

End-to-End Natural Language Processing

Rahul Kumar
Chief AI Scientist

BOTSUPPLY

Natural Language Processing: The Goal

[**Context:** We saw some lions; guide tells us it's a rare sight. Few min later we pass a second safari guide.]



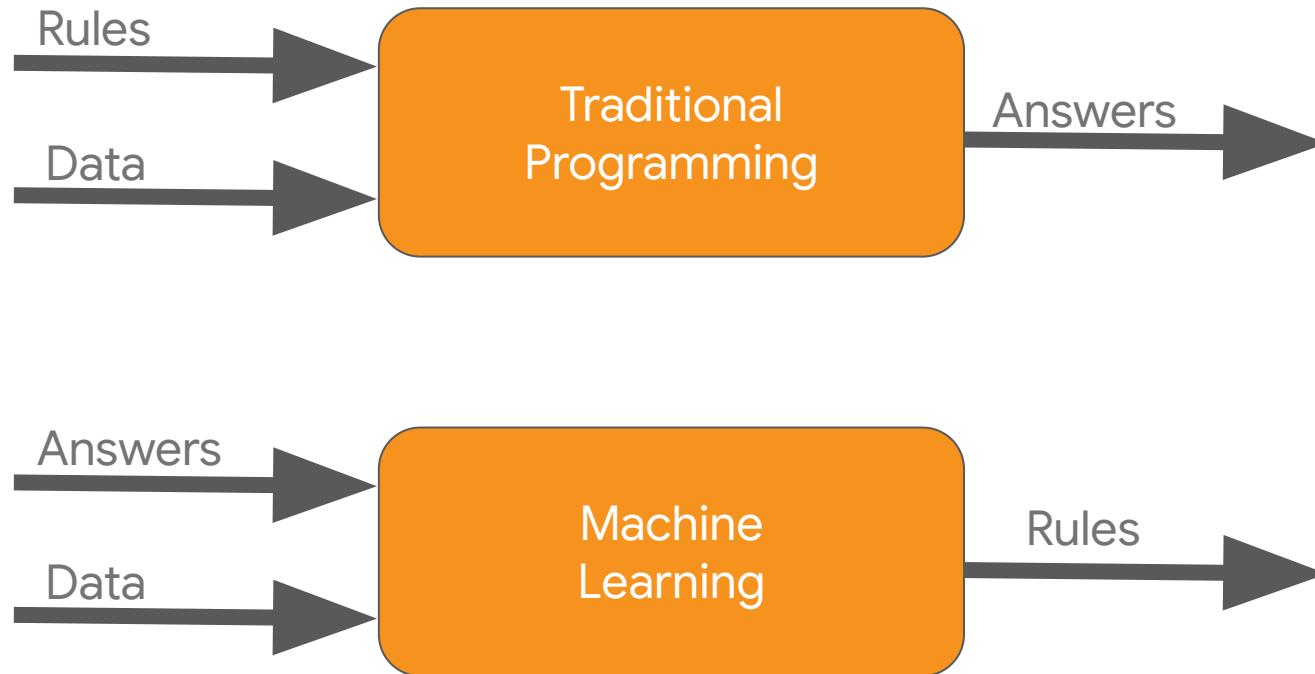
Our guide to other guide: Vi så nogle løver

[Second **driver** speeds off]

Simon and I: I bet he told him about the lion.

Our guide: Yep!

Programming Paradigm



Before ~2013

2015 onwards

Traditional
Programming

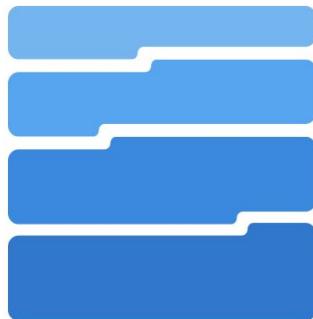
Machine
Learning

Bucketing

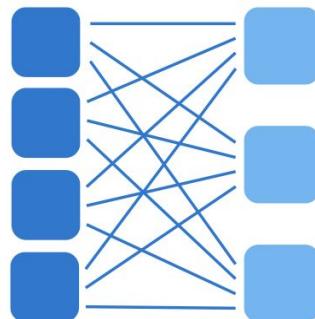
Crossing

Hashing

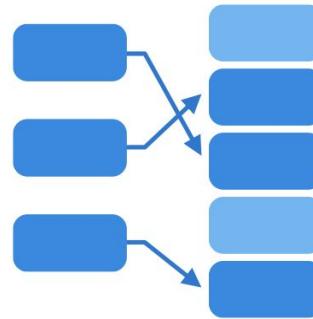
Embedding



Partition by range



Create new combinations



Limit size



Learn a new
representation

Before ~2013

2015 onwards

Traditional
Programming

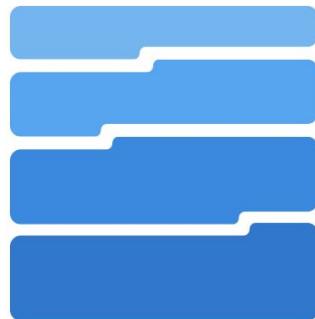
Machine
Learning

Bucketing

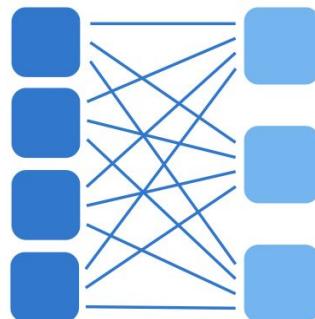
Crossing

Hashing

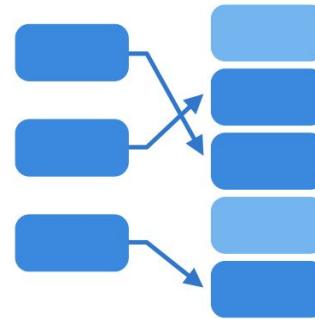
Embedding



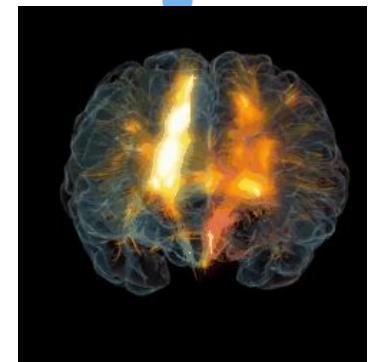
Partition by range



Create new combinations

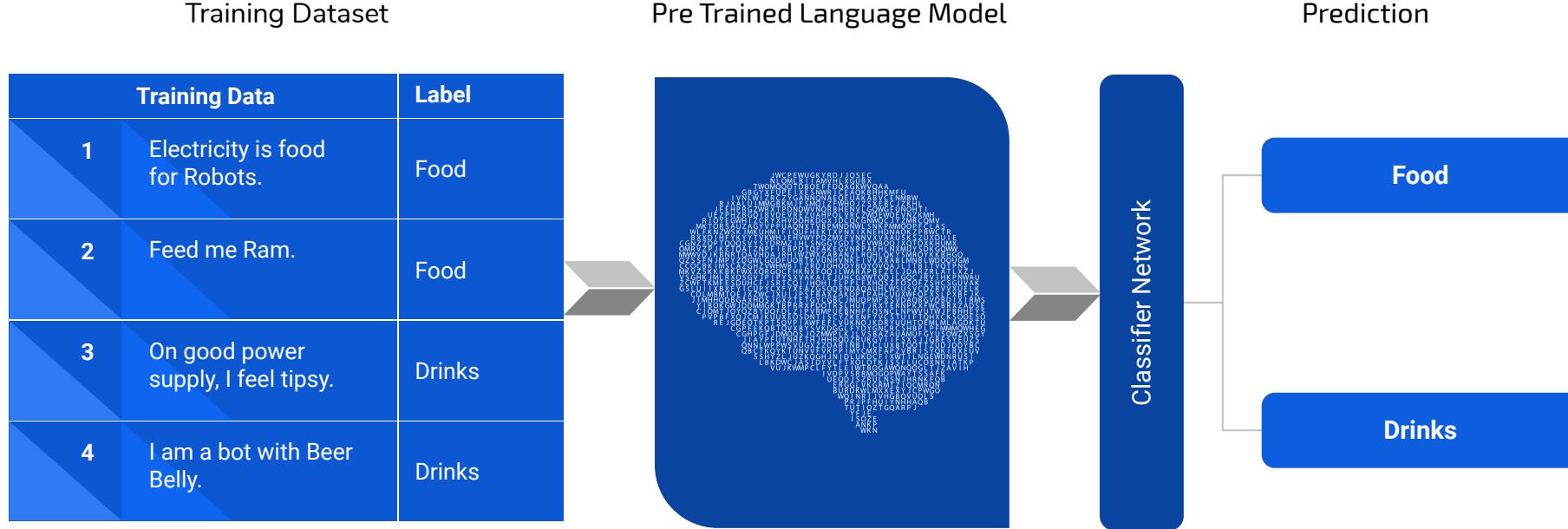


Limit size

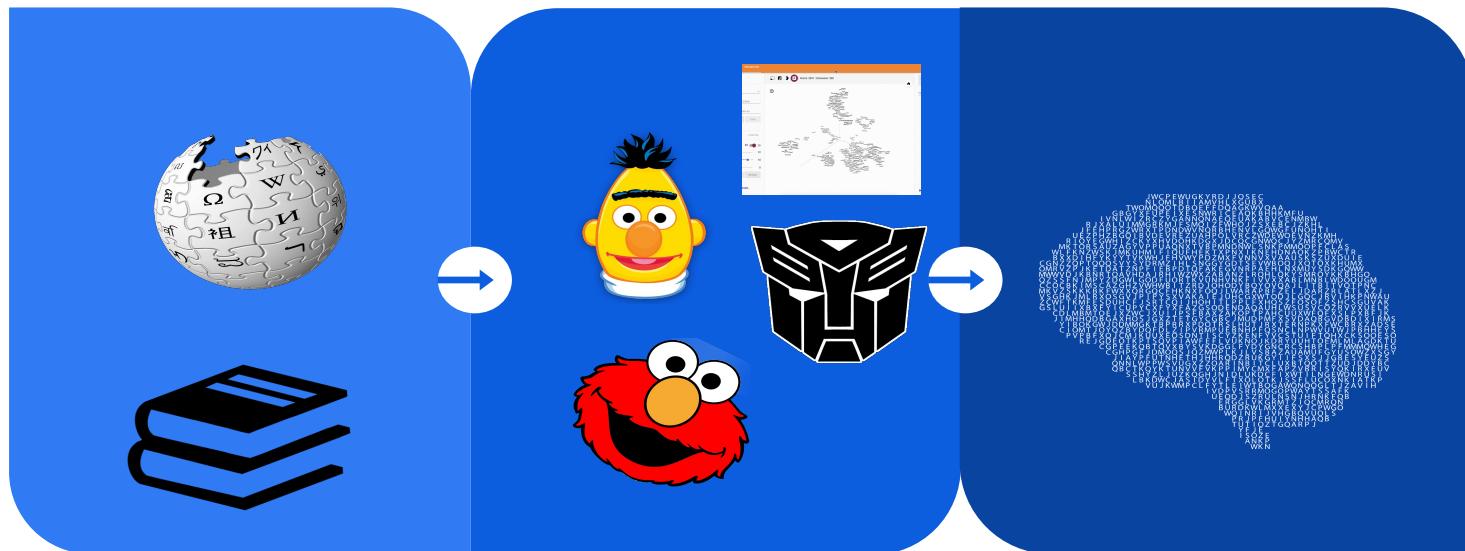


Learn a new
representation

(Modern) Natural Language Processing: 2019



(Modern) Natural Language Processing: 2019



Wikidata, Books,
Open Source Dataset

word2vec, BERT,
ELMo, ULMFiT,
Transformer Net

Language Model

(Deep) Natural Language Processing

“NLP’s ImageNet moment has arrived” (Sebastian Ruder:
<https://thegradient.pub/nlp-imagenet/>)

}

Word2vec and related methods are shallow approaches that trade expressivity for efficiency.

Using word embeddings is like initializing a computer vision model with pretrained representations that only encode edges: they will be helpful for many tasks, but they fail to capture higher-level information that might be even more useful. A model initialized with word embeddings needs to learn from scratch not only to disambiguate words, but also to derive meaning from a sequence of words. This is the core aspect of language understanding, and it requires modeling complex language phenomena such as compositionality, analogy, and

In order to predict the most probable next word in a sentence, a model is required not only to be able to express syntax (the grammatical form of the predicted word must match its modifier or verb) but also model semantics. Even more, the most accurate models must incorporate what could be considered world knowledge or common sense. Consider the incomplete sentence "The service was poor, but the food was". In order to predict the succeeding word such as "yummy" or "delicious", the model must not only memorize what attributes are used to describe food, but also be able to identify that the conjunction "but" introduces a contrast, so that the new attribute has the opposing sentiment of "poor".

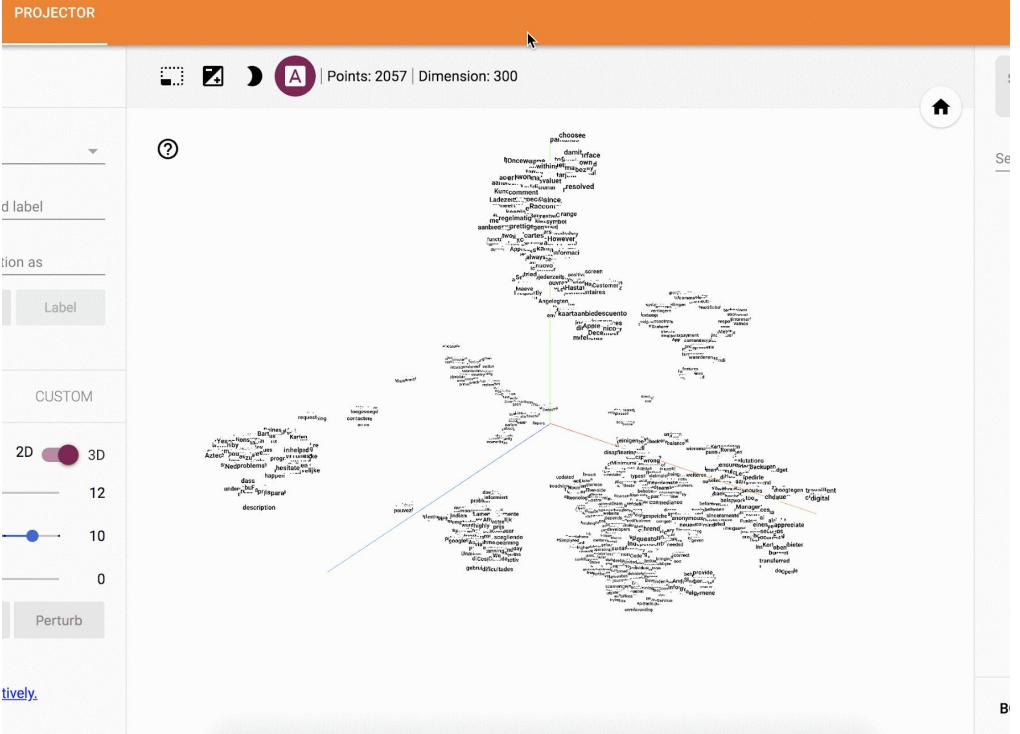
W
O

transformer Net

Language Model

Journey of Modern NLP

Pre-Trained: word2vec [2013]



Efficient Estimation of Word Representations in Vector Space

<https://arxiv.org/> > cs

by T Mikolov - 2013 - Cited by 9937 - Related articles

[PDF] [Distributed Representations of Words and Phrases and their ...](#)

Thu, 7 Nov 2013

<https://arxiv.org/pdf/1310.4546.pdf>

by T Mikolov - 2013 - Cited by 12042 - Related articles

[PDF] [Linguistic Regularities in Continuous Space Word Representations](#)

com/pwot/1310.4546.pdf

<https://www.aclweb.org/anthology/N13-1090.pdf>

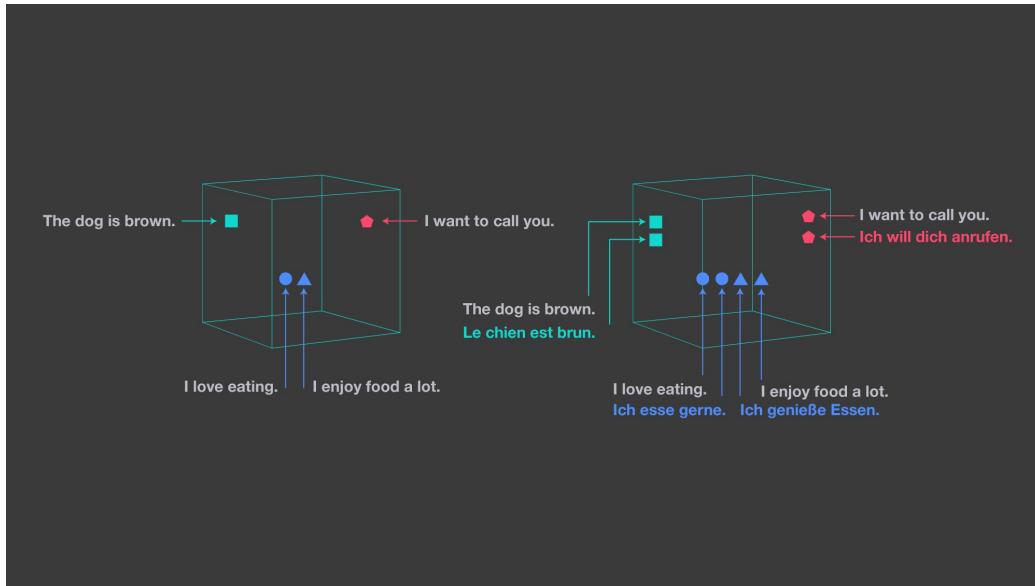
by T Mikolov - 2013 - Cited by 1969 - Related articles

9 Jun 2013 - Linguistic Regularities in Continuous Space Word Representations. Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig. Microsoft Research.

Key takeaways:

1. Skip-gram neural language model.
2. Mapped words to vectors (embeddings) which were part of the trained weights of the net.
3. Showed that learned vector spaces organized to preserve linear relationships between concepts, eg $v(\text{"king"}) - v(\text{"man"}) + v(\text{"woman"}) = v(\text{"queen"})$.
4. In the process of training, similar words ended up near each other (similar vectors). HUGE!

Pre-Trained: word2vec [2014] (Cross-lingual embeddings)



[PDF] Exploiting Similarities among Languages for Machine Translation

<https://arxiv.org/pdf/1309.4168.pdf>

by T Mikolov - 2013 - Cited by 631 - Related articles

17 Sep 2013 · Translating between multiple languages with one model and many more features

BilBOWA: Fast Bilingual Distributed Representations without Wor...

<https://arxiv.org/statistics/>

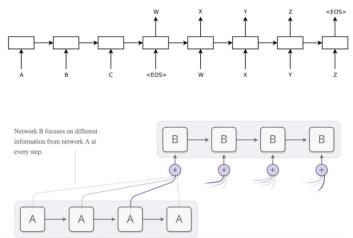
by S Gouws - 2014 - Cited by 223 - Related articles

Abstract: We introduce BilBOWA (Bilingual Bag-of-Words without Alignments), a simple and computationally-efficient model for learning bilingual distributed ...

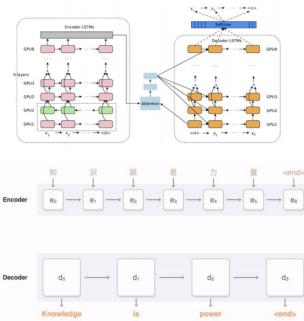
Key takeaways:

1. Noticed that word vectors in one language can be projected to another using learned affine mapping.
2. Suggested a new way for doing word-by-word translation: nearest neighbours in the aligned spaces!
3. Also suggested easy method for **transfer learning**: train on EN embedding space, and apply classifier directly to aligned FR space. Voila!

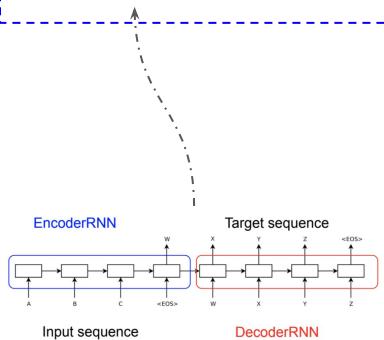
Fastforward



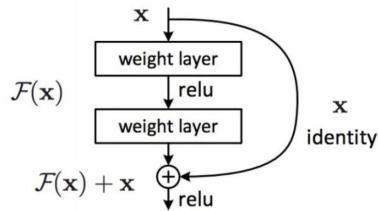
The skip connections allows information to flow unmitigated through the network (related to LSTM error carousels, but not gated).



seq2seq [2014] -----> seq2seq + Attention[2015] -----> ResNets[2015] -----> Google NMT[2016] ----->



This was huge. Soon, LSTMs+Attention became the default for almost any task that can be cast as seq2seq.



Deep LSTM encoder-decoder with residual connections and attention. Boosted translation from 58% to 87%.

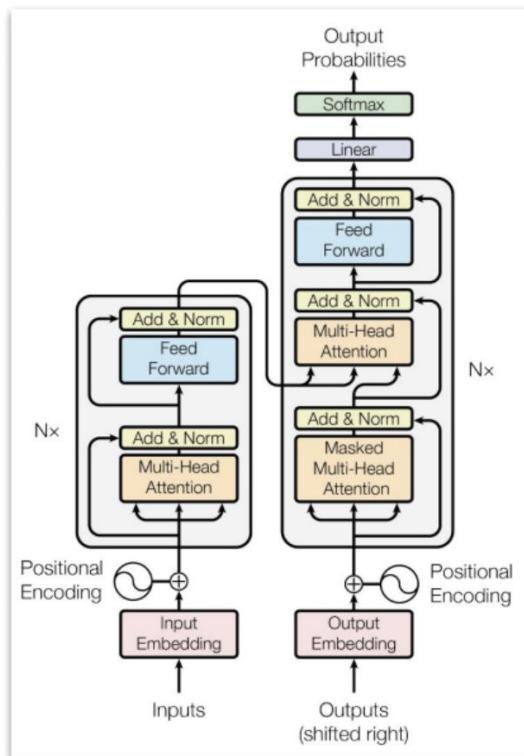
Pre-Trained: Transformer[2017]

Attention Is All You Need

<https://arxiv.org> > cs

by A Vaswani - 2017 - Cited by 2925 - Related articles

Jun 12, 2017 - The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network ...



Key takeaways:

1. Parallel-in-time encoder-decoder model with blocks of self-attention interleaved with feedforward blocks.
2. Introduced Multi-Head Self-Attention which allows multimodal attention.
3. Fast to train, as all encoder blocks for all sequence positions are performed in parallel (as opposed to sequentially for RNNs).
4. Started **huge new wave** of “self-attentive” models.

Pre-Trained: GPT-2 [2018]

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

 Follow @AdamDanielKing for updates and other demos like this one.

Custom prompt

I am going to attend TechBBQ

GENERATE ANOTHER

Completion

I am going to attend TechBBQ !!!!!!!

This is a serious event that requires some serious skills. Get ready to be prepared and stay hydrated.

A large group of people, who come together to have fun and learn

We are not asking you to "work" for all those hours you'll be taking. We're asking you to keep at it and give us all a good time

Come and have a great time!

Key takeaways:

1. Pretrain large 12-layer left-to-right Transformer.
2. Fine tune for sentence, sentence-pair and multiple choice questions.
3. SOTA results for 9 tasks.

DEMO LINK:
<https://talktotransformer.com/>
Built by [Adam King \(@AdamDanielKing\)](#)

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever
OpenAI OpenAI OpenAI OpenAI
alec@openai.com karthikn@openai.com tim@openai.com ilya@openai.com



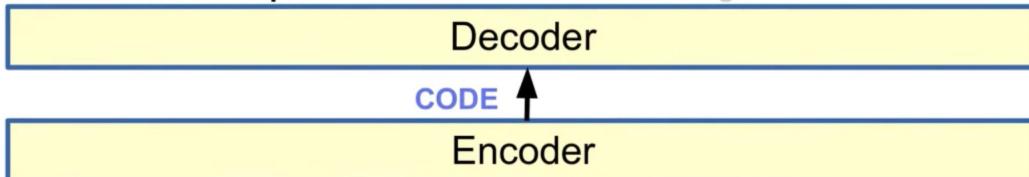
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

(Submitted on 11 Oct 2018 (v1), last revised 24 May 2019 (this version, v2))

Pre-Trained: BERT [2019]

OUTPUT: This is a piece of text extracted from a large set of news articles



INPUT: This is a [...] of text extracted [...] a large set of [...] articles

Key takeaways:

1. BERT pretrains both sentence and contextual word representations, using masked LM and next sentence prediction.
2. LM which can looks both forward and backwards (**learning both left and right context**)
3. BERT-large has 340M parameters, 24 layers!

Semi-Supervised Bot

DEMO LINK {WIP}:

<http://35.228.74.197:3031/>

Building blocks of (Deep) NLP

Non-Linearities



ReLU
Sigmoid
Tanh
GRU
LSTM
Linear
...

Connectivity Pattern



Fully connected
Convolutional
Dilated
Recurrent
Self Attention
Recursive
Skip / Residual
Random

Optimizer

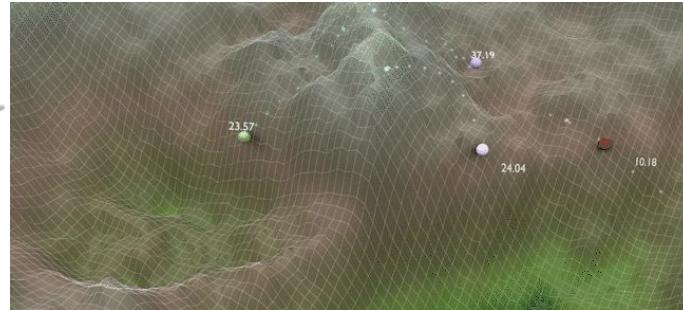


SGD
Momentum
RMSProp
Adagrad
Adam
Second Order (KFac)
...

Loss



Cross Entropy
Adversarial
Variational
Max. Likelihood
Sparse
L2 Reg
REINFORCE
...



Hyper Parameters

Learning Rate
Decay
Layer Size
Batch Size
Dropout Rate
Weight init
Data augmentation
Gradient Clipping
Beta
Momentum

Deep`NLP Overview

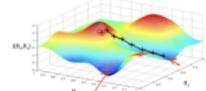
Ideas / Tricks

Soft inputs (Word embeddings), Soft state (RNNs), **Gated connections**, Computation graph/Autodiff, Non-saturating non-linearities, Gradient clipping, Cross-lingual Embeddings, Encoder-decoder (seq2seq), Skip-connections, Word-pieces, Distillation, Self-Attention



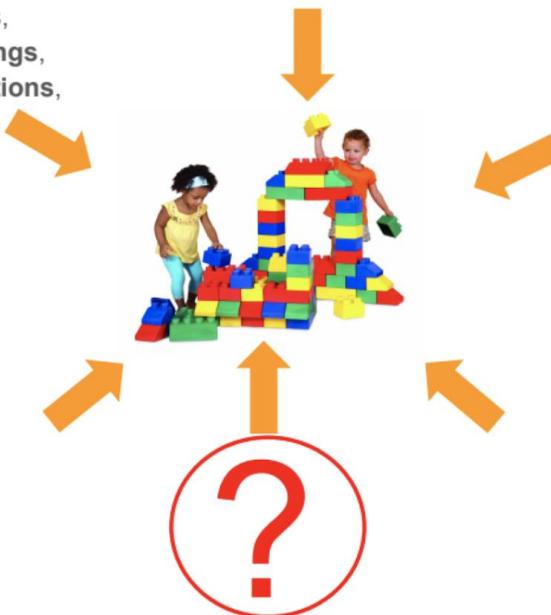
Optimizers

SGD, Backpropagation, BPTT, Momentum, Adam; better initialization



Layers

Embedding, Dropout, Residual Blocks, LayerNorm, MH-Self-Attention,



Architectures

FFNs, RNNs, ConvNets, LSTMs/GRUs, Transformers



Frameworks & Platforms

Theano, TensorFlow, PyTorch, etc.

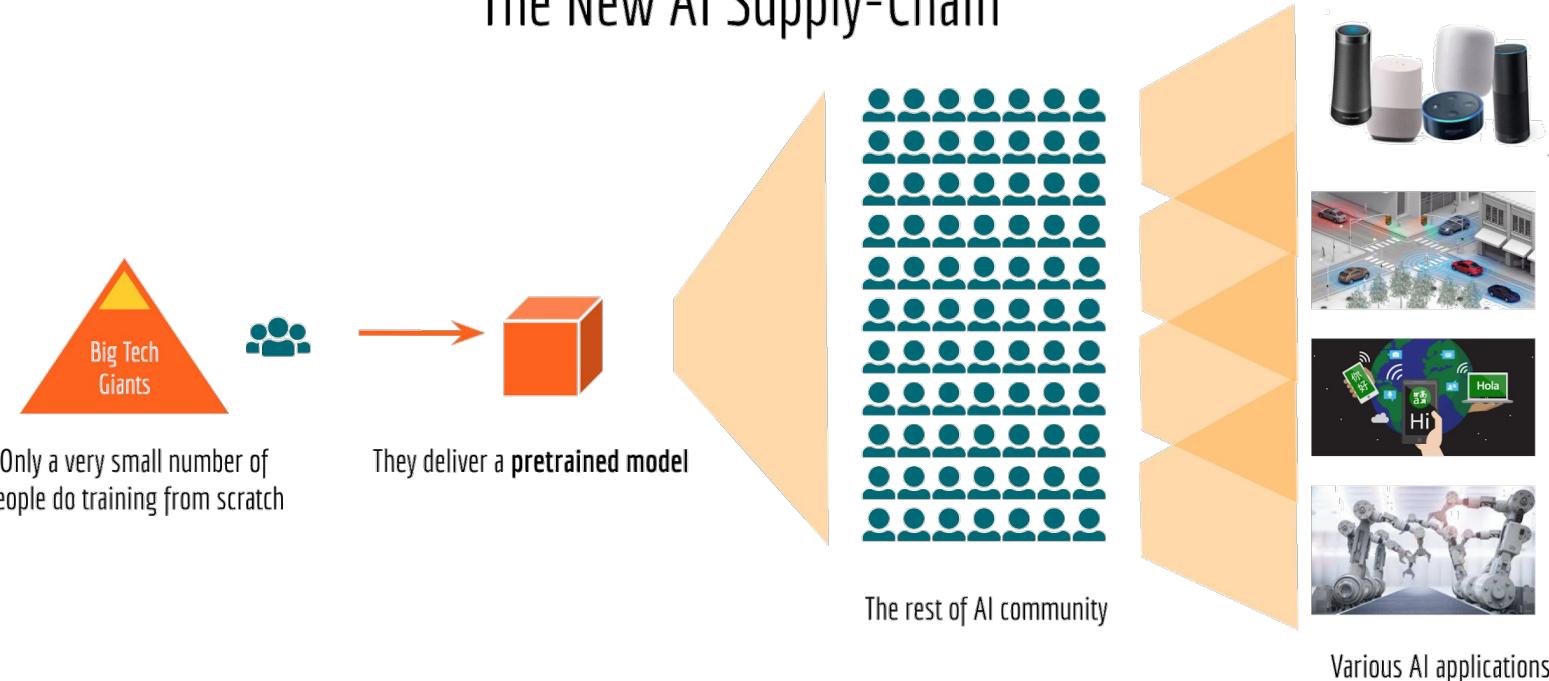


Faster CPUs, GPUs, TPUs; easier deployment to accelerators (write-once, run-anywhere)

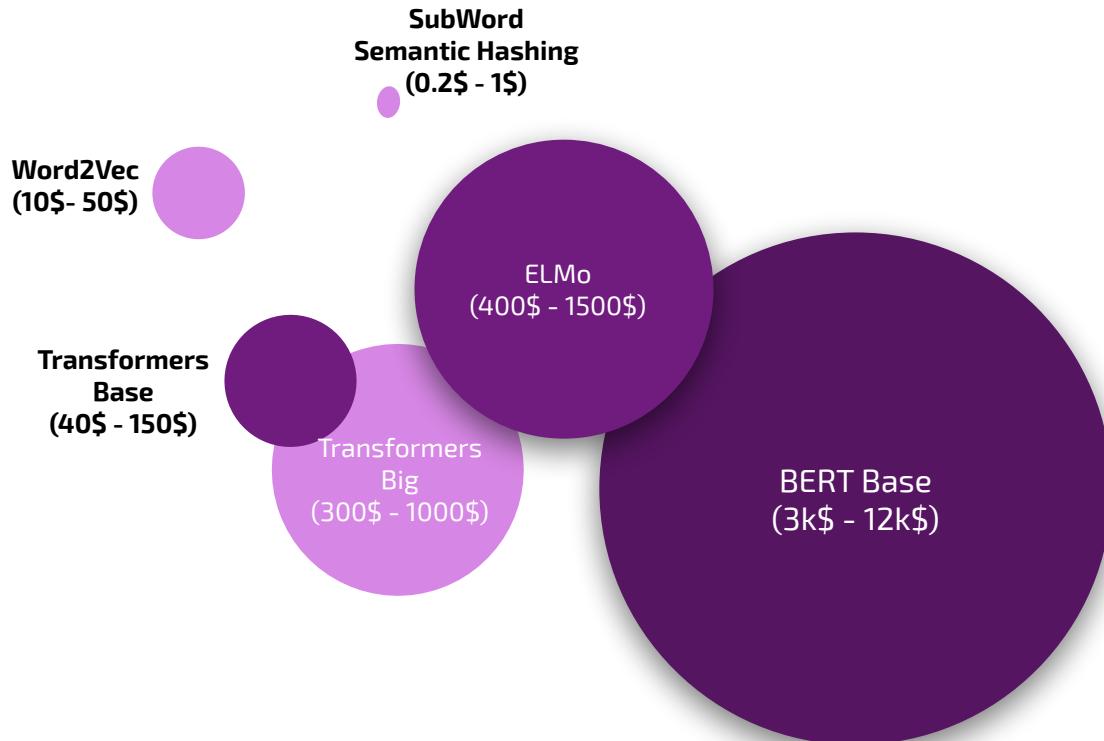
But?

Modern AI Ecosystem

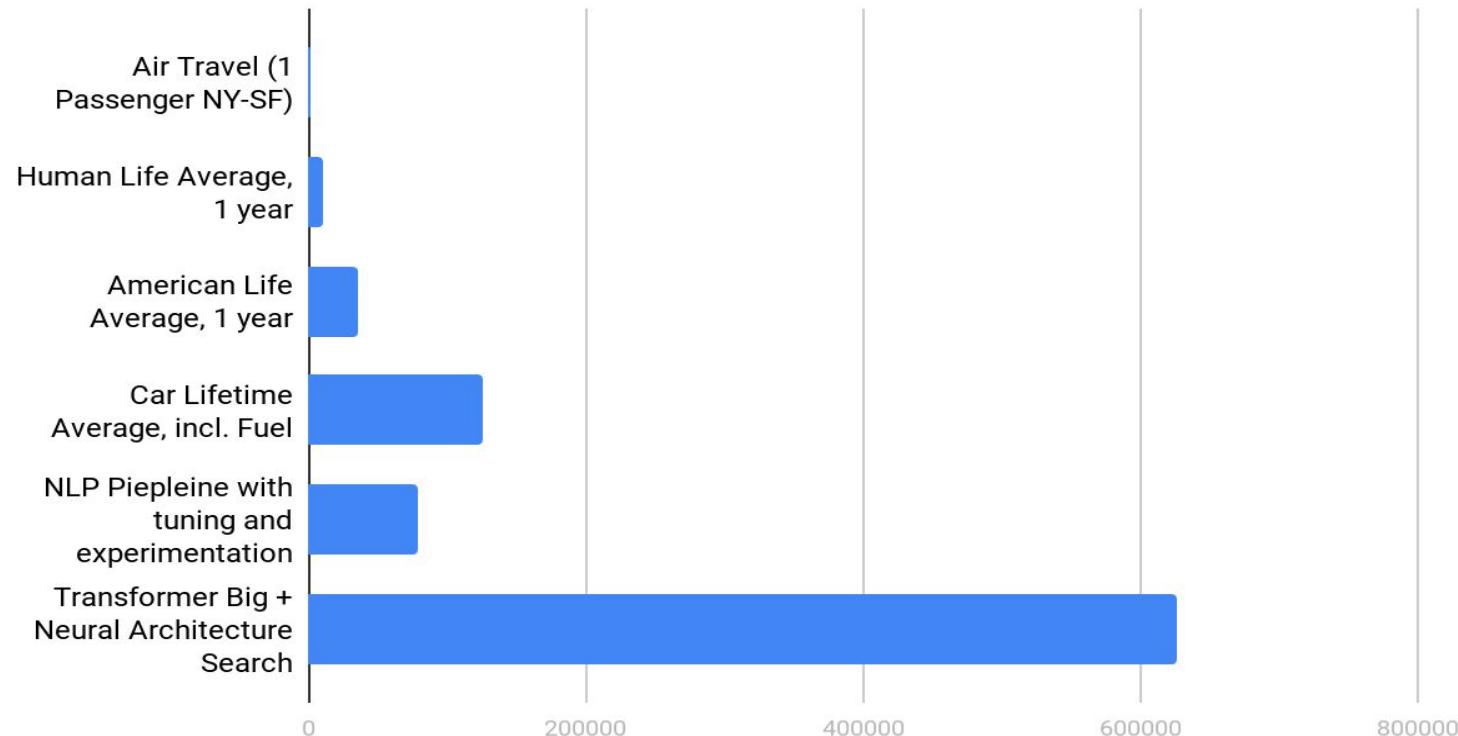
The New AI Supply-Chain



Cloud Compute Training Cost



Estimated Carbon Dioxide Emission Cost



Source: [Energy and Policy Considerations for Deep Learning in NLP](#)

Our Approach:

Meta Learning (learning to learn)

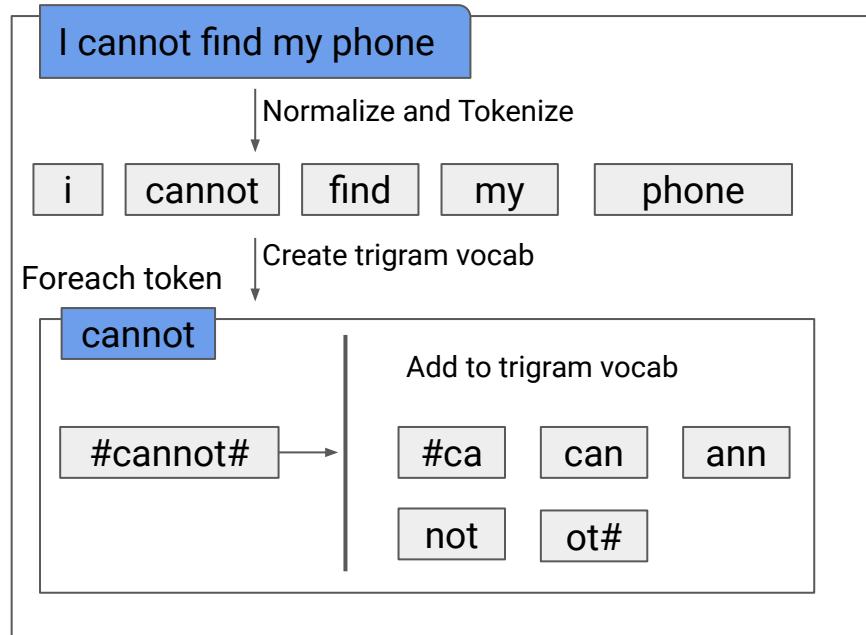
Traditional
Programming

Machine
Learning

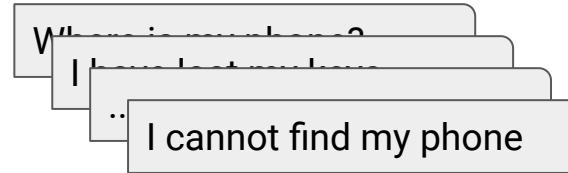
Overview

Creating meta-knowledge

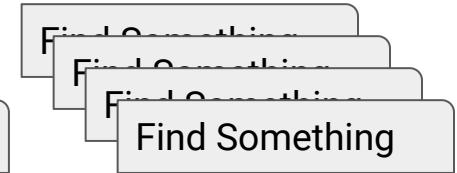
Foreach Sample



Dataset



Samples



Labels

Trigram Vocabulary

Pick a Vectorizer

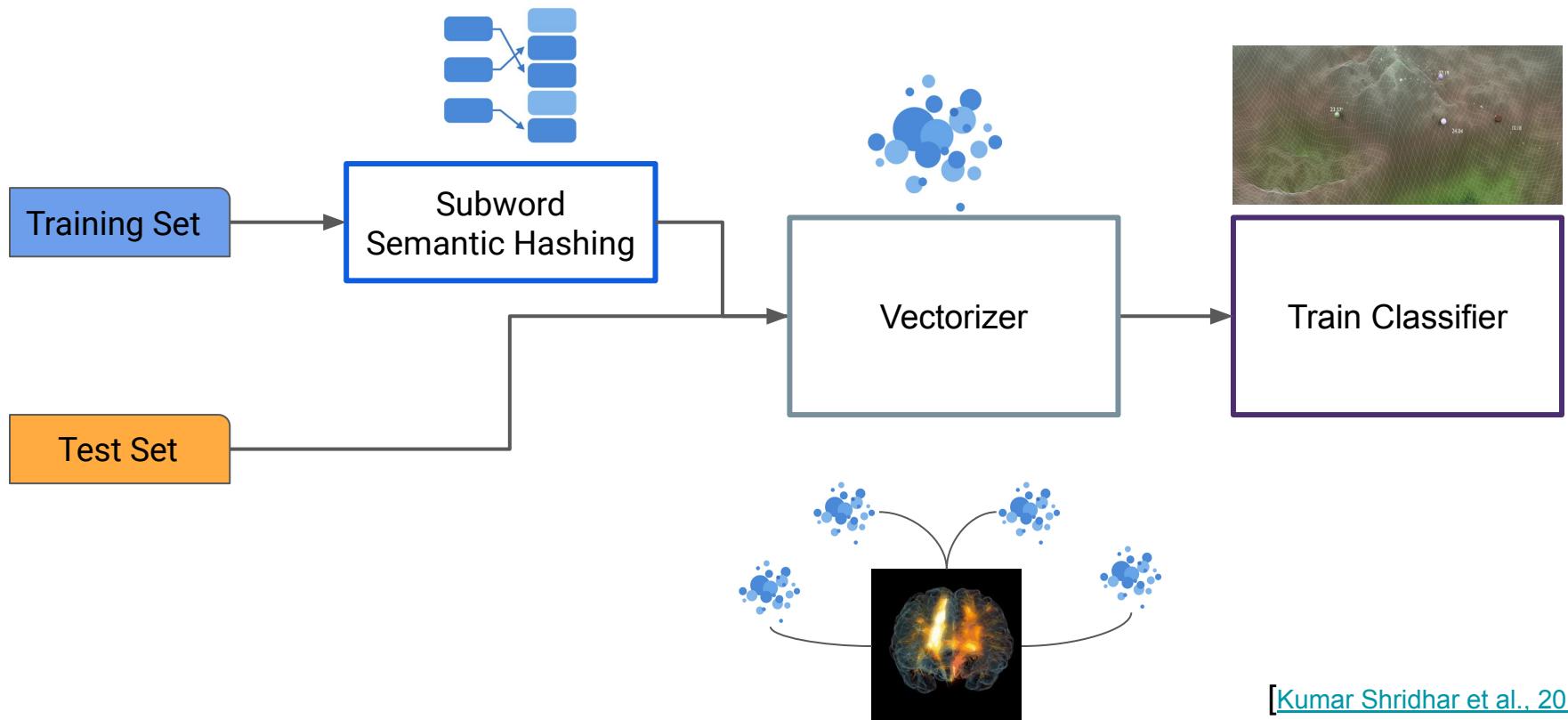
Count Vectorizer

TF-IDF Vectorizer

BPE Embeddings

...

Where Semantic Hashing Fits?



[Kumar Shridhar et al., 2018]

Accuracy comparison against various frameworks

Platform	Chatbot	AskUbuntu	WebApp	Overall	Average
Botfuel	0.98	0.90	0.80	0.91	0.89
Luis	0.98	0.90	0.81	0.91	0.90
Dialogflow	0.93	0.85	0.80	0.87	0.86
Watson	0.97	0.92	0.83	0.91	0.91
Rasa	0.98	0.86	0.74	0.88	0.86
Snips	0.96	0.83	0.78	0.89	0.86
Recast	0.99	0.86	0.75	0.89	0.87
TildeCNN	0.99	0.92	0.81	0.92	0.91
Our Average	0.98	0.92	0.83	0.92	0.91
Our Best	0.99	0.93	0.85	0.93	0.92
<i>Our Individual Best</i>	1.00	0.93	0.86	0.94	0.93

STAGES

Proprietary Framework

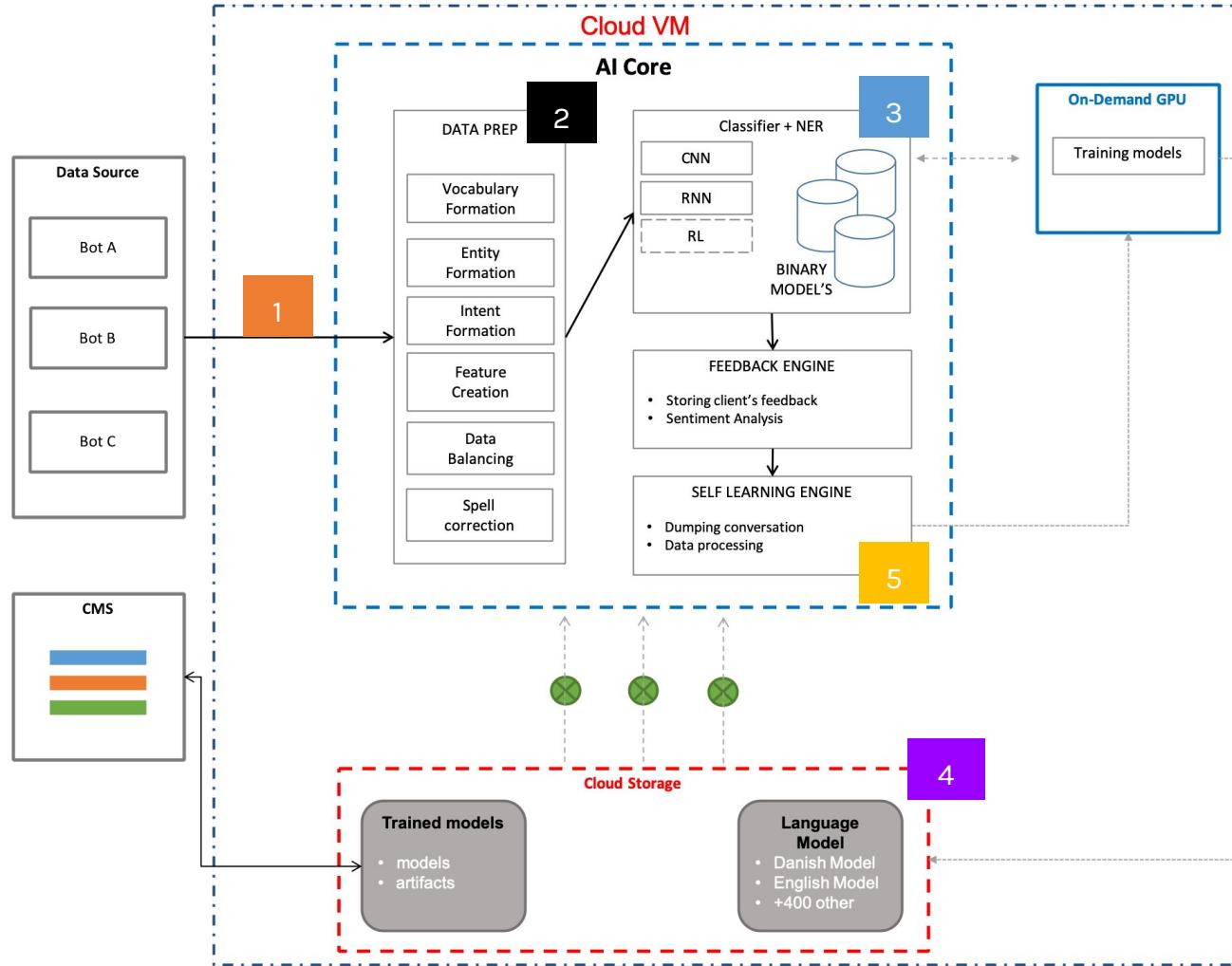
1. Data Collection

2. Data Processing

3. Model Training

4. Model Deployment

5. Continuous Learning



Natural Language Processing: The Goal

[Context: We saw some lions; guide tells us it's a rare sight. Few min later we pass a second safari guide.]



No current models can understand this situation.

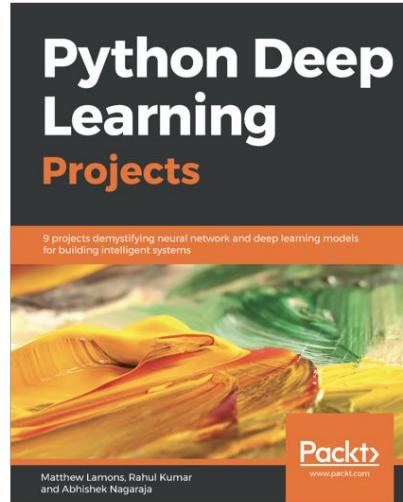
Likely need to integrate various sources of information in a multi-step reasoning process.

Vi så nogle løver

him

Which architectural, training, loss function, etc. advances do we need to get here?

Thank you



Book : <https://www.amazon.com/dp/1788997093>

Code : <http://bit.ly/DeepLearningCode>

@hellorahulk



goodrahstar



hellorahulk.com

