How Has Deep Learning Revolutionized Human Language Technology?

Tao Ma, Ph.D.

Vice President of the Language Technology Group
ASAPP Inc.
12/6/2020

ASAPP

Human Language Communication

- Communication is transferring information, through encoding/decoding/channels.
- Human language communication deals with information on a conceptual level, integrated with knowledge of speech production, perception and linguistics.

Human Language Technology

How computer can analyze, produce, modify or respond to human texts and speech.

Speech Recognition

Automatic Speech Recognition (ASR) converts spoken words into text. It detects spoken sounds and recognize them as words.

Speech Synthesis

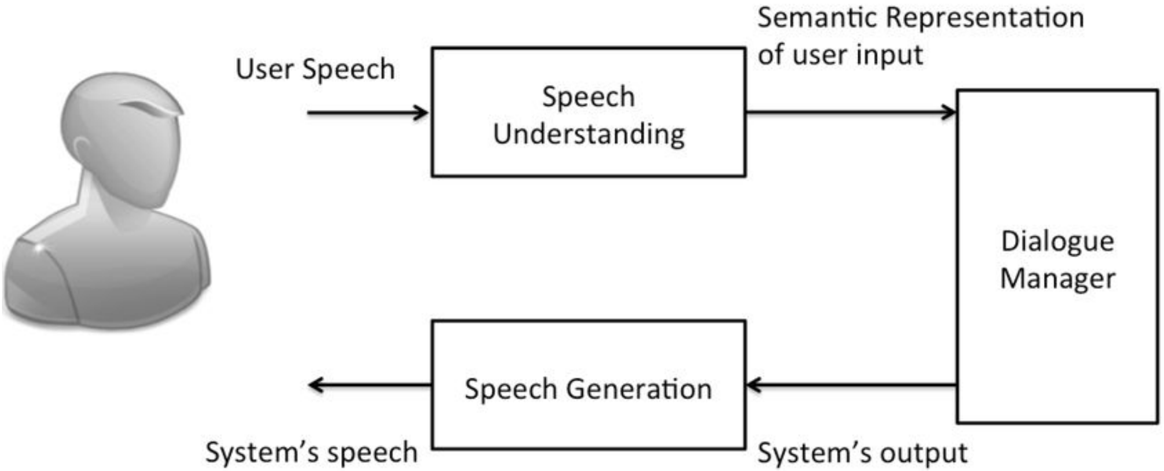
Text To Speech (TTS) converts text to voice. It's a technology that makes computer talk.

Natural Language Processing

Natural Language Processing (NLP) helps computers understand, interpret and manipulate human language.

Behind Siri/Alexa/Google Assistant/Cortana

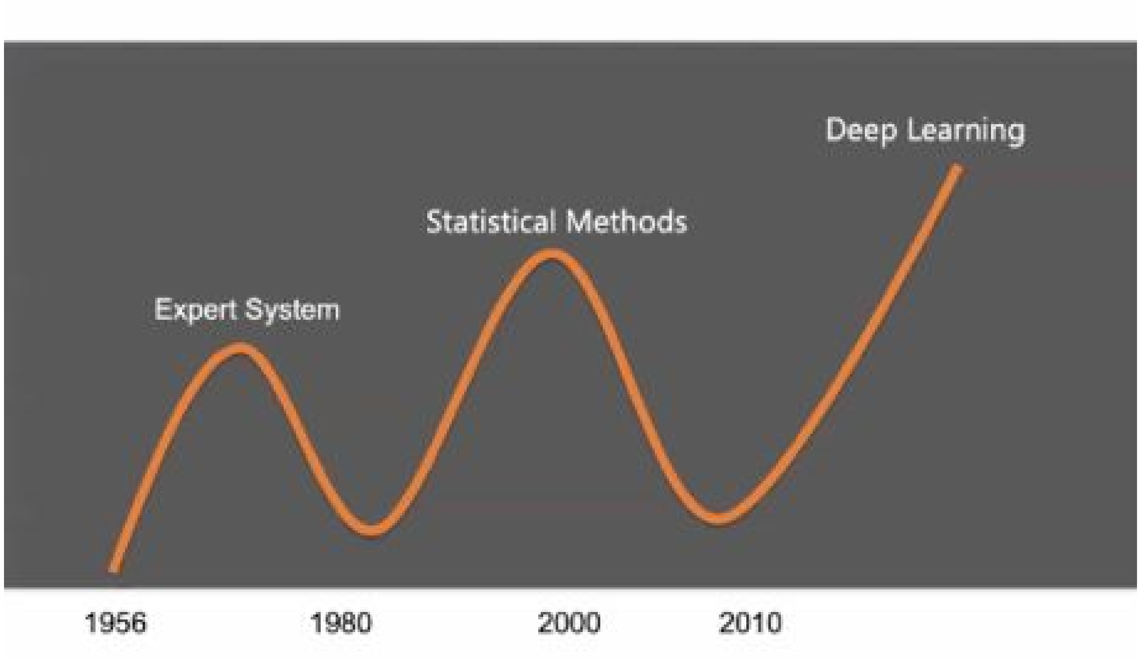
How AI assistants work? The architecture of a Spoken Dialogue System.



https://www.researchgate.net/publication/306091536_Statistical_Task_Modeling_of_Activities_of_Daily_Living_for_Rehabilitation

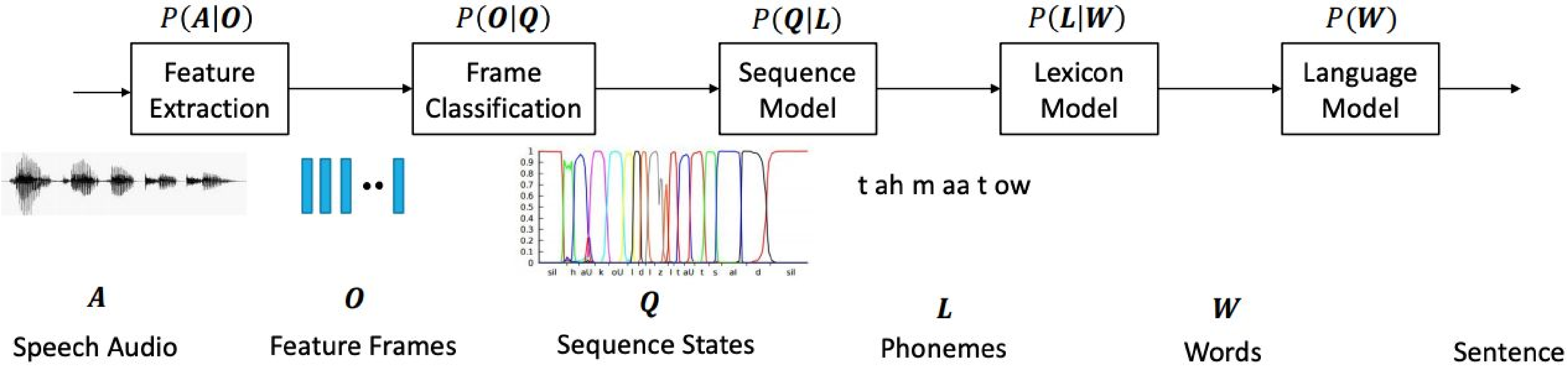
Deep Learning: 3rd wave of Artificial Intelligence

In 2011, *AlexNet* achieved big wins in Computer Vision, then deep learning is quickly being adopted in Speech and NLP.



Speech Recognition

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{W}|\boldsymbol{O}) = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{A}|\boldsymbol{O})P(\boldsymbol{O}|\boldsymbol{Q})P(\boldsymbol{Q}|\boldsymbol{L})P(\boldsymbol{L}|\boldsymbol{W})P(\boldsymbol{W})$$



$P(\boldsymbol{A}|\boldsymbol{O})$ Feature Extraction → $P(\boldsymbol{O}|\boldsymbol{Q})$ Frame Classification → $P(\boldsymbol{Q}|\boldsymbol{L})$ Sequence Model → $P(\boldsymbol{L}|\boldsymbol{W})$ Lexicon Model → $P(\boldsymbol{W})$ Language Model →

t ah m aa t ow

$\boldsymbol{A}$ Speech Audio

$\boldsymbol{O}$ Feature Frames

$\boldsymbol{Q}$ Sequence States

$\boldsymbol{L}$ Phonemes

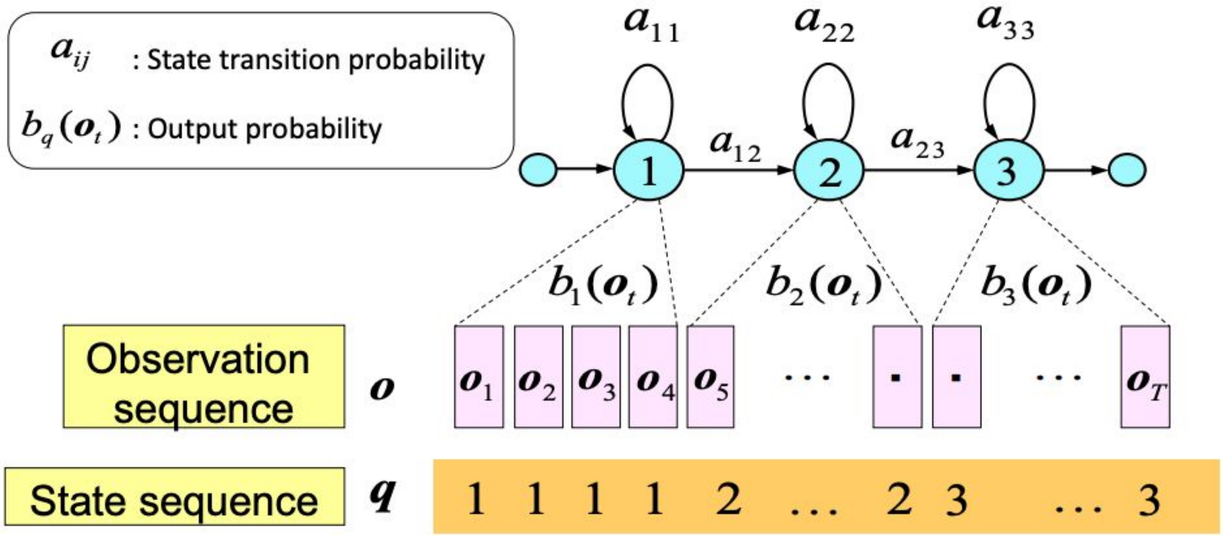
$\boldsymbol{W}$ Words

Sentence

http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Before Deep Learning: Hidden Markov Models (HMM)

The Markov chain whose state sequence is unknown.

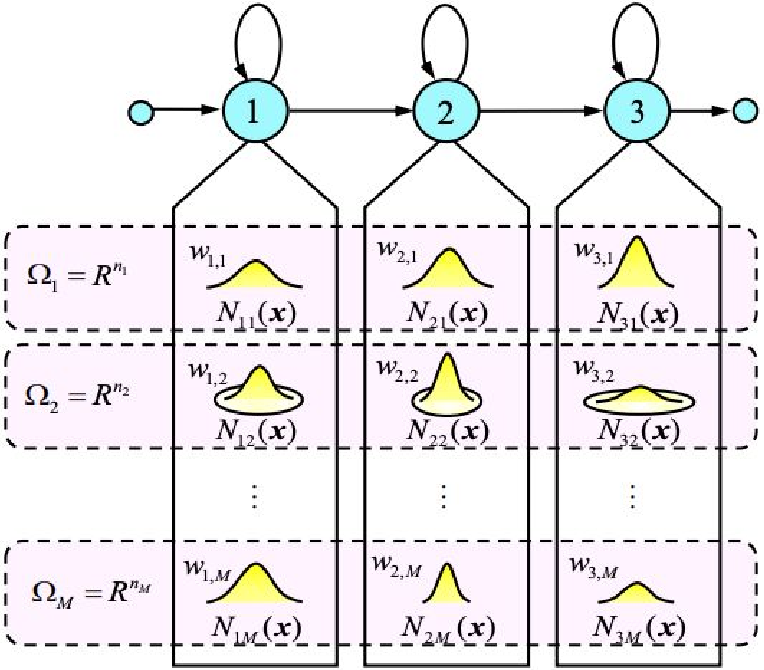


ASAPP

http://hts.sp.nitech.ac.jp/archives/2.3/HTS_Slides.zip

Before Deep Learning: Gaussian Mixture Models (GMM)

Output probability is modeled by Gaussian mixture models.

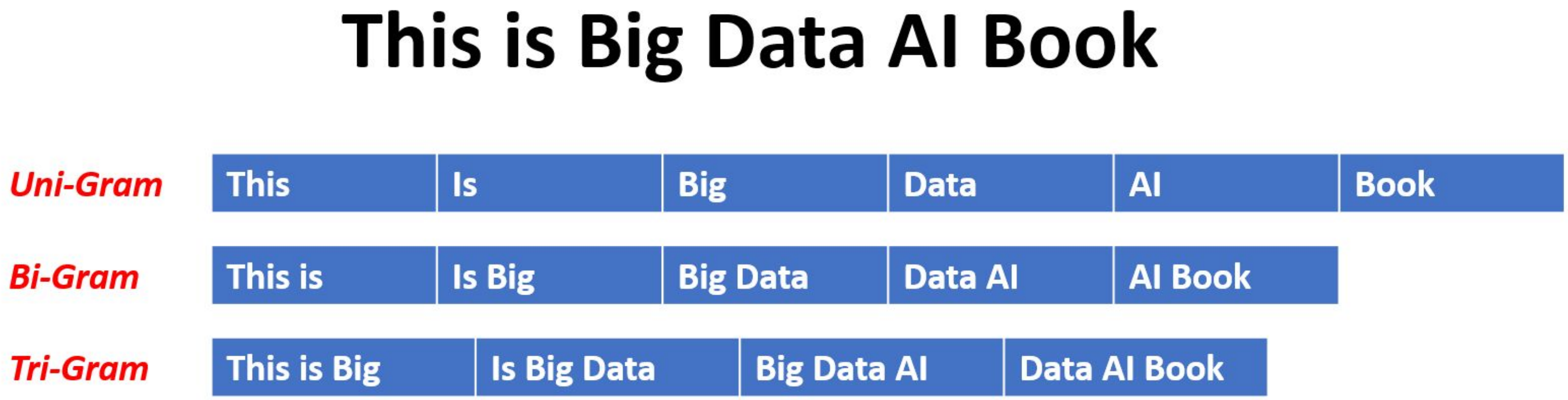


ASAPP

http://hts.sp.nitech.ac.jp/archives/2.3/HTS_Slides.zip

Before Deep Learning: N-gram Language Model

Given a sequence of N-1 words, an N-gram model predicts the most probable word that might follow this sequence.



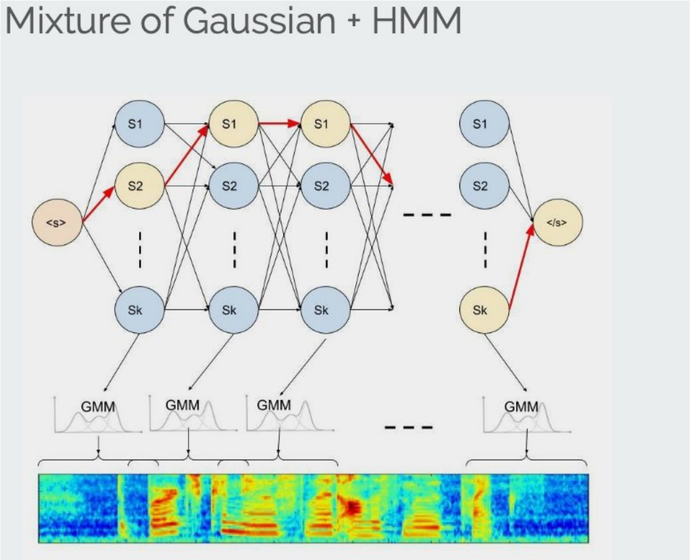
This is Big Data AI Book

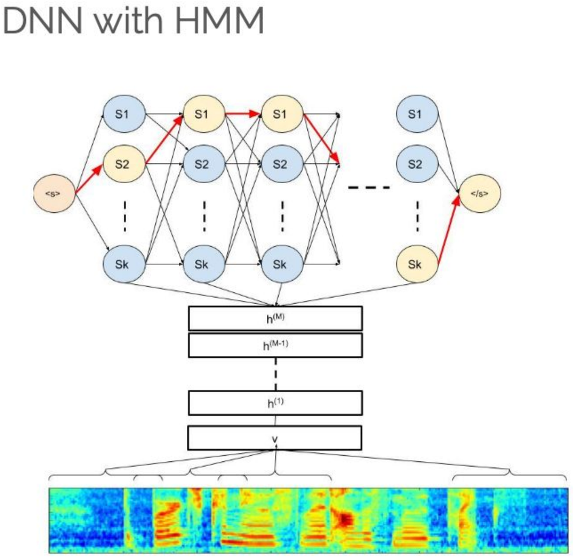
Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

https://devopedia.org/images/article/219/7356.1569499094.png

Acoustic Modeling: GMM-HMM to DNN-HMM

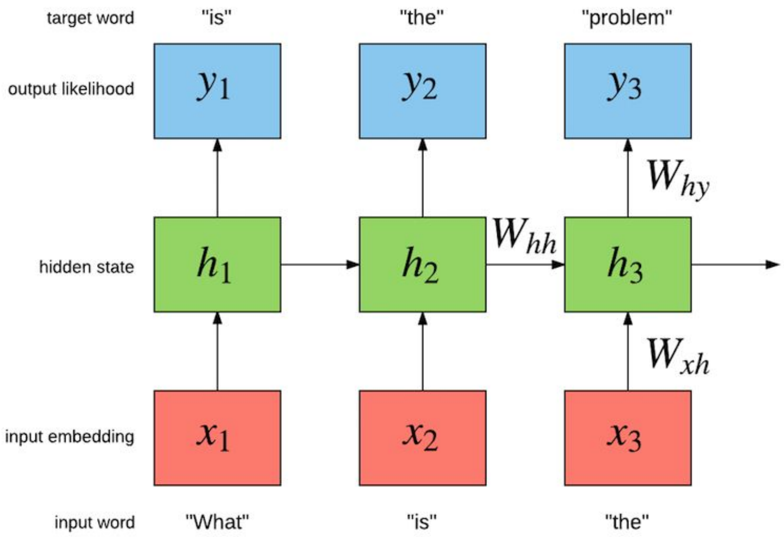
Deep Neural Network (DNN) replaced Gaussian Mixture Model (GMM) for audio frame classification.





Language Modeling: N-gram to NNLM

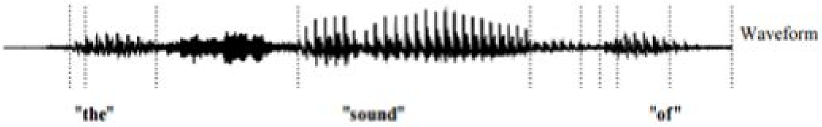
Neural network language model (NNLM) replaced N-gram for language modeling.

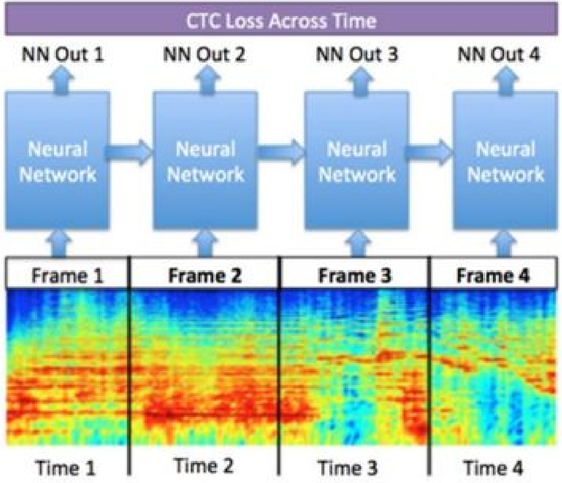


http://torch.ch/blog/2016/07/25/nce.html

Sequence Modeling: HMM to CTC

Connectionist Temporal Classification (CTC) is a method for labeling unsegmented data sequences.



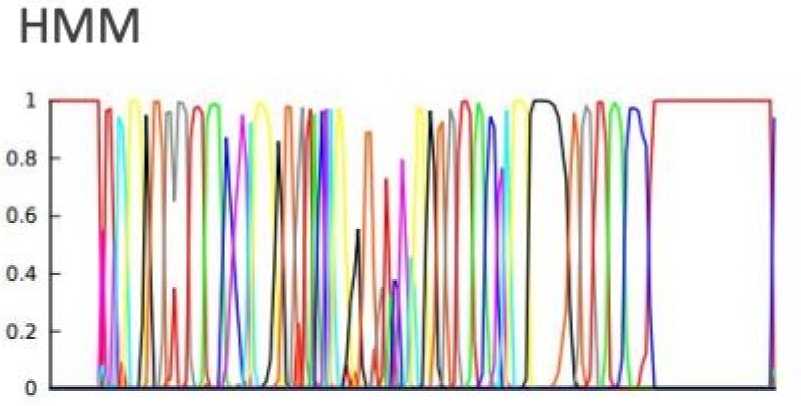


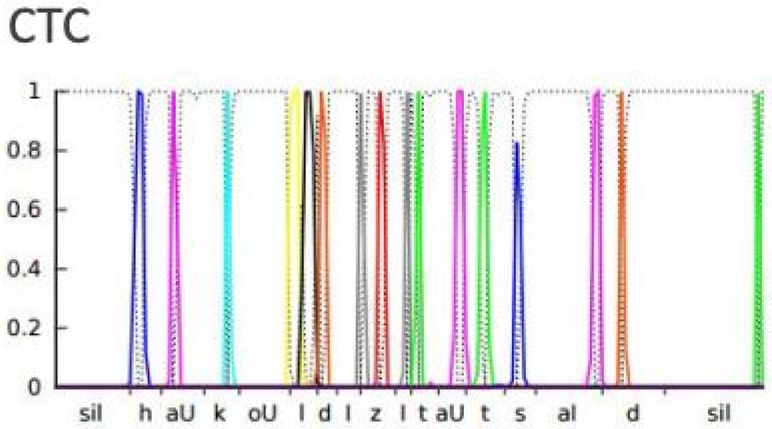
ftp://ftp.idsia.ch/pub/juergen/icml2006.pdf

Sequence Modeling: HMM to CTC

Output probability of LSTM-HMM vs. LSTM-CTC.

CTC has spiky predictions, more discriminable between states than HMM.



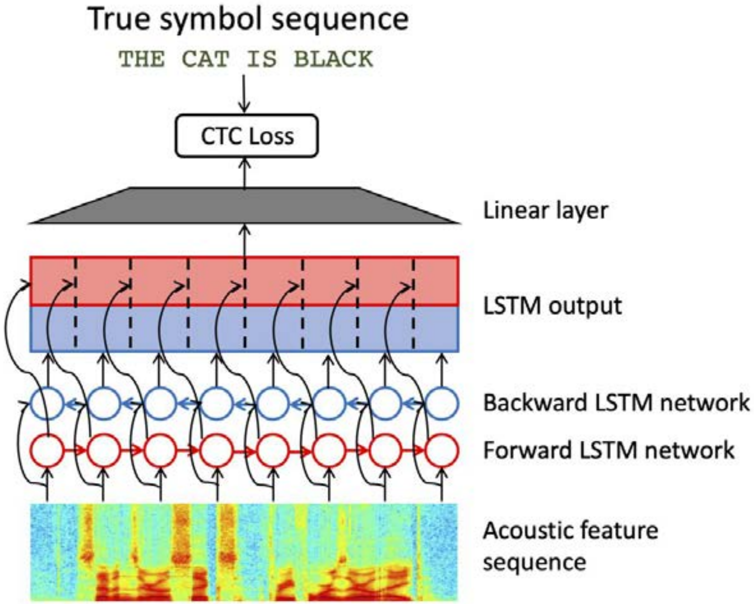


Sak, Haşim, et al. "Learning acoustic frame labeling for speech recognition with recurrent neural networks." ICASSP, 2015.

Hybrid ASR to End-to-end ASR

Hybrid ASR: separated acoustic model, language model, and pronunciation model.

End-to-end ASR: a neural network to tackle audio frame sequence to word sequence mapping as a sequence-to-sequence learning problem.



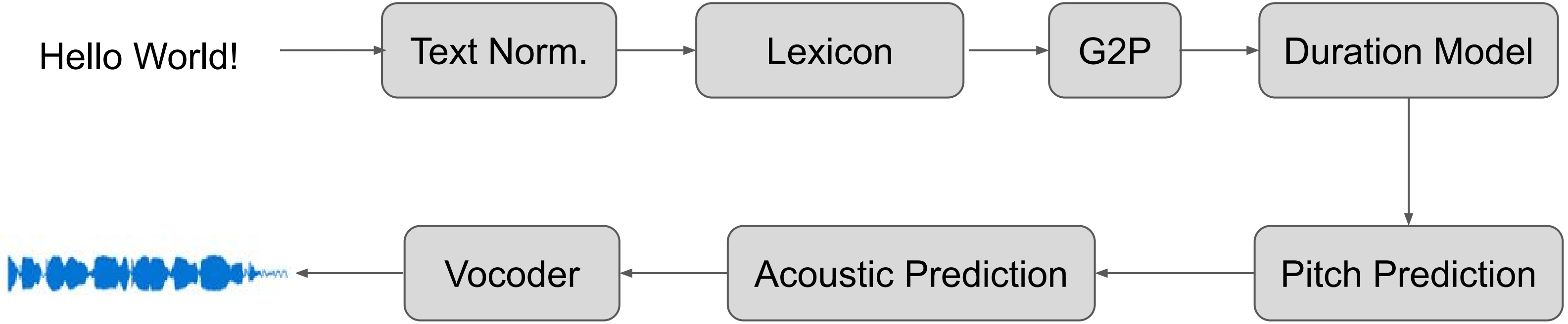
ASAPP

https://www.ibm.com/blogs/research/2019/10/end-to-end-speech-recognition/

Text-to-speech (TTS)

Traditional Paradigms:

- Concatenative synthesis
- Parametric synthesis



ASAPP

Sequence-to-sequence Modeling

- **Speech Recognition (continuous -> discrete):**

→ Hello World!

- **Speech Synthesis (discrete -> continuous):**

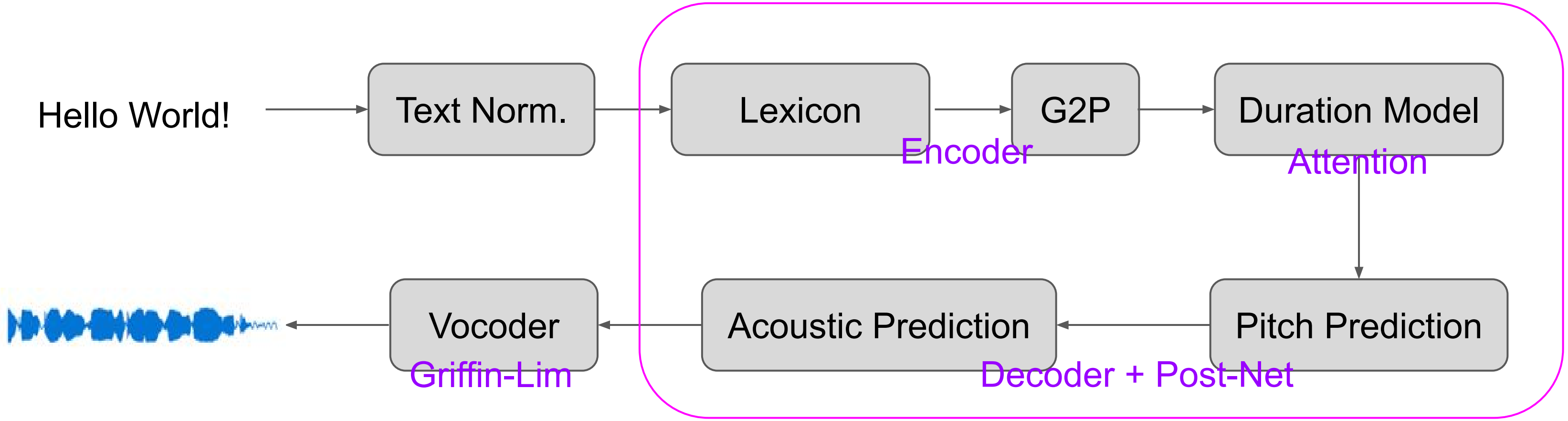
Hello World! →

- **Machine Translation (discrete -> discrete):**

Hello World! → Hola Mundo!

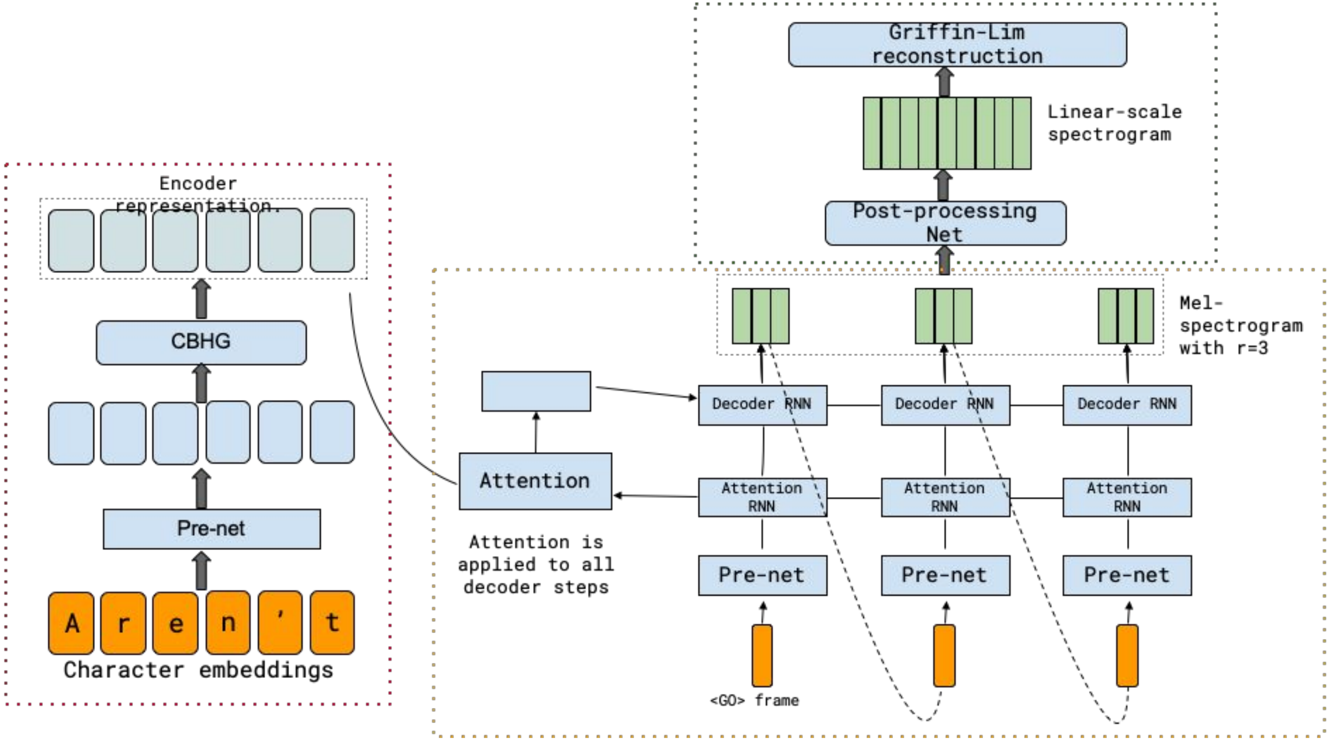
End-to-end TTS

A neural network to tackle word sequence to audio frame sequence mapping as a sequence-to-sequence learning problem.



ASAPP

Tacotron: Google's End-to-end TTS architecture



ASAPP

https://arxiv.org/pdf/1703.10135.pdf

Common NLP Tasks

NLP enables computers to perform a wide range of natural language related tasks at all levels, ranging from parsing and part-of-speech (POS) tagging, to machine translation and dialogue systems.

Sentence Parsing

- Constituency parsing
- Semantic parsing
- Dependency parsing

Word Tagging

- Word segmentation
- Shallow syntax-chunking
- Named entity recognition
- Part-of-speech tagging
- Semantic role labeling
- Word sense disambiguation

Text Classification

- Sentiment analysis
- Text classification
- Temporal processing
- Coreference resolution

Text Pair Matching

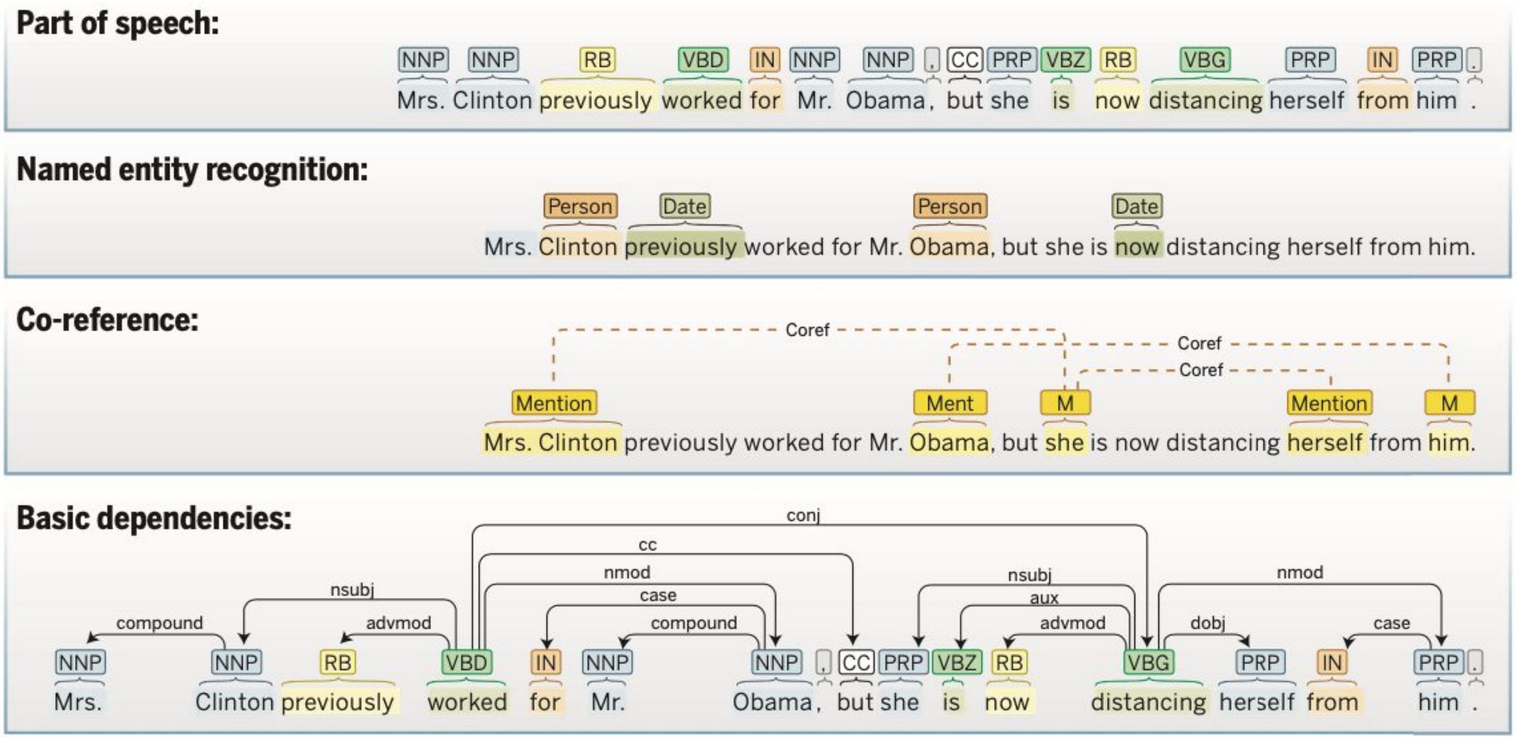
- Semantic textual similarity
- Natural language inference
- Relation prediction

Text Generation

- Language modeling
- Machine translation
- Simplification
- Summarization
- Dialogue
- Question answering

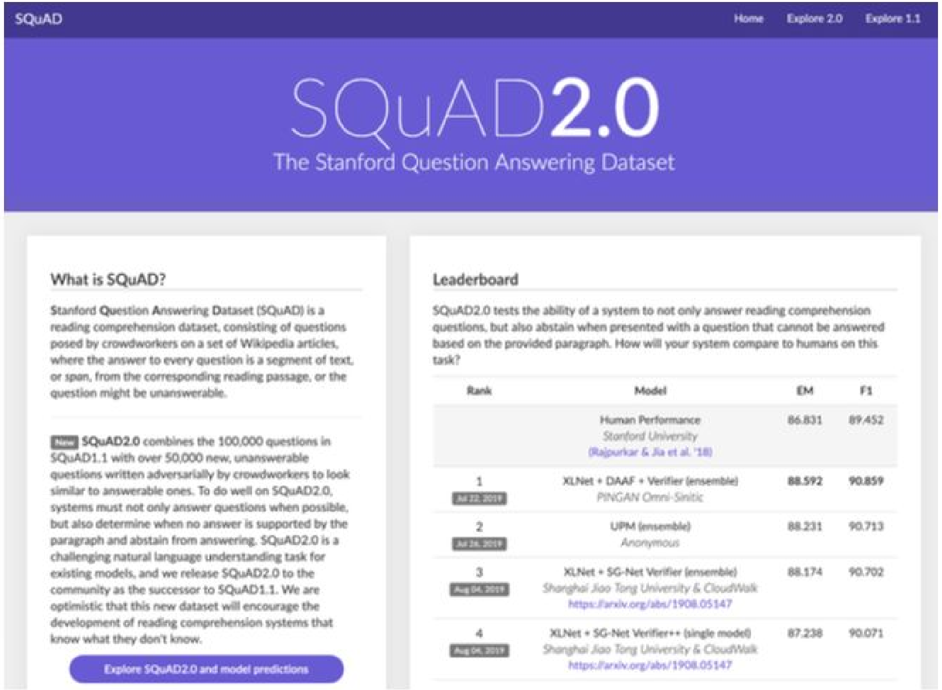
https://mobidev.biz/blog/natural-language-processing-nlp-use-cases-business

An Example of NLP Linguistic Structure Analysis


Part of speech:
NNP NNP RB VBD IN NNP NNP , CC PRP VBZ RB VBG PRP IN PRP .
Mrs. Clinton previously worked for Mr. Obama , but she is now distancing herself from him .
Named entity recognition:
Person Date Person Date
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.
Co-reference:
Coref
Coref
Coref
Mention Ment M Mention M
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.
Basic dependencies:
conj
cc
nmod
nsubj
case
compound
advmod
compound
nsubj
aux
advmod
nmod
dobj
case
NNP NNP RB VBD IN NNP NNP , CC PRP VBZ RB VBG PRP IN PRP .
Mrs. Clinton previously worked for Mr. Obama , but she is now distancing herself from him .

https://cs224d.stanford.edu/papers/advances.pdf

An Example of NLP Question Answering



SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

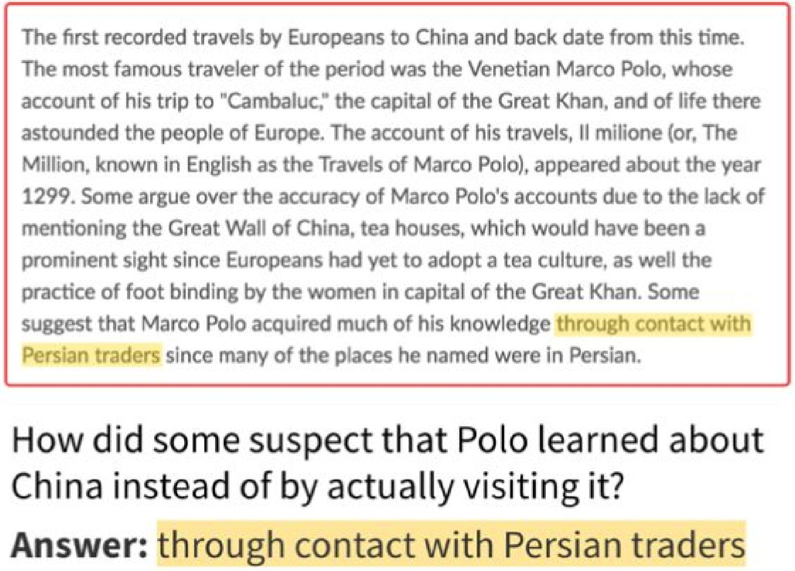
Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

New SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	88.592	90.859
2 Jul 26, 2019	UPM (ensemble) Anonymous	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	87.238	90.071



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

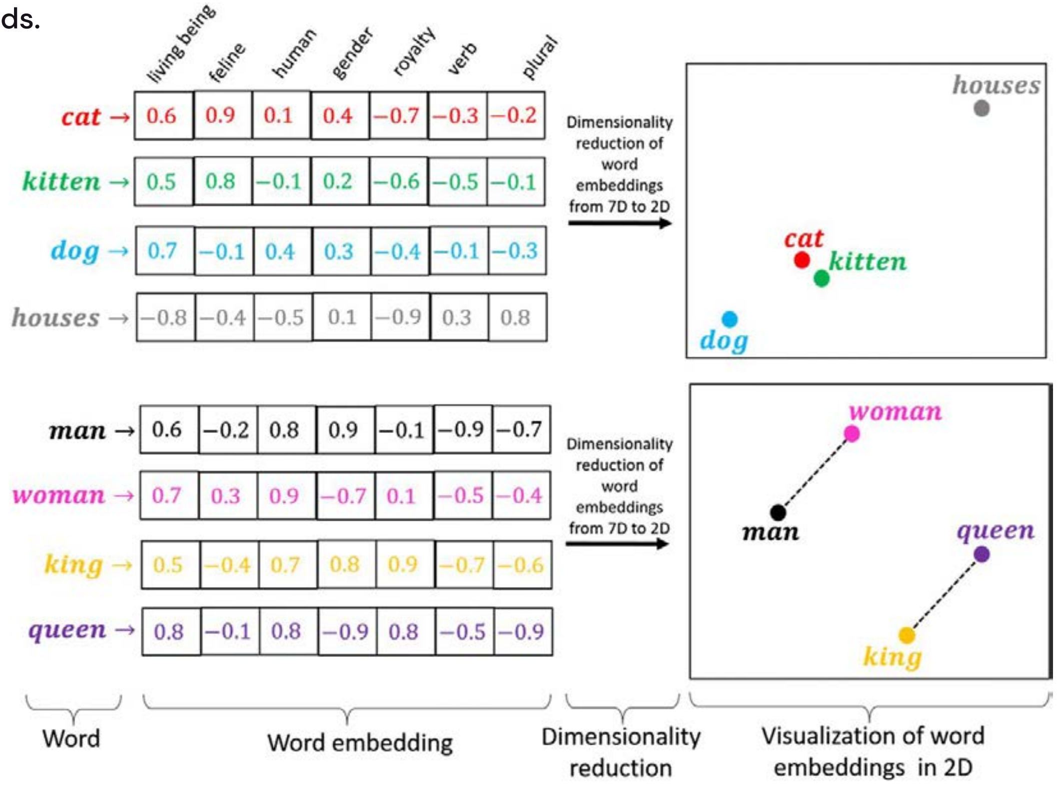
How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

https://rajpurkar.github.io/SQuAD-explorer/

Word Embeddings

As the first data processing layer in a deep learning model, distributional vectors has advantage to capture similarity between words.



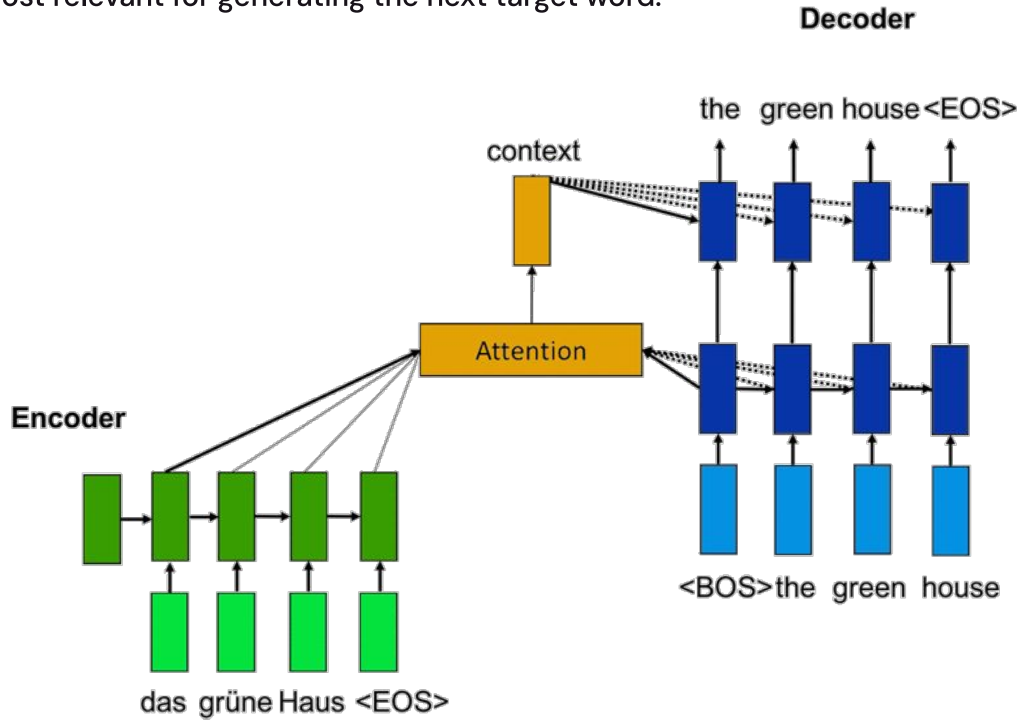
	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

https://medium.com/@hari4om/word-embedding-d816f643140

Attention Mechanism

In a classic Neural Machine Translation model, using the attention mechanism, the decoder can decide which words are most relevant for generating the next target word.

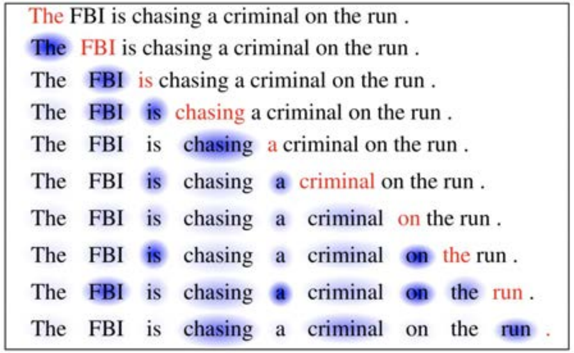


https://aws.amazon.com/blogs/machine-learning/train-neural-machine-translation-models-with-sockeye/

Self-attention & Transformer Architecture

Self-attention: allows the inputs to interact with each other and find out who they should pay more attention to.

The Transformer: first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution.



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

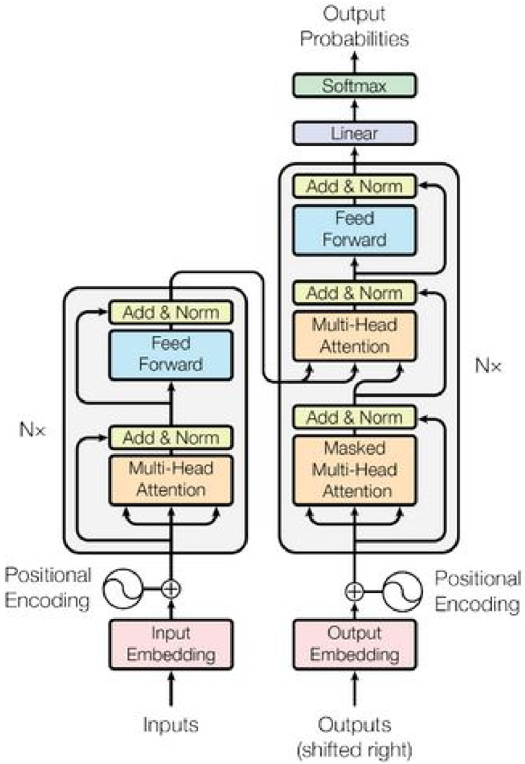
The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

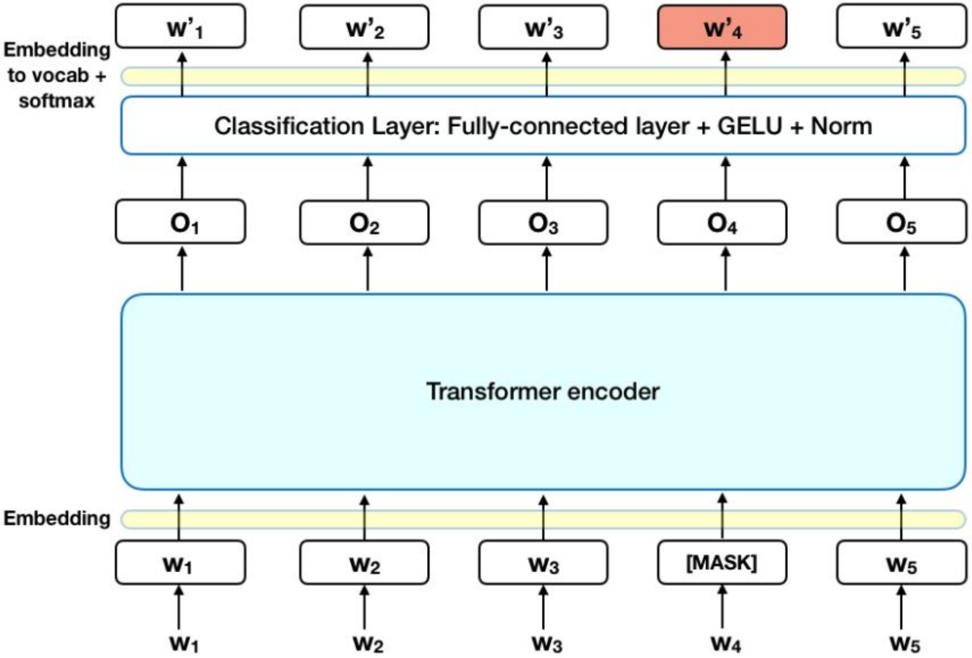


Vaswani et al. "Attention Is All You Need".
https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

BERT: State-of-the-art Language Model for NLP

BERT: Bidirectional Encoder Representations from Transformers, from Google.

Trained using Masked LM technique, BERT applied the bidirectional training of Transformer to language modelling and achieved state-of-the-art performance for many NLP tasks.


w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

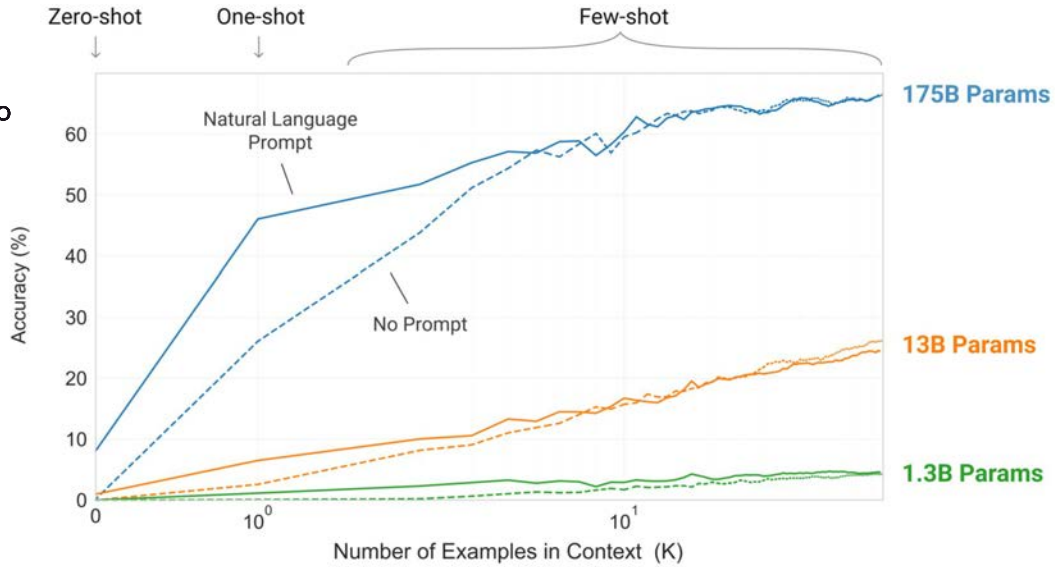
ASAPP

Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding."

GPT-3: Predicts Anything that has a Language Structure

GPT-3: Generative Pre-trained Transformer 3, from OpenAI.

The largest language model ever trained, with 175 billion parameters. The quality of the text generated by GPT-3 is so high that it is difficult to distinguish from that written by a human.

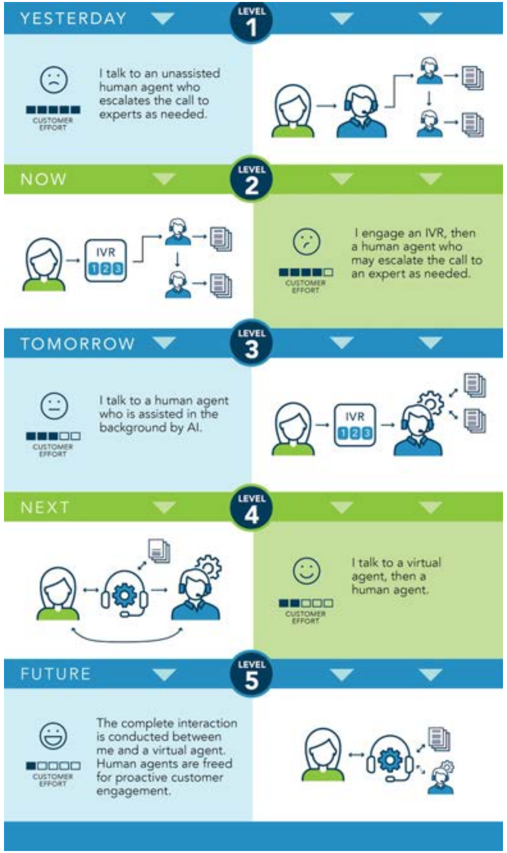


ASAPP

Brown et al. "Language Models are Few-Shot Learners".

Language Technology Revolution is transforming Contact Center

Amid the COVID-19 crisis, the global market for Call Centers estimated at $339.4 Billion in the year 2020.



I talk to an unassisted human agent who escalates the call to experts as needed.

I engage an IVR, then a human agent who may escalate the call to an expert as needed.

I talk to a human agent who is assisted in the background by AI.

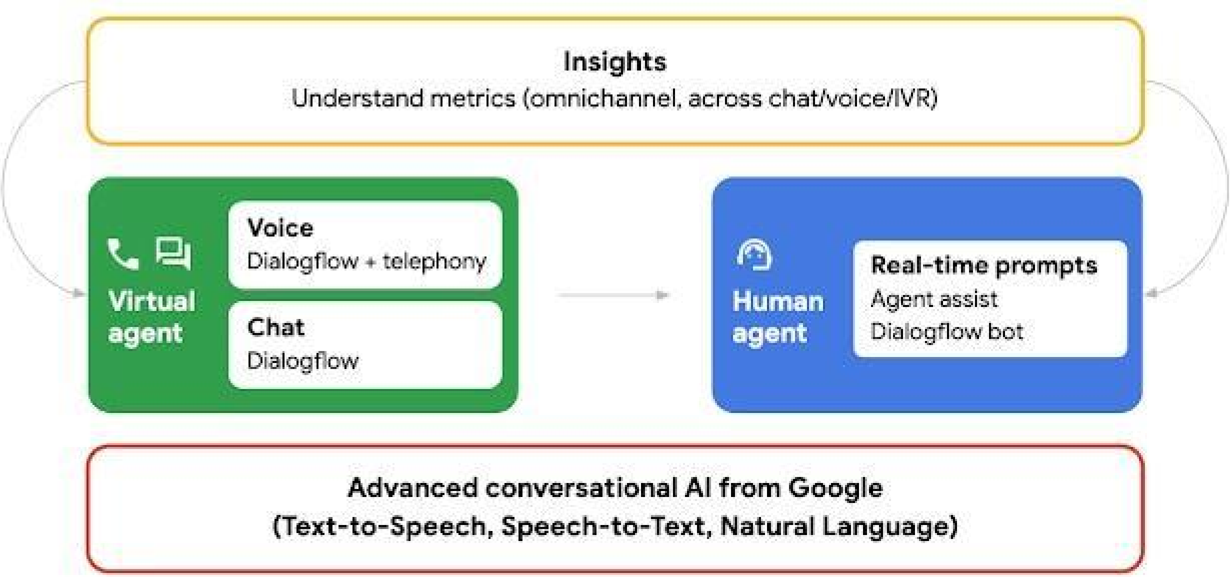
I talk to a virtual agent, then a human agent.

The complete interaction is conducted between me and a virtual agent. Human agents are freed for proactive customer engagement.

https://www.five9.com/blog/five-levels-of-contact-center-agent-augmentation
https://www.globenewswire.com/news-release/2020/07/17/2063820/0/en/Global-Call-Centers-Industry.html

Google Contact Center AI

Improve customer service with AI that understands, interacts, and talks.



Insights
Understand metrics (omnichannel, across chat/voice/IVR)

Virtual agent

Voice
Dialogflow + telephony

Chat
Dialogflow

Human agent

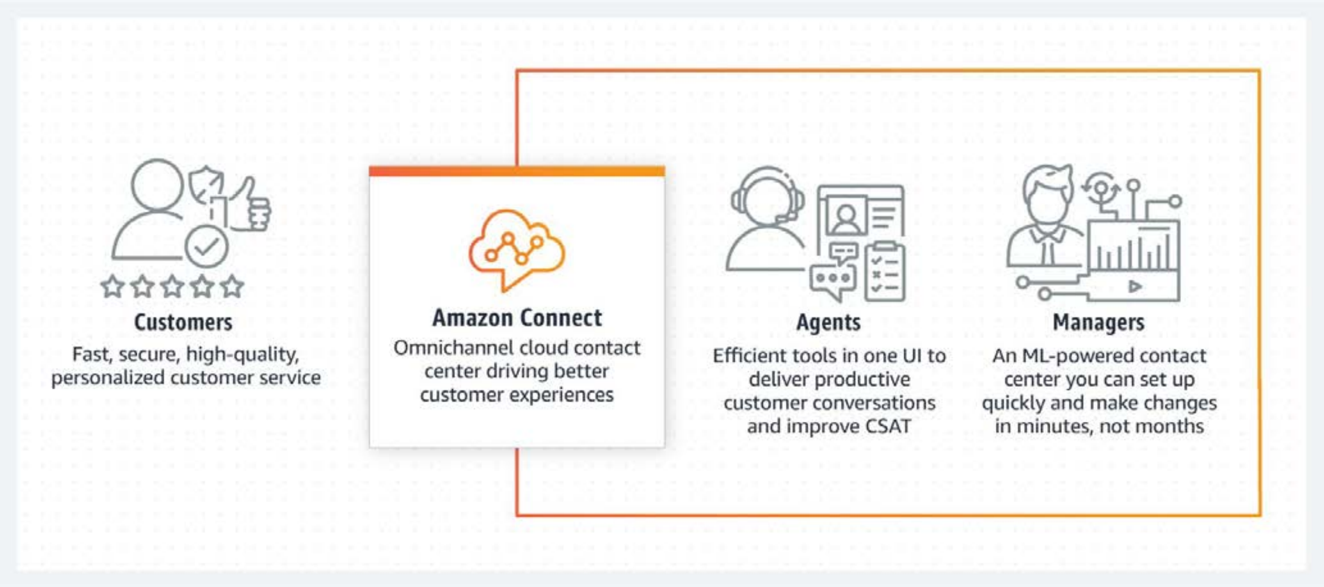
Real-time prompts
Agent assist
Dialogflow bot

Advanced conversational AI from Google
(Text-to-Speech, Speech-to-Text, Natural Language)

https://cloud.google.com/solutions/contact-center

Amazon Connect

An omnichannel cloud contact center that helps you provide superior customer service at a lower cost.



Customers

Fast, secure, high-quality, personalized customer service

Amazon Connect

Omnichannel cloud contact center driving better customer experiences

Agents

Efficient tools in one UI to deliver productive customer conversations and improve CSAT

Managers

An ML-powered contact center you can set up quickly and make changes in minutes, not months

https://aws.amazon.com/connect/

Future Directions & Opportunities

"*Any sufficiently advanced technology is indistinguishable from magic.*"
-- Arthur C. Clarke

Joint Modeling of Speech & NLP

The *Status Quo* of separated Speech modeling and NLP modeling constraints language technology performance. Technology maturity of end-to-end deep learning architectures is opening up opportunities for a holistic language processing.

Multimodal Communication

With contact center adopting new communication technologies such as WebRTC, a unified multimodal interaction platform across voice, text, and video starts emerging.

Thank you!

ASAPP