- Alternative - Train word embeddings from scratch using domain-specific data (supervised embeddings)
- How to choose?
- Challenges - similar intents, multiple intents, skewed data, OOV words

Parts of Speech Tagging

- Eight parts of speech taught in English but more can be used for practical purposes in NLP
- Use-Cases : NER, IE, TTS pronunciation, input to a parser
- Useful features -
  - Knowledge of neighbouring words
  - Word probabilities
  - Word structure (prefix, suffix, capitalisation, symbols, periods, wird shapes, etc.)

Information Extraction(IE)

- Goals of Information Extraction-
  - Organise information so that it can be consumed by people
  - Convert information into a precise semantic format on which computer algorithms can run inferences.

- Simple task - Extract clear, factual information from documents
- Example - Mail clients automatically detect dates and offer to schedule meeting/block calendar

- Difficult - Word meaning Disambiguation and combining different sources of related data to derive inferences

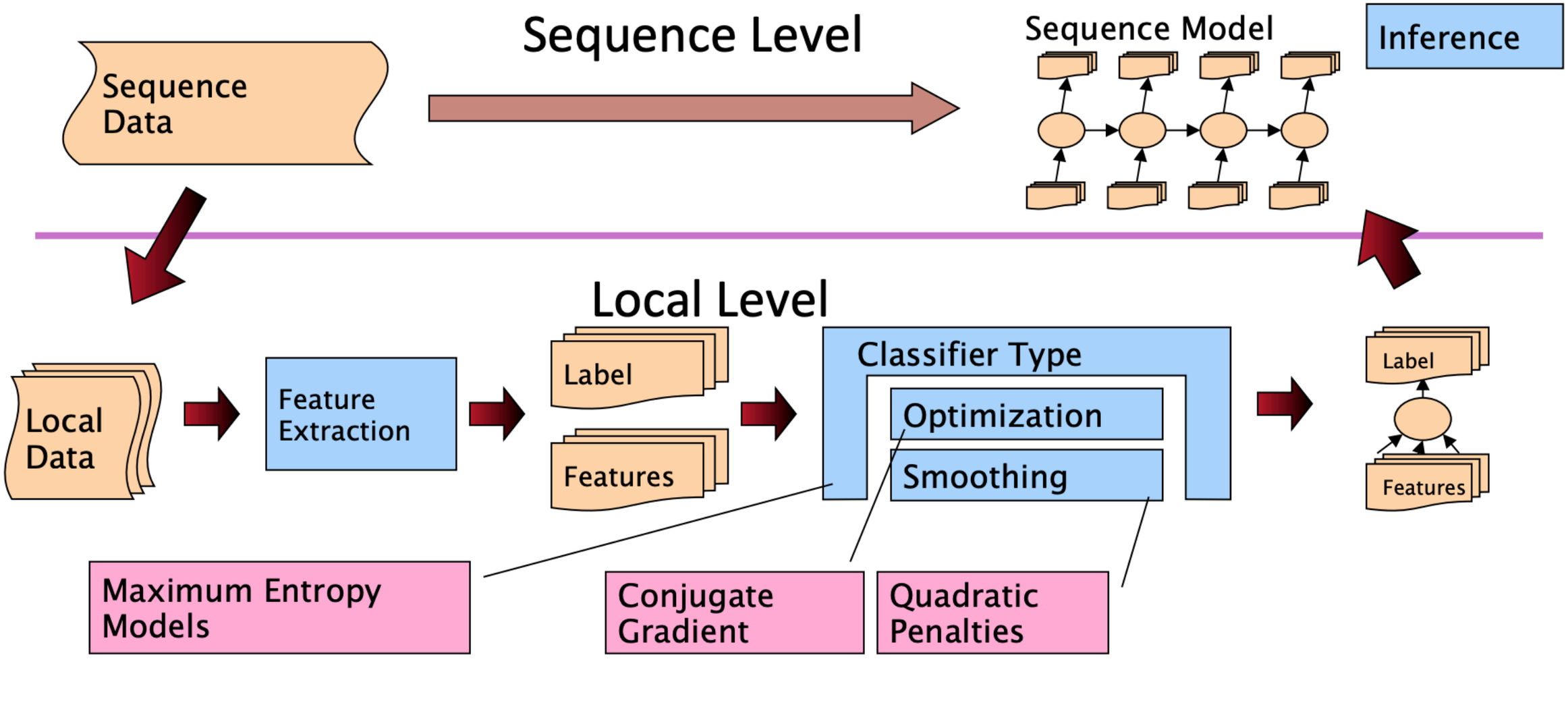
NLU : Named Entity Recognition (NER)

- Sub-task of IE - Identify and classify ‘entities’ in texts
- What are entities? How can we use them in chatbots?
- Rule-based : Facebook’s duckling (demo) - ordinal, duration, date, etc.
- Pre-trained models : SpaCy (Try here) - person, organisation, place, etc.
- Custom entity detection (annotation)
- Challenges - fuzzy entities, extracting addresses, and mapping of extracted entities

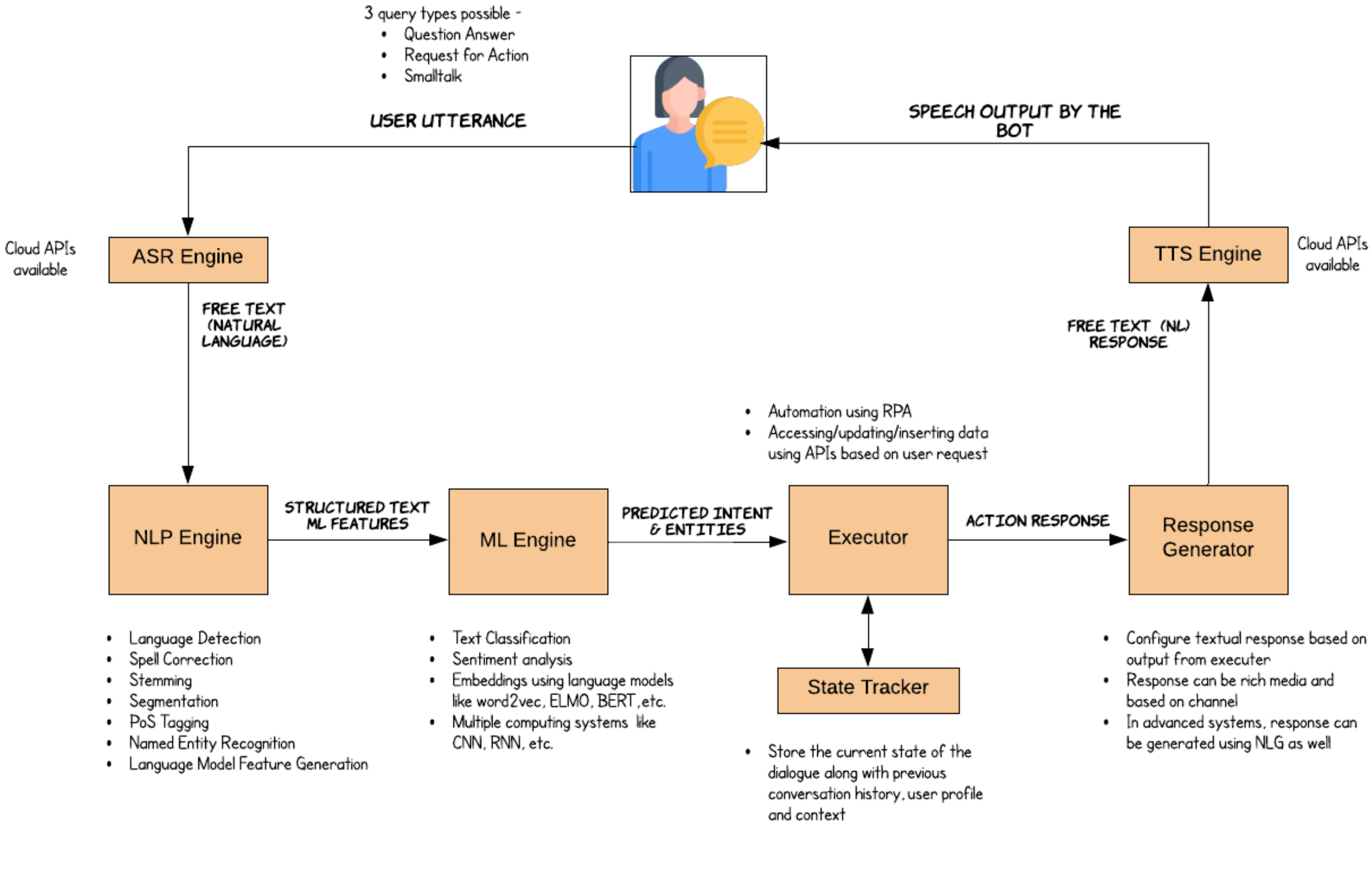
Sequencing using Conditional Markov Models

Christopher Manning

Inference in Systems



Now let us look at this again!



Further Reading

- Stanford's Intro to NLP course by Dan Jurafsky - link
- Spacy crash course - link
- We could not discuss Text Classification - Google's Crash course link
- Metablog by Pratik Bhavsar (if you want to go Ninja) - link

yellowmessenger

We are Hiring!

Shubhi Saxena
shubhi@yellowmessenger.com