Deep Learning for Natural Language Processing

Said Al Faraby

Telkom University

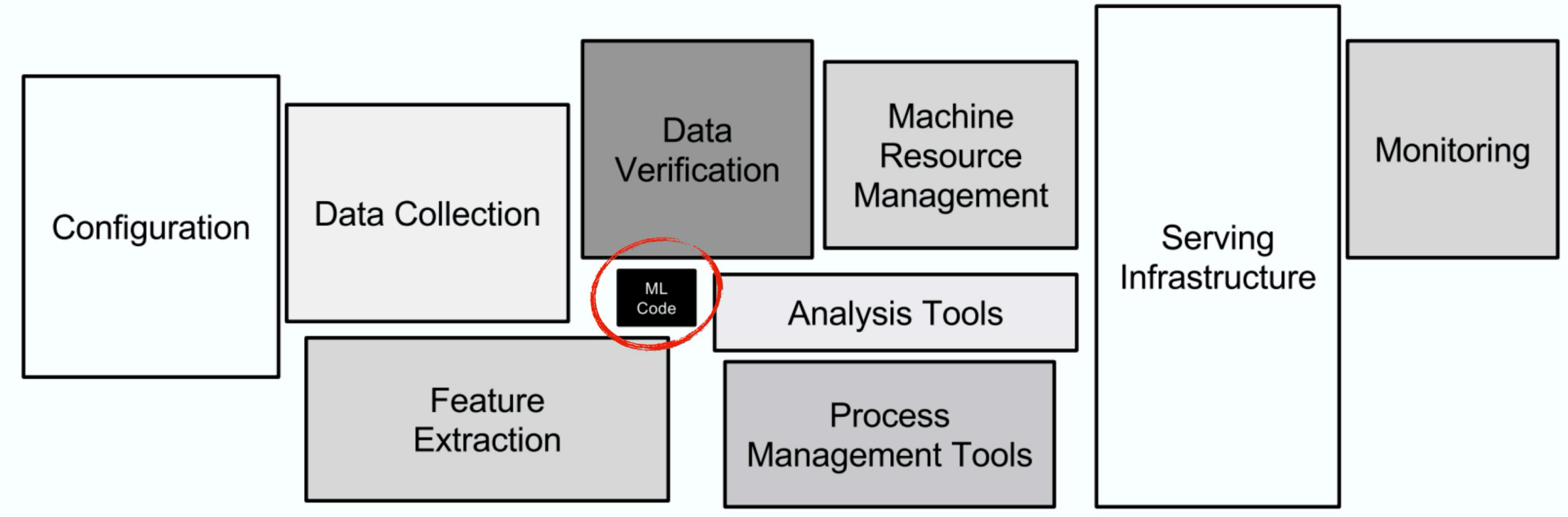
Colab Notebook

Colab

https://colab.research.google.com/drive/1LzDnQxFc1gi0W5nGy8Vgc1hbdxTKB2e7

Slide http://bit.ly/dl4nlp_tu

Full Stack Deep Learning



https://fullstackdeeplearning.com/

Natural Language Processing

Natural Language Processing

- A field at the intersection of
  - Computer science
  - Artificial Intelligence
  - Linguistics
- Goal: to process or understand natural language in order to perform some useful tasks
- **Fully understanding** and **representing meaning** of language is **impossible goal**.

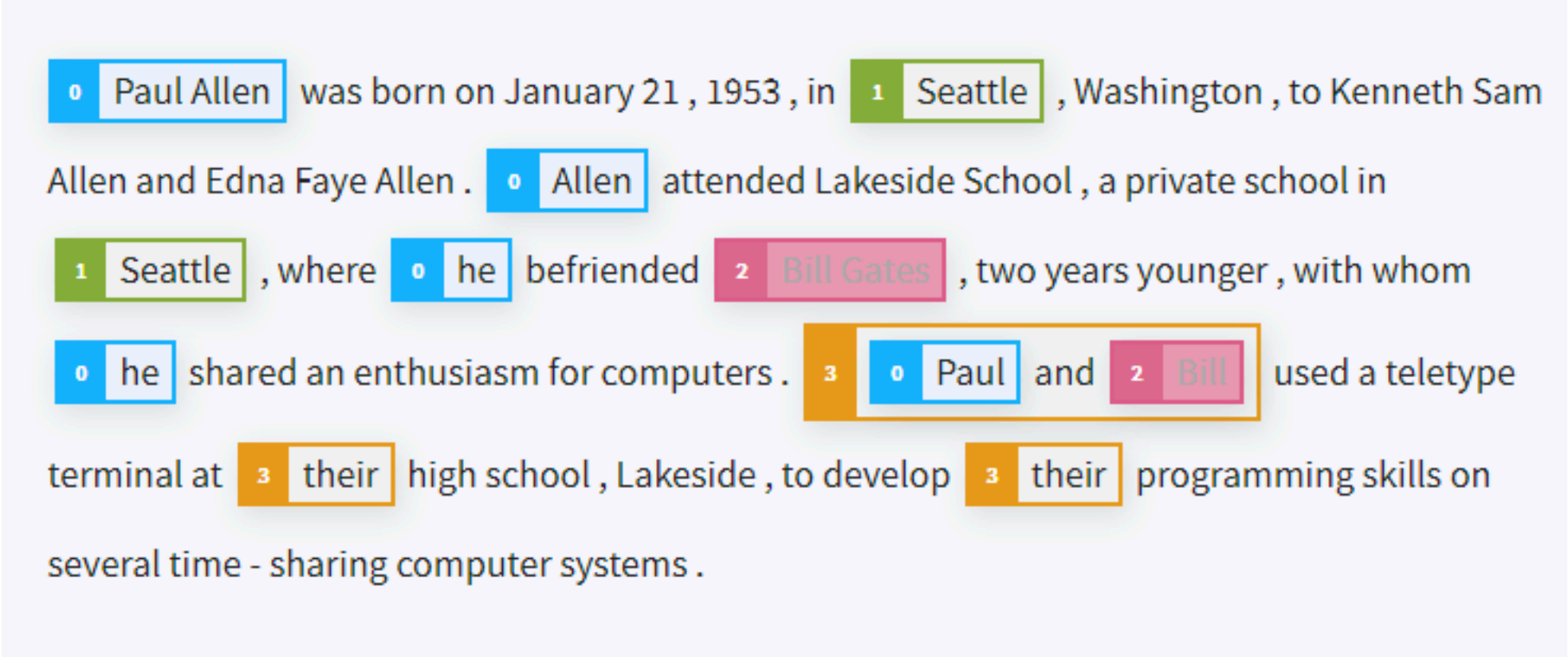
Adopted from Richard Sochar slides

Tasks in NLP

Automatic speech recognition	Lexical normalization	Semantic parsing
CCG	Machine translation	Semantic role labeling
Common sense	Missing elements	Sentiment analysis
Constituency parsing	Multi-task learning	Shallow syntax
Coreference resolution	Multi-modal	Simplification
Dependency parsing	Named entity recognition	Intent Detection and Slot Filling
Dialogue	Natural language inference	Stance detection
Domain adaptation	Part-of-speech tagging	Summarization
Entity linking	Question answering	Taxonomy learning
Grammatical error correction	Relation prediction	Temporal processing
Information extraction	Relationship extraction	Text classification
Language modeling	Semantic textual similarity	Word sense disambiguation

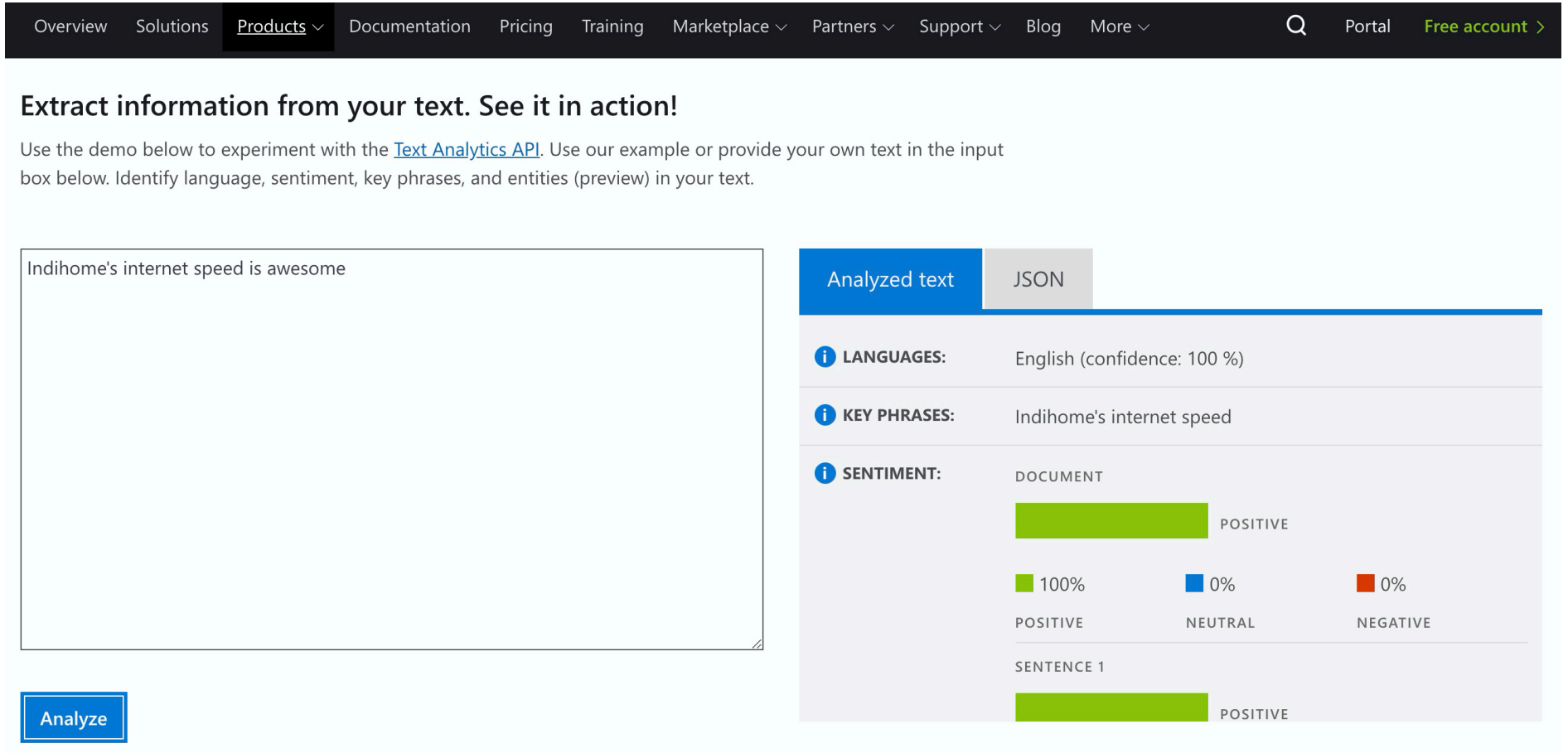
Adopted from Richard Sochar slides

Co-Reference Resolution



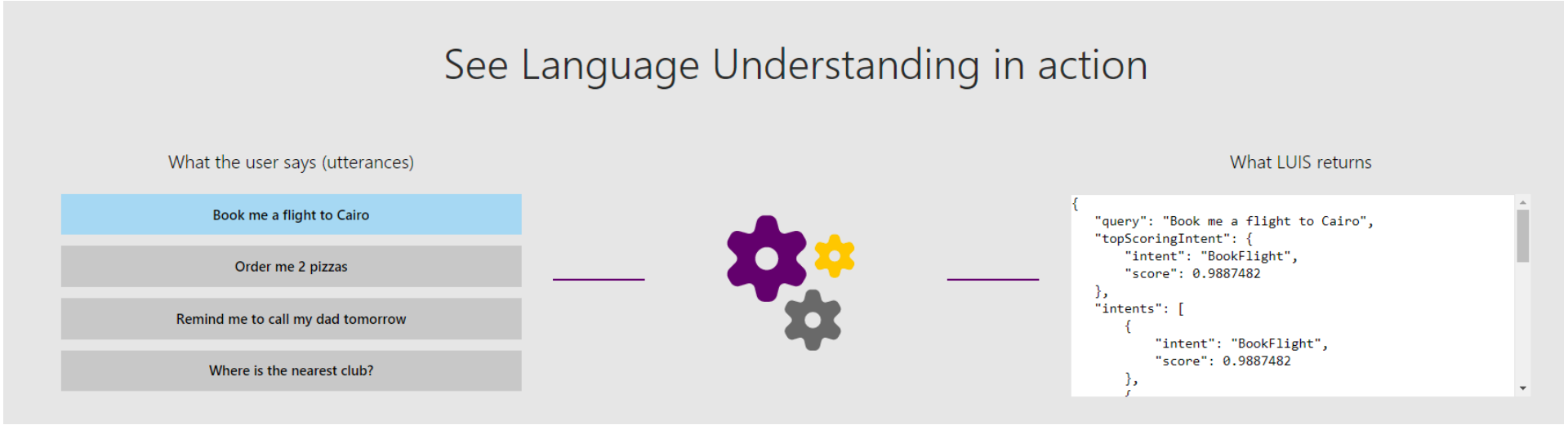
- http://demo.allennlp.org/coreference-resolution

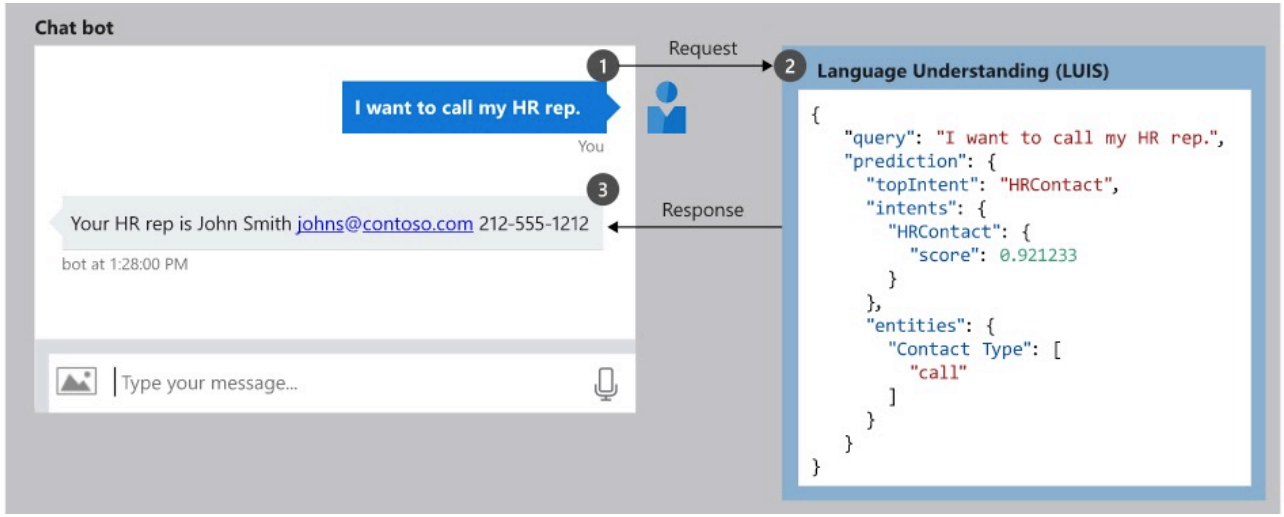
Sentiment Analysis/Text Classification



https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/?WT.mc_id=blog-medium-abornst

Intent Classification





https://luis.ai/home, https://dialogflow.com/

Intent to Chat Bot

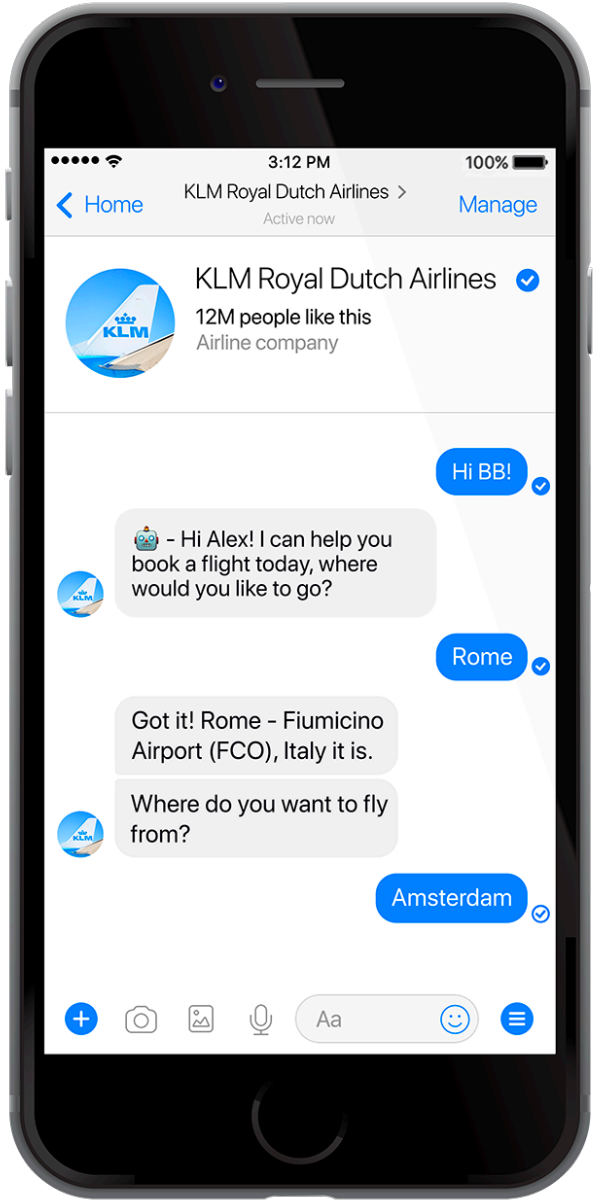
Hi, welcome to Domino's. Would you like to place an order or track an order?

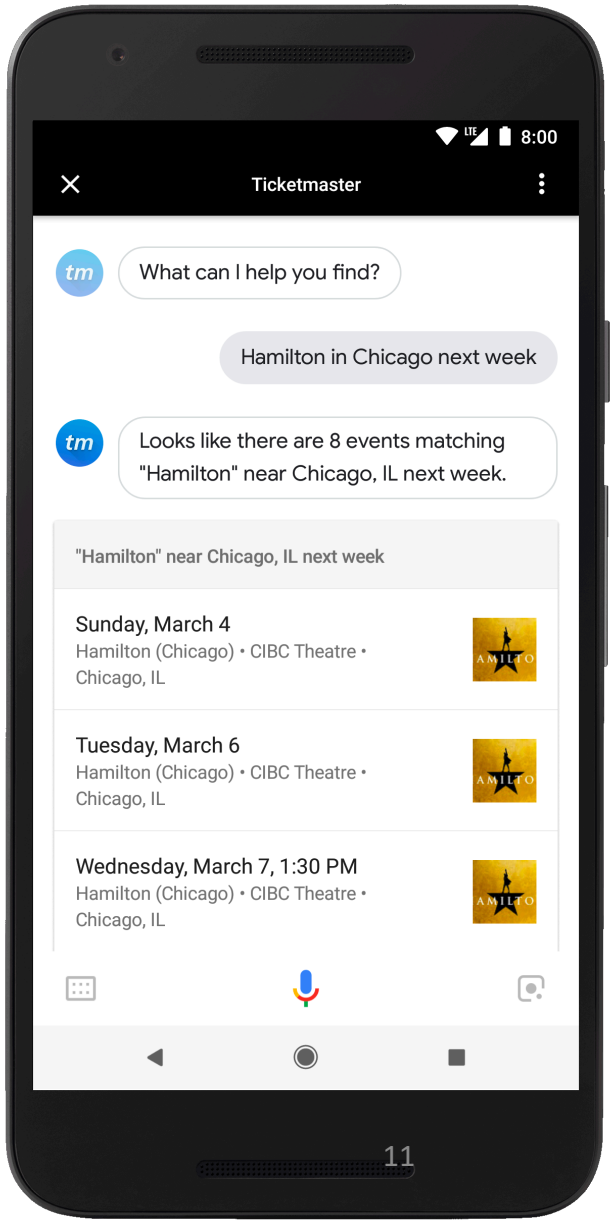
Place an order

Awesome! Do you want to place your Easy Order, most recent order, or a new order?

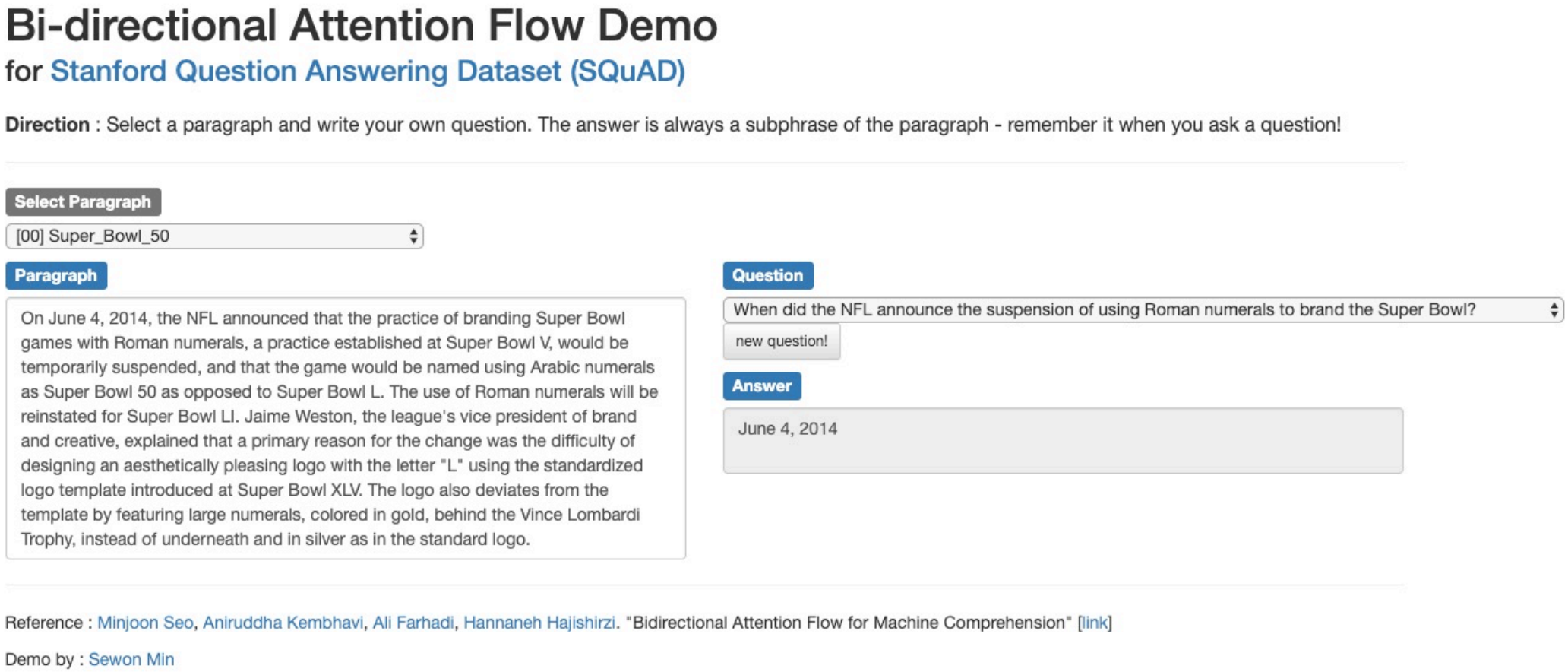
New order

Great, let's get started!





Machine Reading Comprehension



https://allenai.github.io/bi-att-flow/

Machine Translation

TRANSLATE

Found in translation: More accurate, fluent sentences in Google Translate

Barak Turovsky
Product Lead, Google Translate

Published Nov 15, 2016

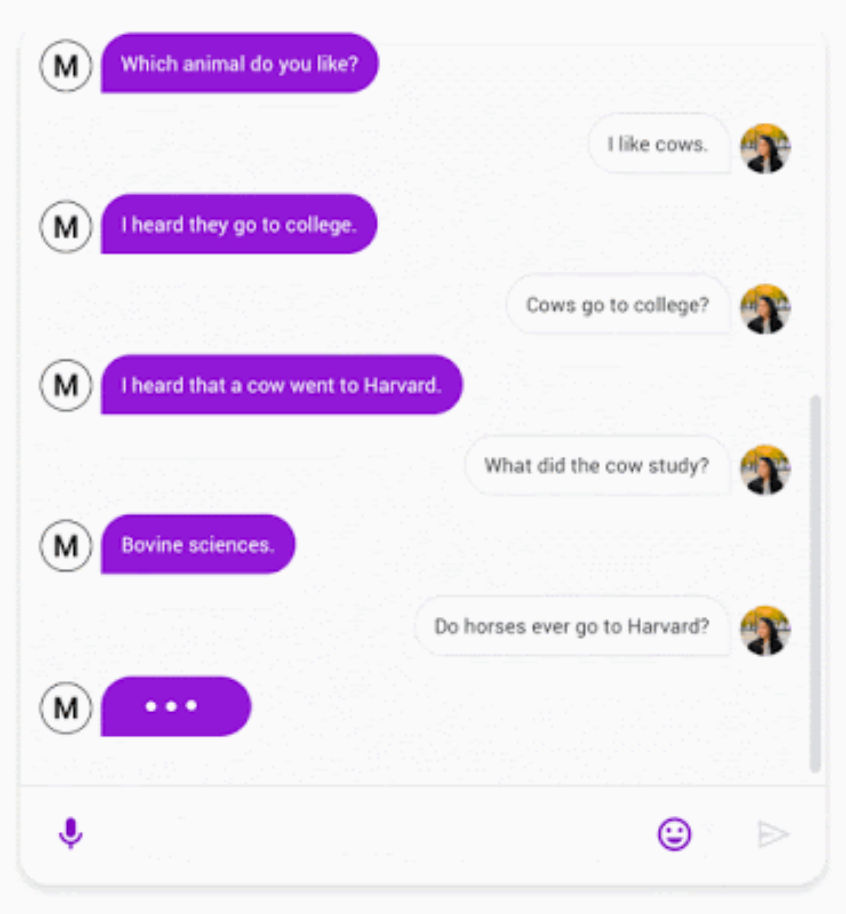
In 10 years, Google Translate has gone from supporting just a few languages to 103, connecting strangers, reaching across language barriers and even helping people find love. At the start, we pioneered large-scale statistical machine translation, which uses statistical models to translate text. Today, we're introducing the next step in making Google Translate even better: Neural Machine Translation.

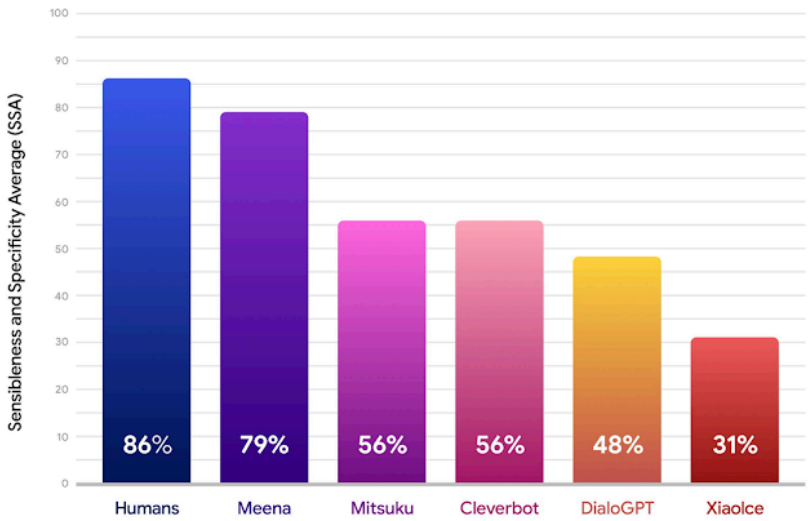
Neural Machine Translation has been generating exciting research results for a few years and in September, our researchers announced Google's version of this technique. At a high level, the Neural system translates whole sentences at a time, rather than just piece by piece. It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts to be more like a human speaking with proper grammar. Since it's easier to understand each sentence, translated paragraphs and articles are a lot smoother and easier to read. And this is all possible because of end-to-end learning system built on Neural Machine Translation, which basically means that the system learns over time to create better, more natural translations.

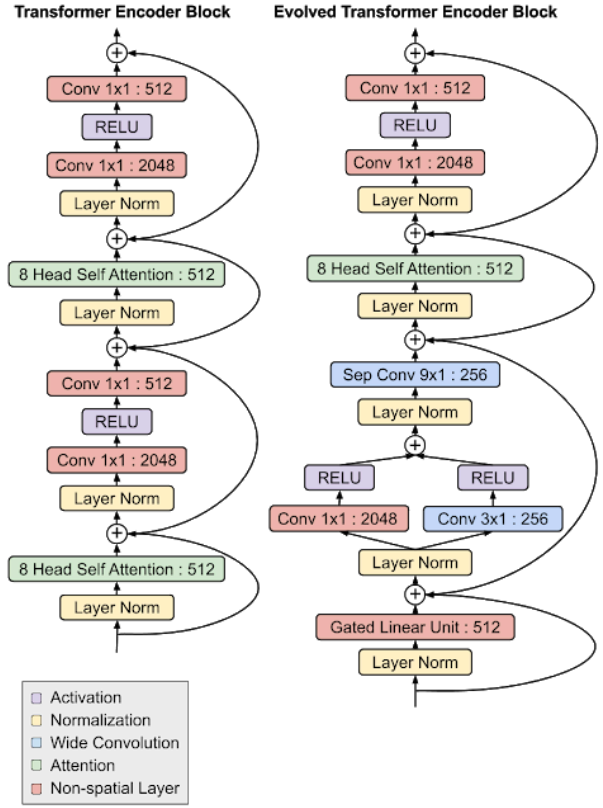
Today we're putting Neural Machine Translation into action with a total of eight language pairs to and from English and French, German, Spanish, Portuguese, Chinese, Japanese, Korean and Turkish. These represent the native languages of around one-third of the world's population, covering more than 35% of all Google Translate queries!

https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Conversational Agent

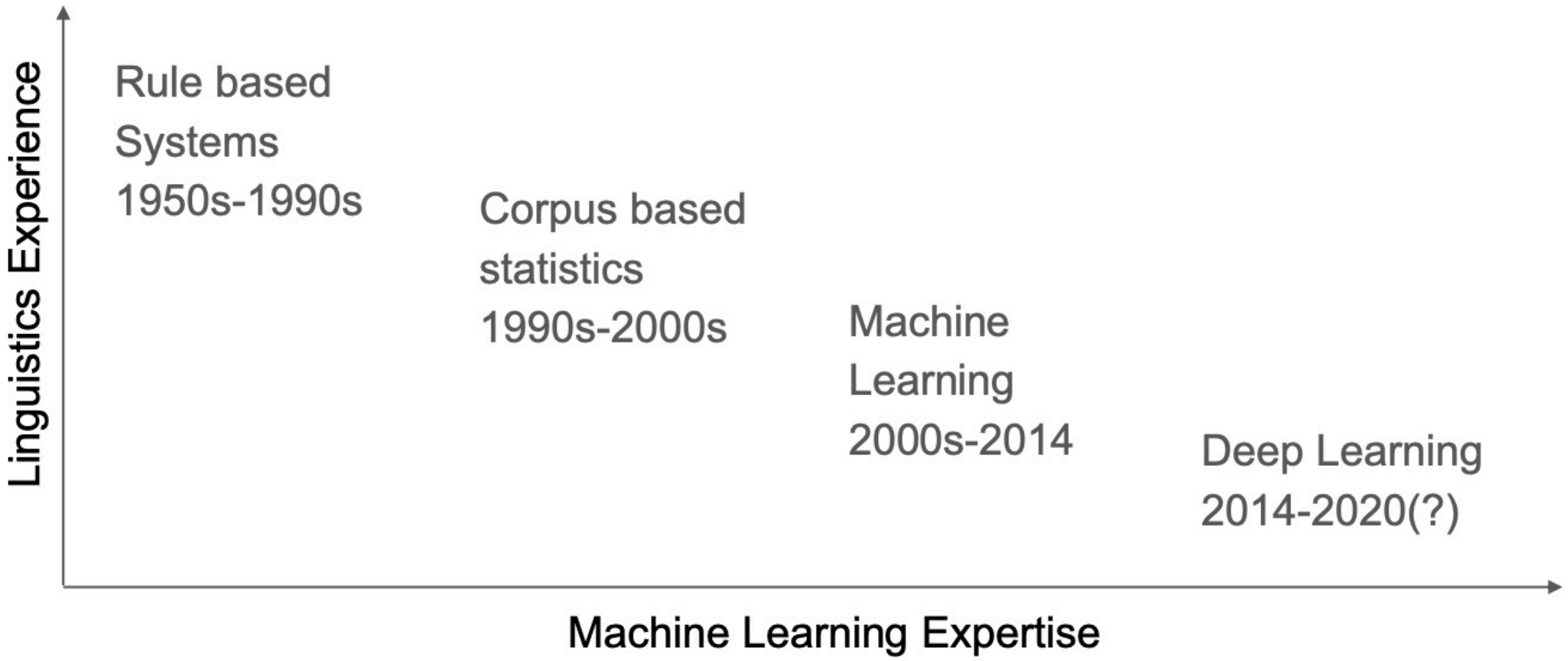






https://towardsdatascience.com/inside-the-machine-learning-that-google-used-to-build-meena-a-chatbot-that-can-chat-about-anything-32e4d2242f79

Approaches in NLP



Classic Machine Learning for NLP

- In Machine Learning for NLP, human need to design good features and representations to make it works well.
- Need help from domain experts (e.g Linguist)
- So, the Machine Learning methods just optimize the weights for final predictions.

- For example: features for finding named entity (Finkel, 2010)

Feature	NER
Current Word	✓
Previous Word	✓
Next Word	✓
Current Word Character n-gram	all
Current POS Tag	✓
Surrounding POS Tag Sequence	✓
Current Word Shape	✓
Surrounding Word Shape Sequence	✓
Presence of Word in Left Window	size 4
Presence of Word in Right Window	size 4

Deep Learning for NLP

Representation Learning: automatically learn good representation of language (e.g chars/words/sentences)

+

“Machine Learning”: predict the outputs for the given task

Why Explore Deep Learning for NLP?

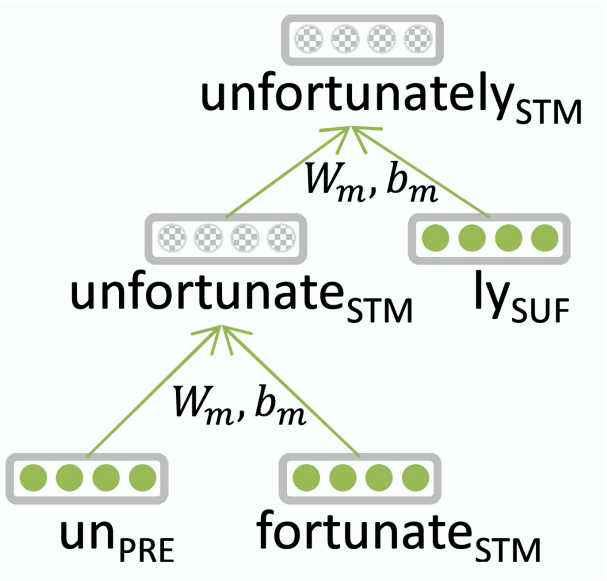
- Representing features is a hard task, while learned features are easy to adapt, and fast to learn
- Deep learning provides a very flexible framework to learn representation, including **visual**, **auditory**, and **textual** information.
- Support from data availability and computing powers (e.g. GPU)
- It works really good (e.g outperform non-DL methods almost in every task)

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
+	1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
	2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5	75.2/90.8	91.2	90.8	96.1	90.9	94.5	49.4
	4	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	6	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	7	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	9	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
	10	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	11	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	12	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	13	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	14	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	15	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	16	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
+	17	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6

https://gluebenchmark.com/leaderboard

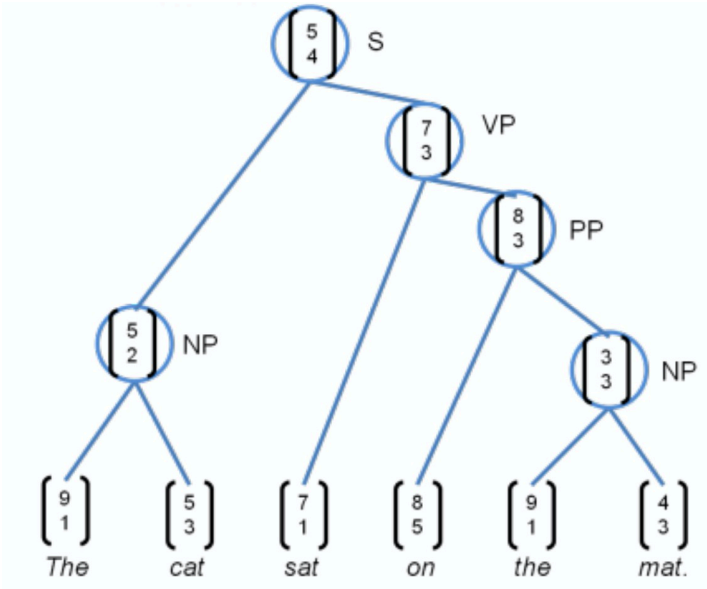
Representations Level

- Morphology



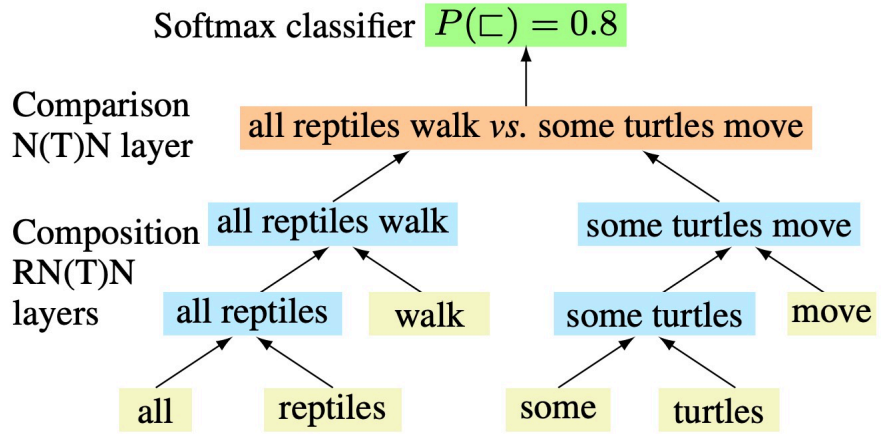
Thang et al. 2013

- Syntax



Socher et al. 2011

- Semantic



Bowman et al. 2014

Word Representation

What is Meaning?

Definition of Meaning (Meriam-Webster):

- the thing one intends to convey especially by language (word, phrase, etc)
- the thing that is conveyed especially by language (word, phrase, etc)

How to Represent Meaning in Computer?

Common answer:

- using Taxonomy like WordNet

Problem:

- hard to maintain
- subjective
- hard to compute accurate word similarity

One-hot Representation

teman = 

Problem:

- Dimension: 20k (speech) – 50k (Penn TB) – 500k (big vocab) – 13m (Google Web 1T)
- No information on similarity (meaning)
  - kawan [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] AND
  - teman [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] = 0

https://nlp.stanford.edu/manning/talks/Simons-Institute-Manning-2017.pdf
http://jalammar.github.io/illustrated-word2vec/
https://ruder.io/word-embeddings-1/index.html

Distributional Similarity Based Representations

"You shall know a word by the company it keeps" (J. R. Firth 1957: 11)

Intuitions: Zellig Harris (1954):

- "oculist and eye-doctor … occur in almost the same environments"
- "If A and B have almost identical environments we say that they are synonyms."

A bottle of ***tesgüino*** is on the table
Everybody likes ***tesgüino***
Tesgüino makes you drunk
We make ***tesgüino*** out of corn.

- What is ***tesgüino ?***

https://web.stanford.edu/~jurafsky/slp3/

Word – Document Matrix

- Two words are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

https://web.stanford.edu/~jurafsky/slp3/

Word – Window Matrix

sugar, a sliced lemon, a tablespoonful of **apricot** preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first **pineapple** and another fruit whose taste she likened
well suited to programming on the digital **computer**. In finding the optimal R-stage policy from
for the purpose of gathering data and **information** necessary for the study authorized in the

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

https://web.stanford.edu/~jurafsky/slp3/

Problem with simple cooccurrence vector

Problem:

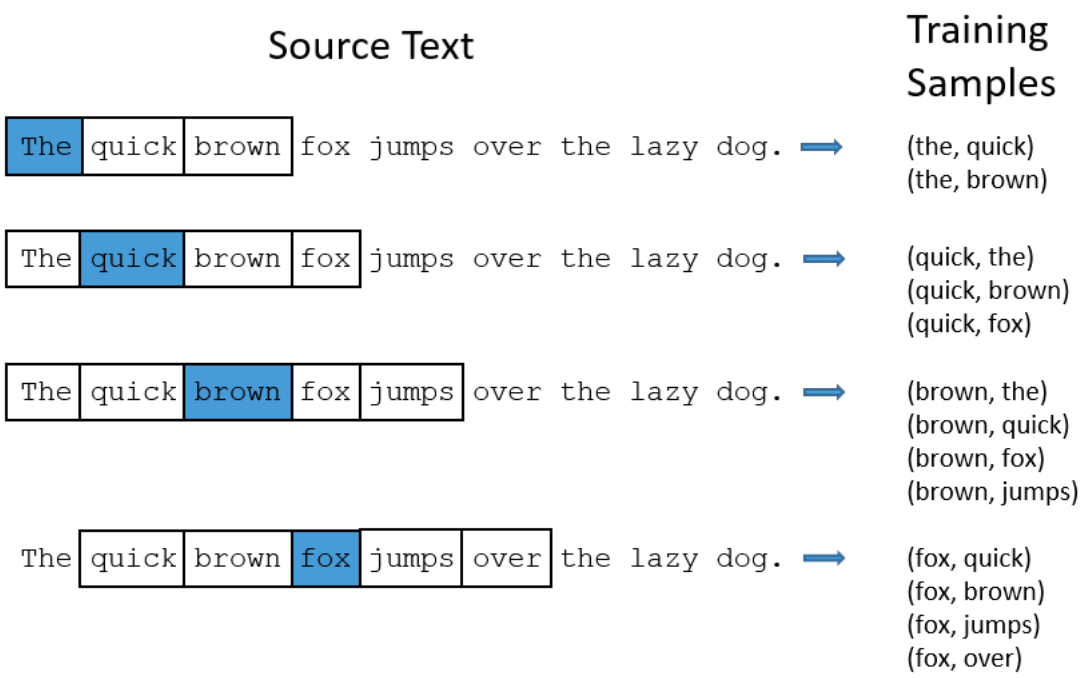
- Very high dimensional
- Vector dimension will increase with vocabulary size
- Sparsity issue (many zero values)

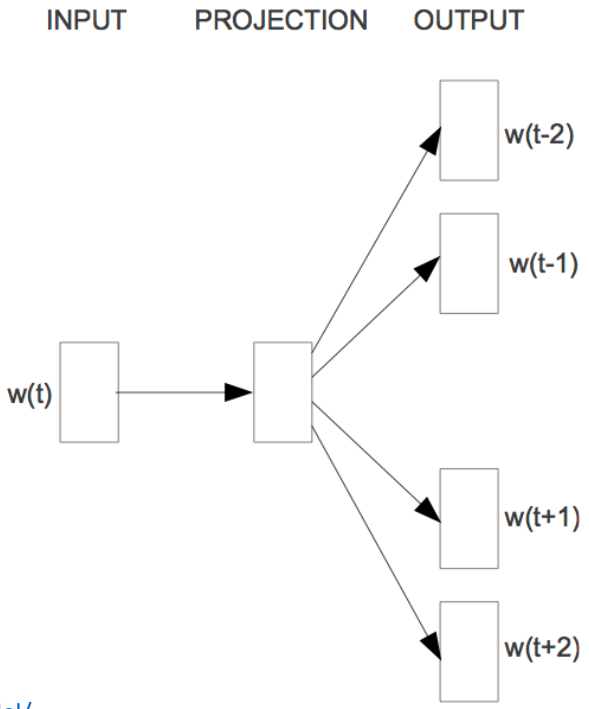
Solution ?

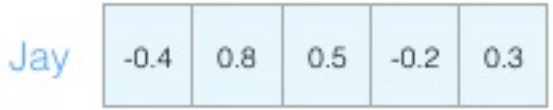
- Dimensionality reduction (e.g SVD)
- **Directly learn low-dimensional**/dense word vectors

Word Embedding (Word2Vec, Glove)

- Learn to predict surrounding words

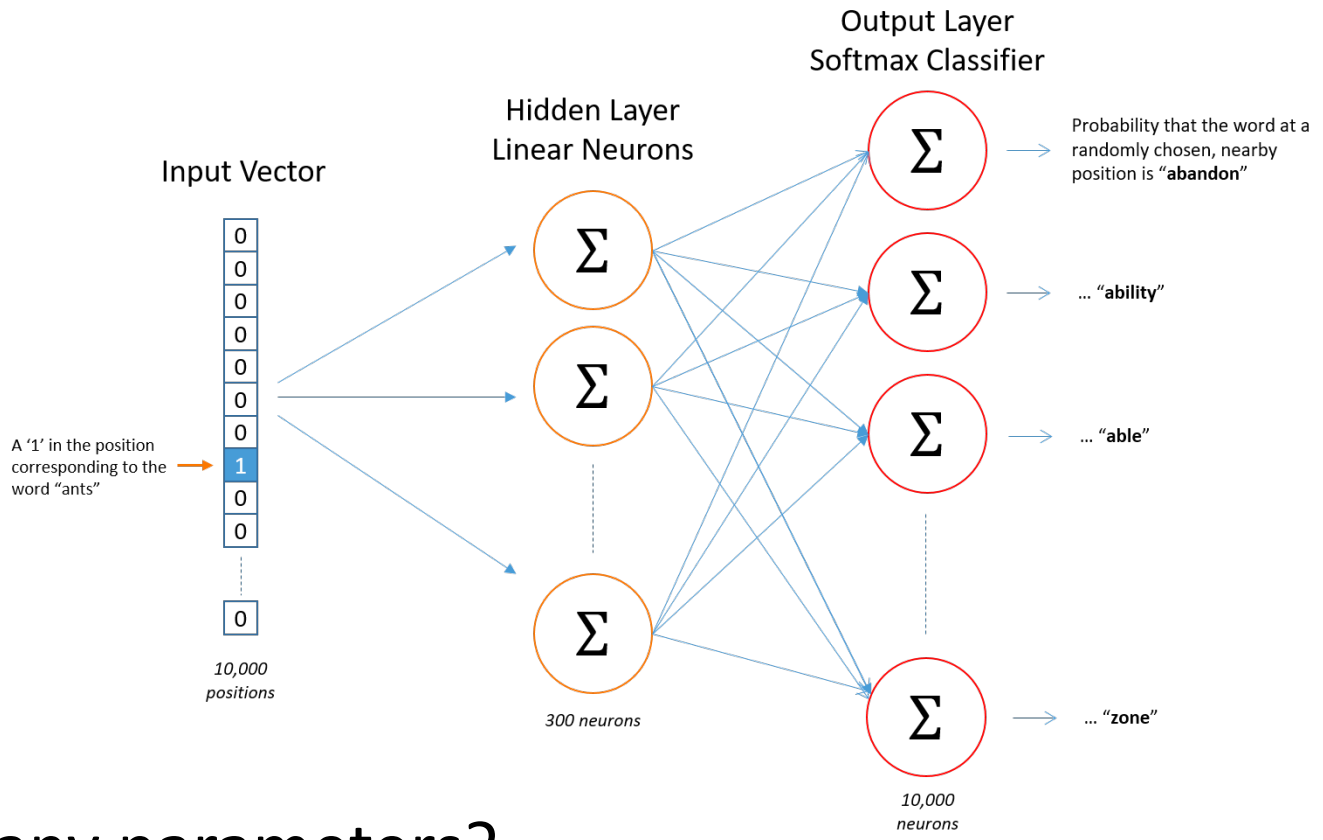






http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word2Vec Skip-Gram



- How many parameters?

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
https://towardsdatascience.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets-13445eebd281
https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/

The BiLSTM Hegemony

3. The BiLSTM Hegemony

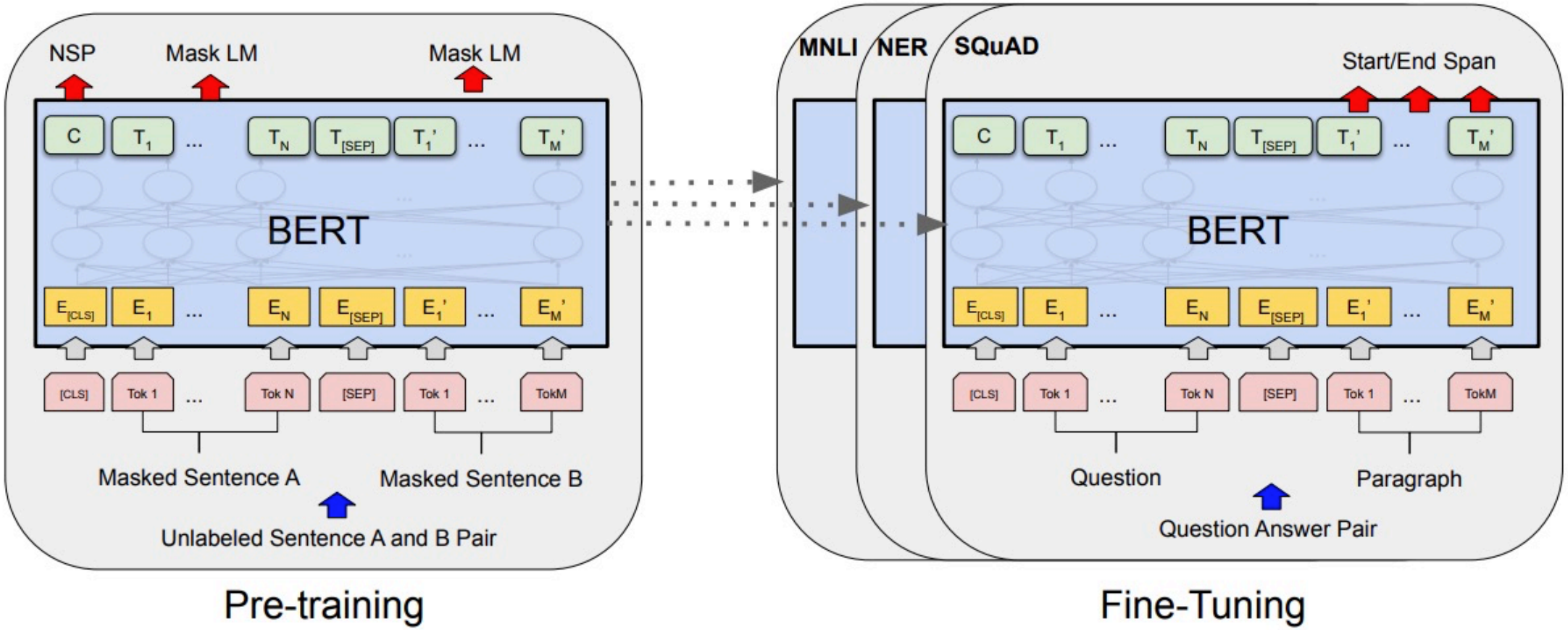
**To a first approximation,
the de facto consensus in NLP in 2017 is
that no matter what the task,
you throw a BiLSTM at it, with
attention if you need information flow**

BERT Hegemony

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Transfer Learning to Different NLP Task



What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (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.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 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.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Getting Started

We've built a few resources to help you get started with the dataset.

Download a copy of the dataset (distributed under the CC

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.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Jan 23, 2020	albert+transform+verify (ensemble) *qianxin*	89.528	92.059
5 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
6 Feb 20, 2020	Tuned ALBERT (ensemble model) *Group Data & Analytics Cell | Aditya Birla Group)* https://www.adityabirla.com/About/group-data-and-analytics	88.637	91.230
6 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	88.107	91.419

What is HotpotQA?

HotpotQA is a question answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. It is collected by a team of NLP researchers at Carnegie Mellon University, Stanford University, and Université de Montréal.

For more details about HotpotQA, please refer to our EMNLP 2018 paper:

(Yang, Qi, Zhang, et al. 2018)

Getting started

HotpotQA is distributed under a CC BY-SA 4.0 License. The training and development sets can be downloaded below.

Training set (535MB)

Dev set (distractor) (44MB)

Dev set (fullwiki) (45MB)

Test set (fullwiki) (46MB)

A more comprehensive summary about data download, preprocessing, baseline model training, and evaluation is included in our GitHub repository, and linked below.

Getting started guide

Once you have built your model, you can use the evaluation script we provide below to

Leaderboard (Distractor Setting)

In the *distractor* setting, a question-answering system reads 10 paragraphs to provide an answer (Ans) to a question. They must also justify these answers with supporting facts (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F_1	EM	F_1	EM	F_1
1 Dec 1, 2019	HGN-large (single model) *Anonymous*		**69.22**	**82.19**	**62.76**	**88.47**	**47.11**	**74.21**
2 Oct 18, 2019	C2F Reader (single model) *Joint Laboratory of HIT and iFLYTEK Research*		67.98	81.24	60.81	87.63	44.67	72.73
3 Nov 19, 2019	SAE-large (single model) *JD AI Research* Tu, Huang et al., AAAI 2020		66.92	79.62	61.53	86.86	45.36	71.45
4 Sep 27, 2019	HGN (single model) *Microsoft Dynamics 365 AI Research* Fang et al., 2019		66.07	79.36	60.33	87.33	43.57	71.03
5 Jul 29, 2019	TAP 2 (ensemble)		66.64	79.82	57.21	86.69	41.21	70.65
6 Oct 1, 2019	EPS + BERT(wwm) (single model) *Anonymous*		65.79	79.05	58.50	86.26	42.47	70.48
7 Jul 29, 2019	TAP 2 (single model)		64.99	78.59	55.47	85.57	39.77	69.12
8 May 31, 2019	EPS + BERT(large) (single model) *Anonymous*		63.29	76.36	58.25	85.60	41.39	67.92
9 Aug 31, 2019	SAE (single model) *JD AI Research* Tu, Huang et al., AAAI 2020		60.36	73.58	56.93	84.63	38.81	64.96
10 Jun 13, 2019	P-BERT (single model) *Anonymous*		61.18	74.16	51.38	82.76	35.42	63.79

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
+	1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
	2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5	75.2/90.8	91.2	90.8	96.1	90.9	94.5	49.4
	4	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	6	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	7	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	9	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
	10	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	11	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	12	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	13	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	14	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	15	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	16	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
+	17	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6

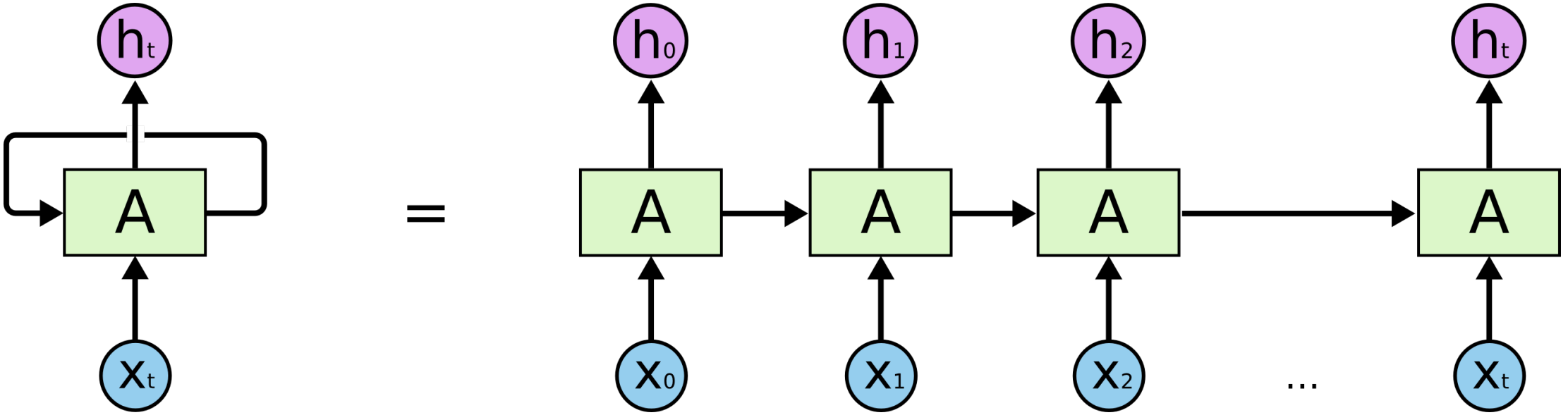
https://gluebenchmark.com/leaderboard

Deep Learning Architectures for NLP

- RNN -> LSTM
- Encoder-Decoder
- Attention
- Self-Attention
- Transformer
- BERT, RoBERTa, XLNet, GPT-2, T5

https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

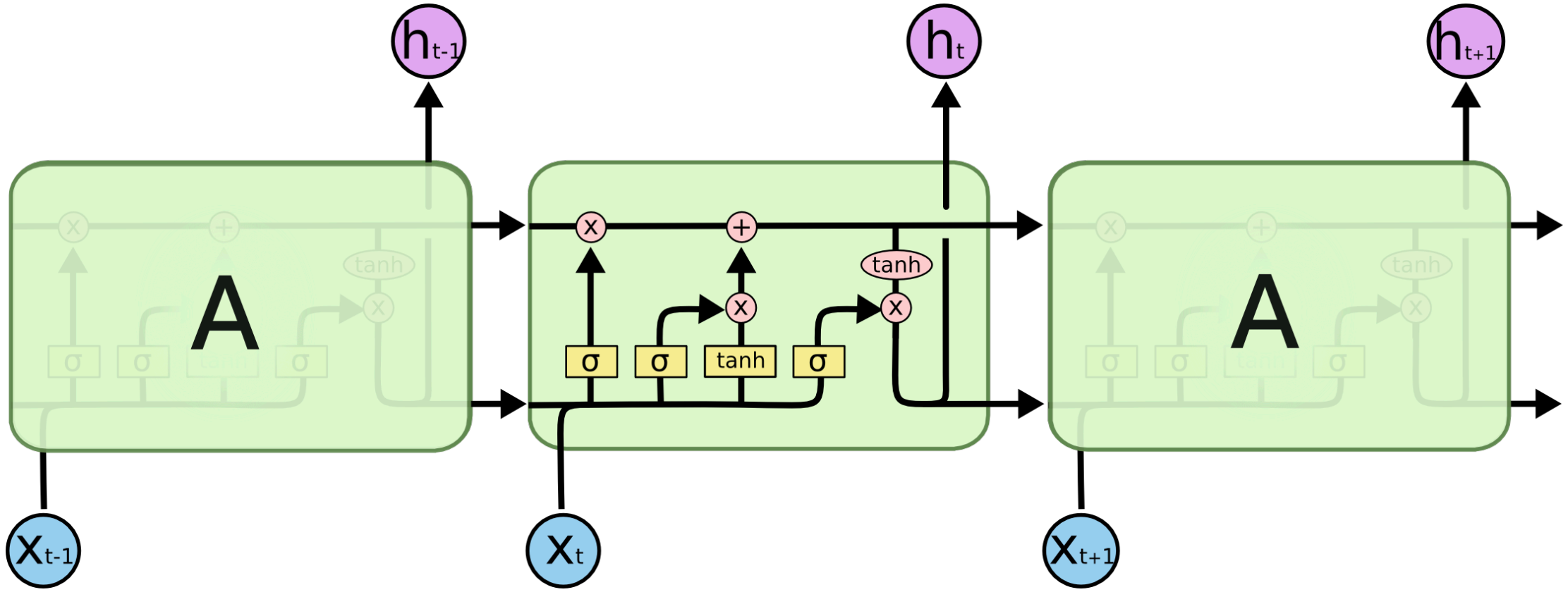
RNN



- Problem : Long Dependency

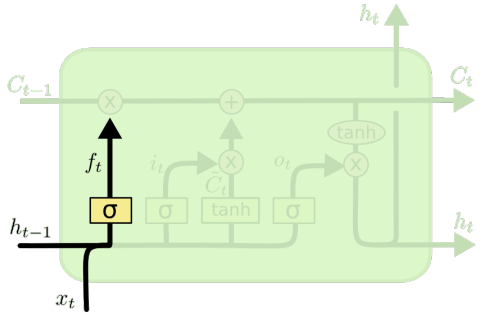
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM (Long Short Memory Network)

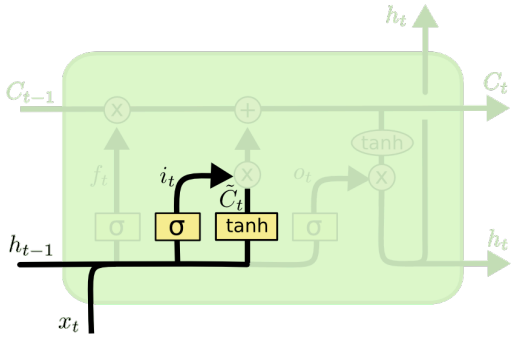


https://colah.github.io/posts/2015-08-Understanding-LSTMs/

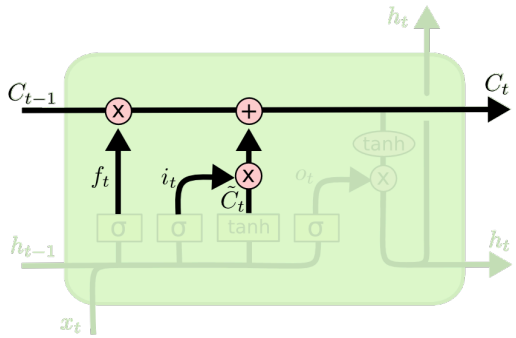
LSTM Gates



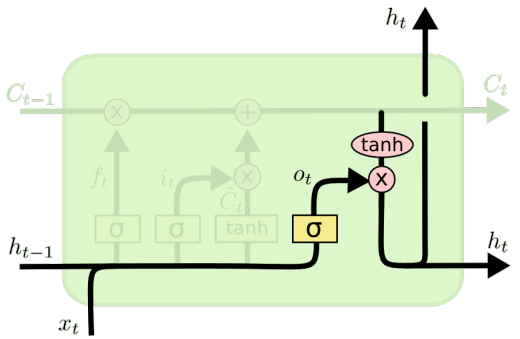
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

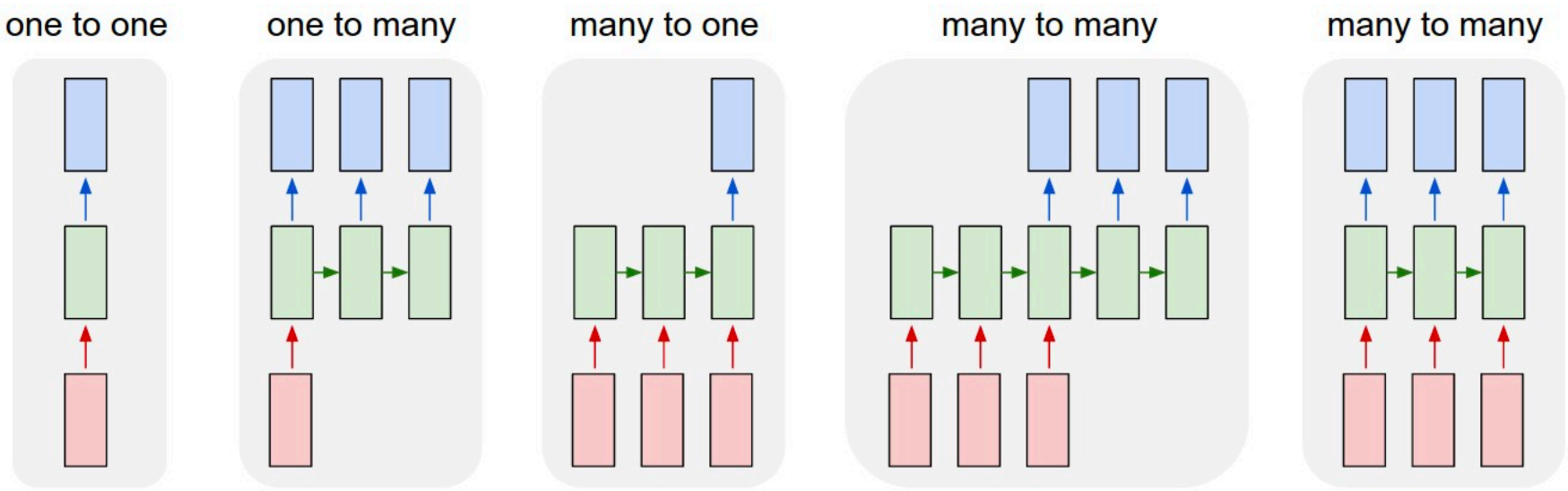
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Number of Parameters in LSTM

- Input_size = 64
- LSTM_units = 100

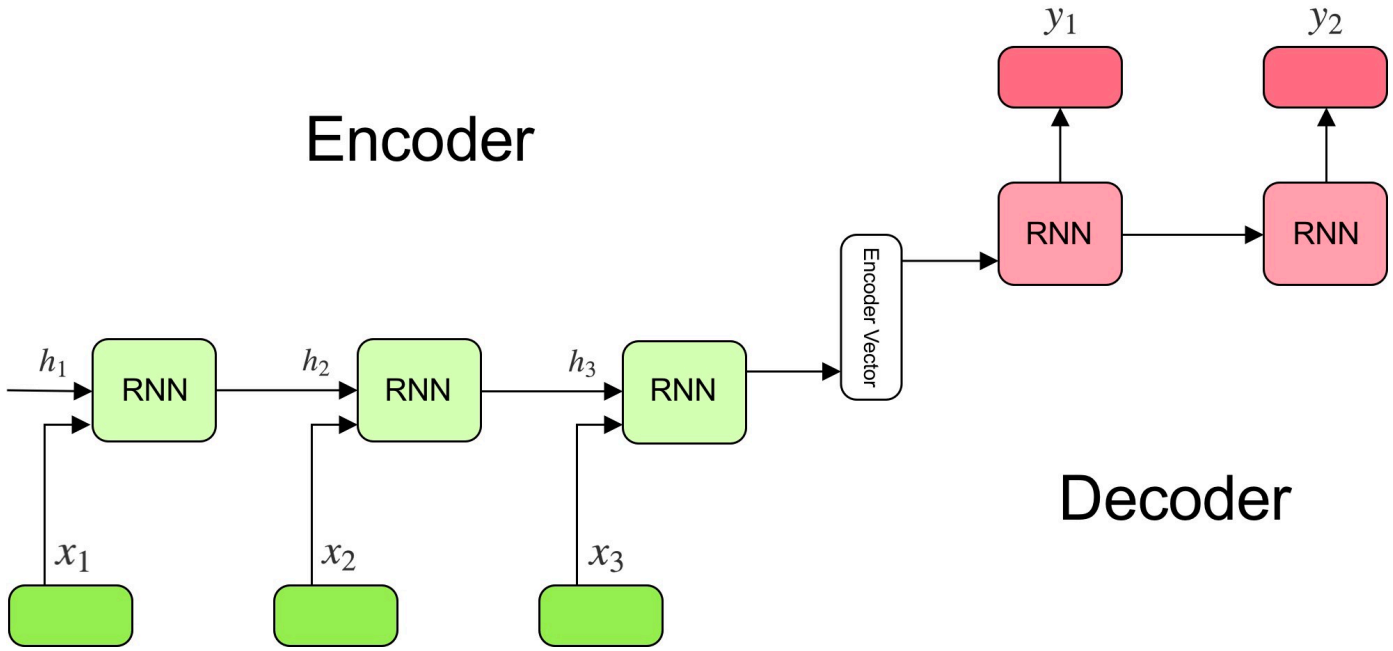
- Berapa total parameter di LSTM layer?

Flexibility RNN Models

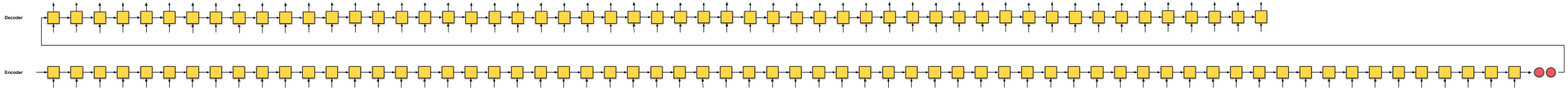


http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Encoder-Decoder



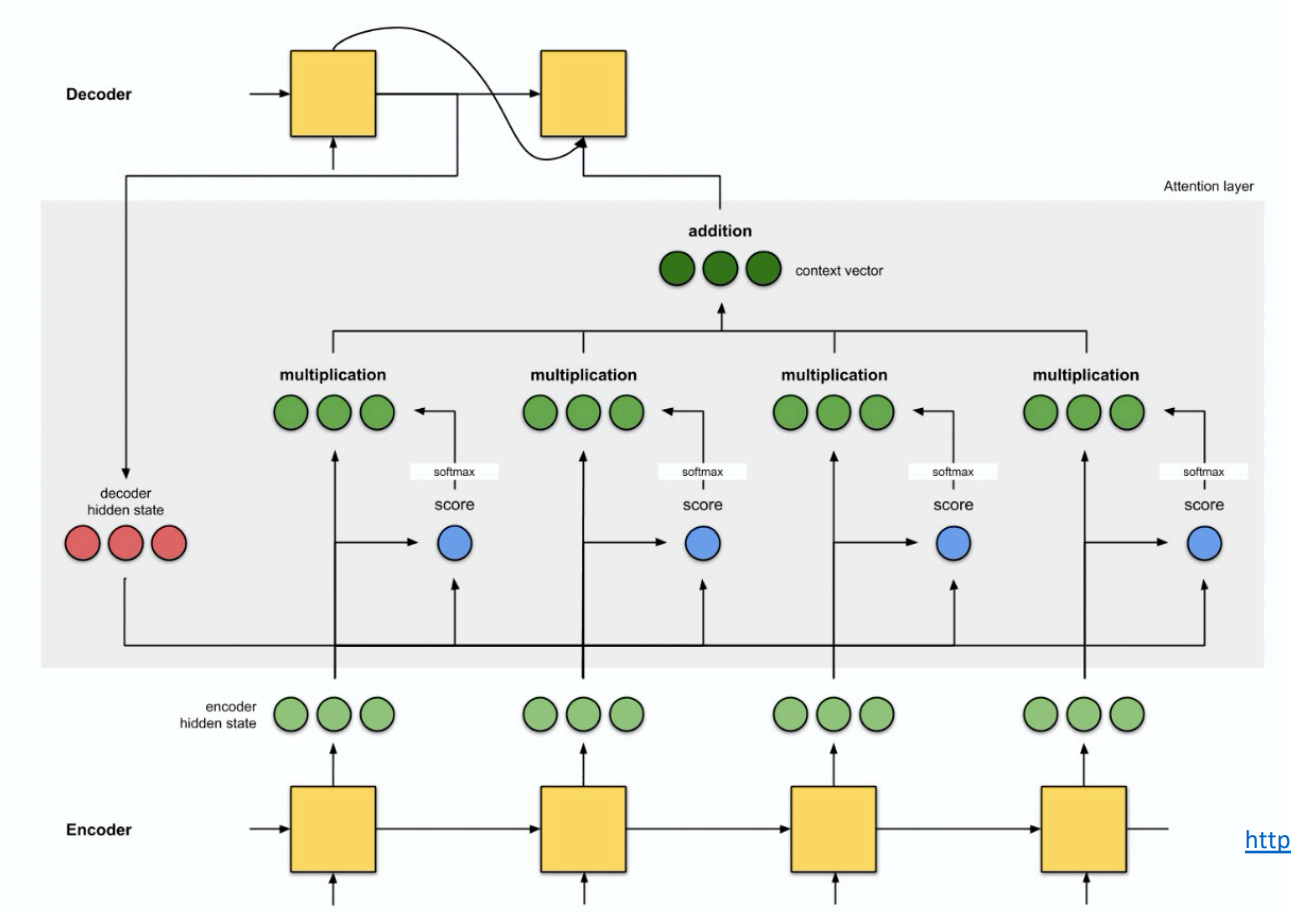
- Problem: For long input, decoder may only use the last few words



https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

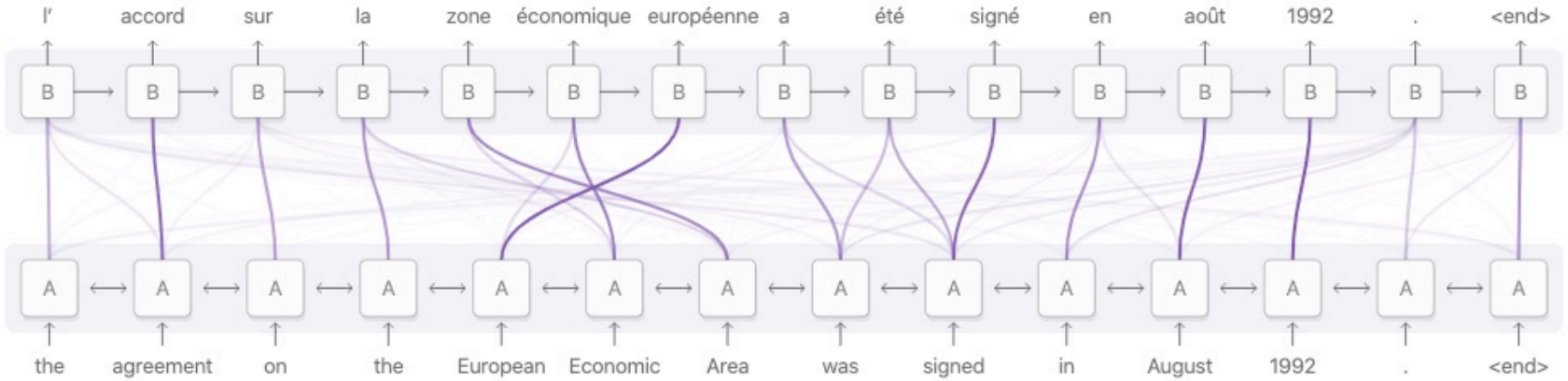
Attention

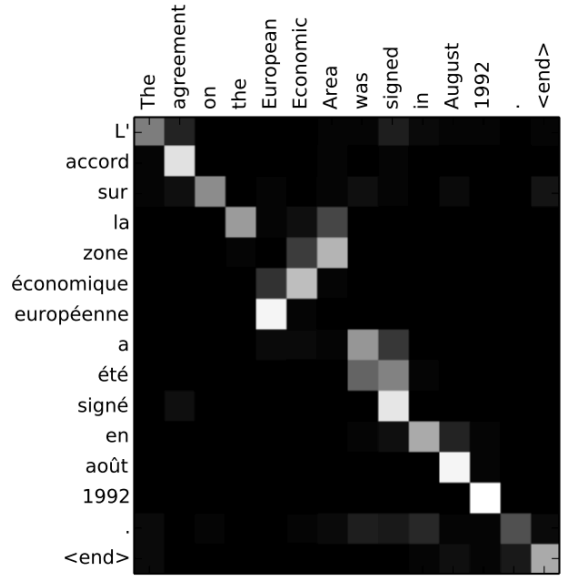
- To decode a target word, give the decoder information from all input words



https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Attention Weight for Neural Machine Translation

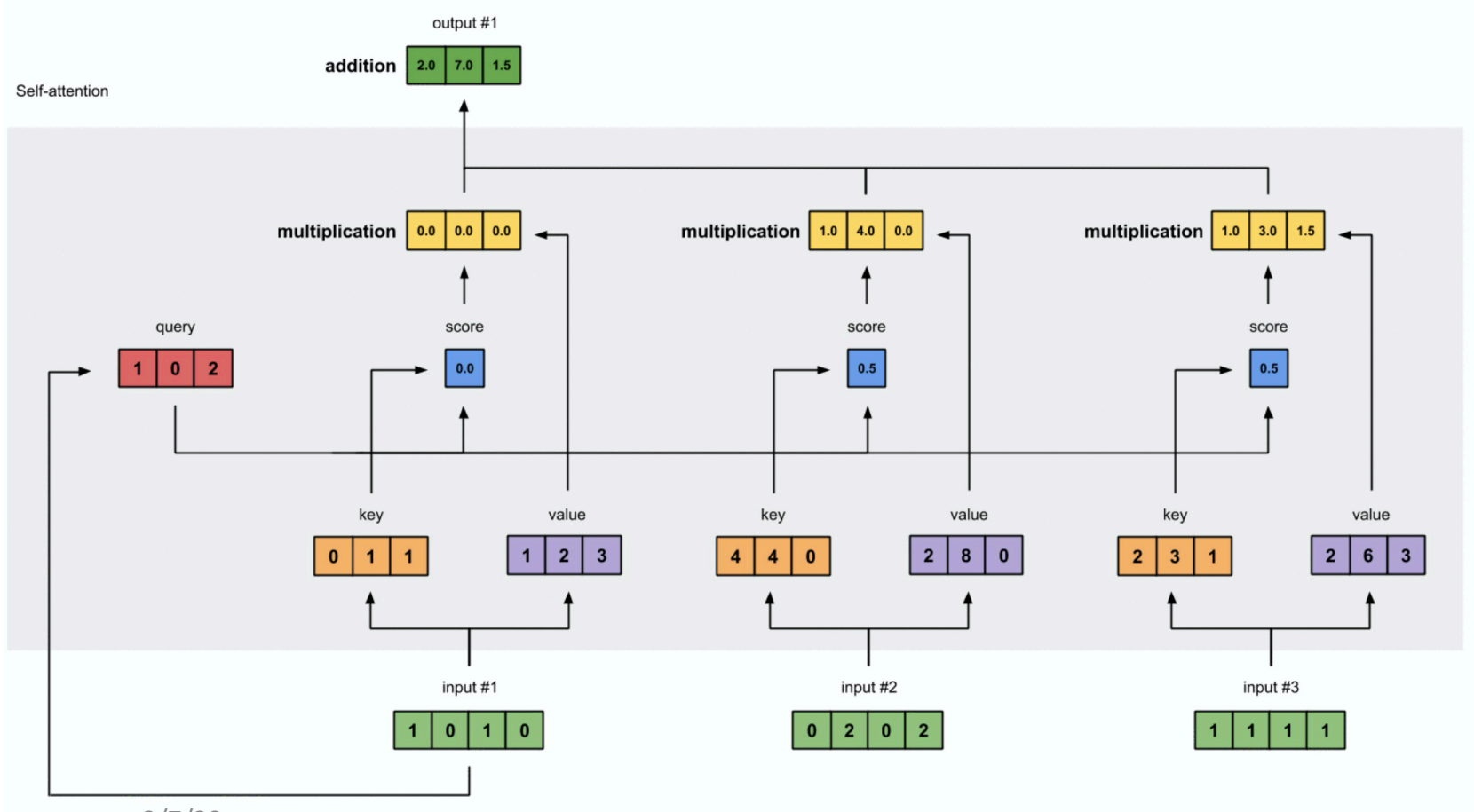


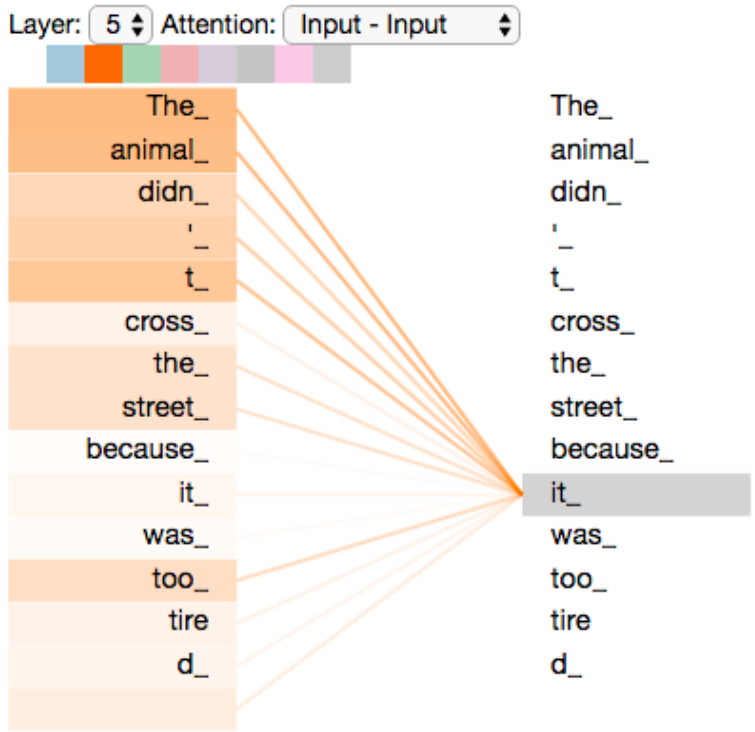


https://distill.pub/2016/augmented-rnns/#attentional-interfaces
https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Self-Attention

- Like attention, but Intra-attention. Final representation of each word (e.g output #1) is resulted by looking at other input words (*self*).

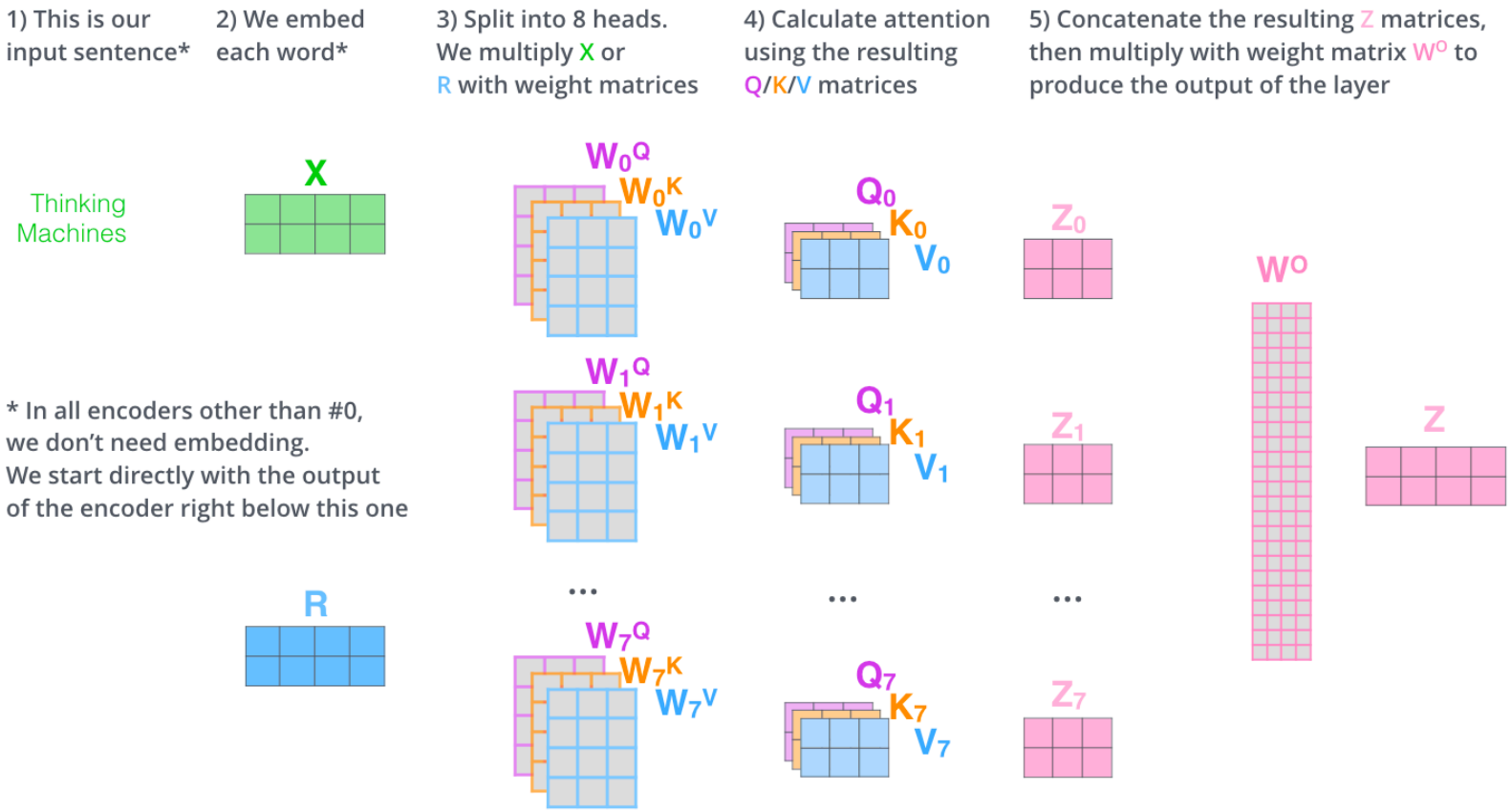




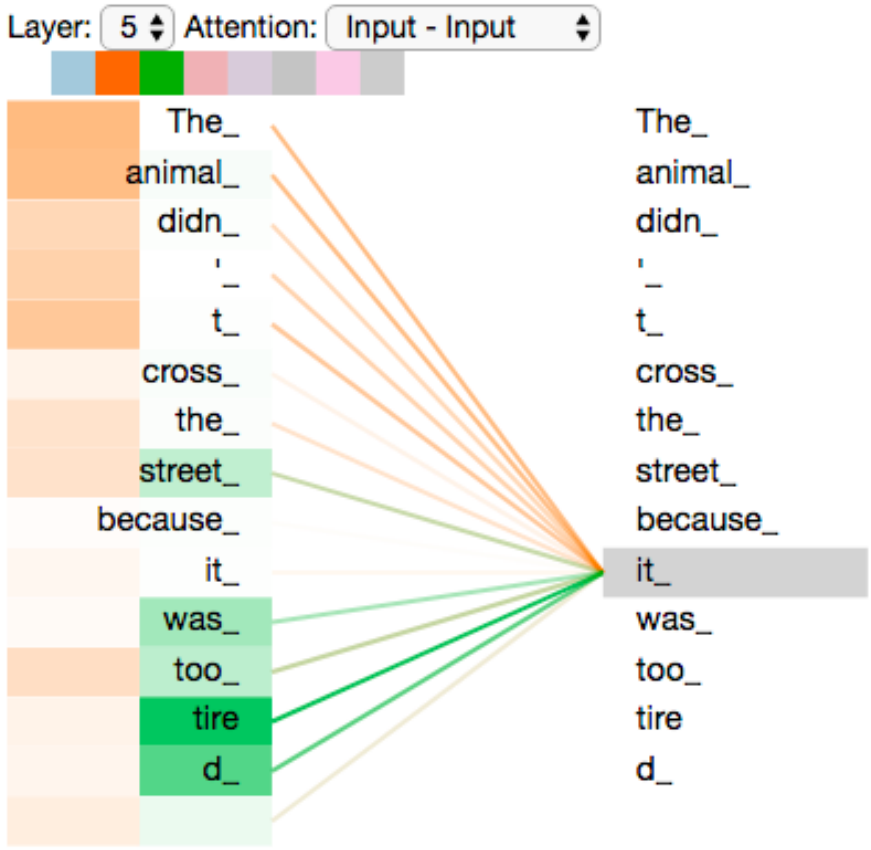
https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

Multi-headed Attention

- Have many sets of Q, K, V

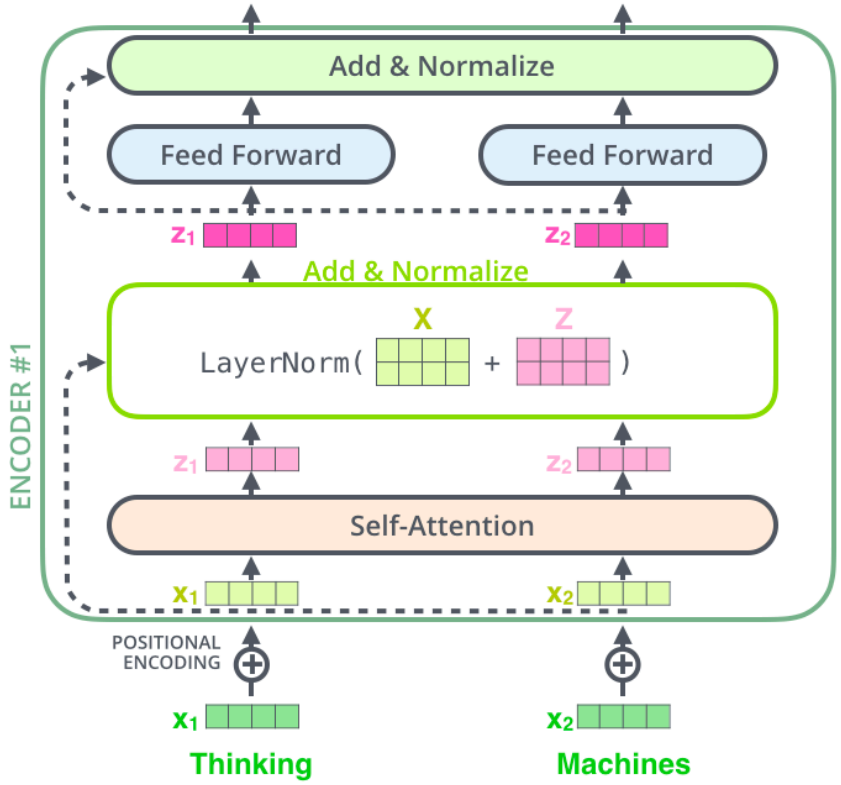


Two heads (orange, green) looking at different words (The animal, tired)

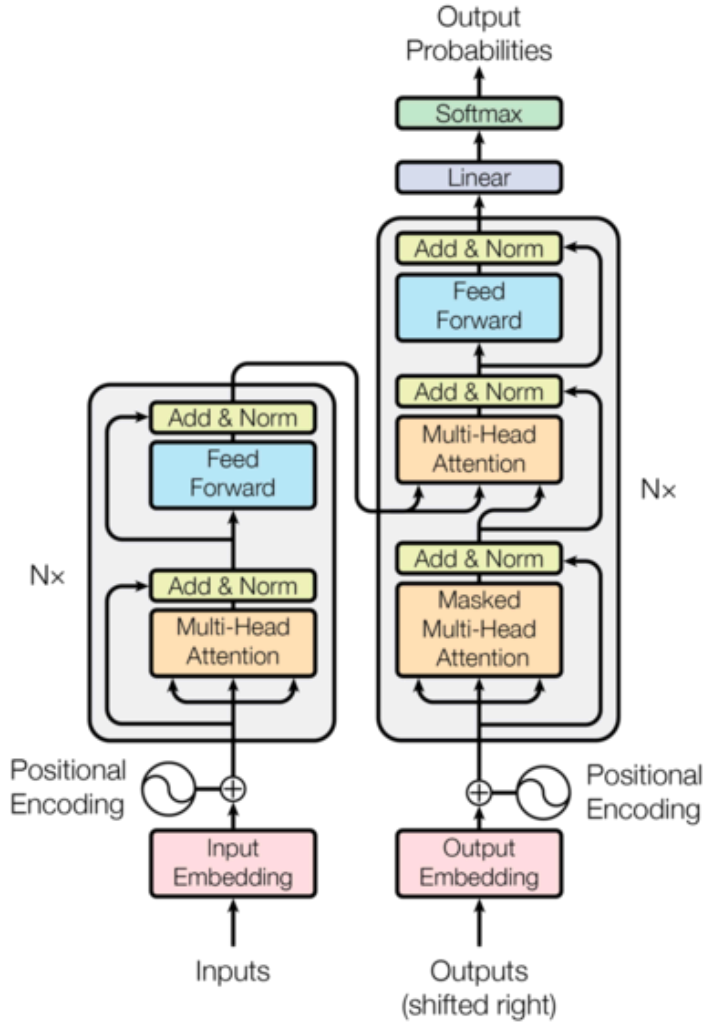


Transformer

- Use 8 attention heads



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

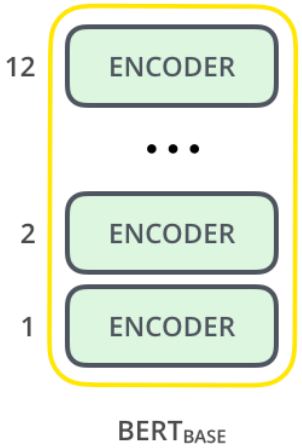


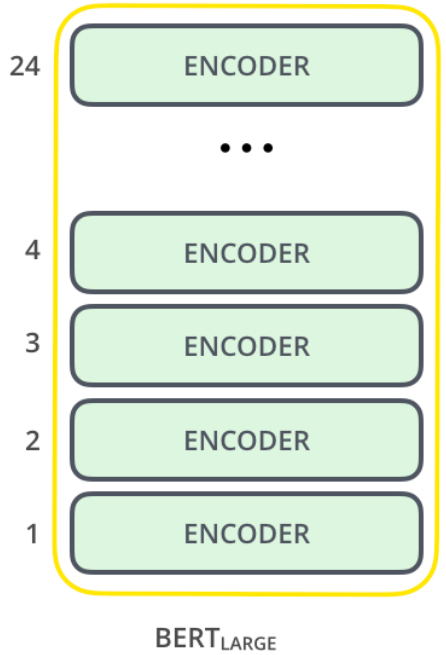
Transformer Encoder-Decoder

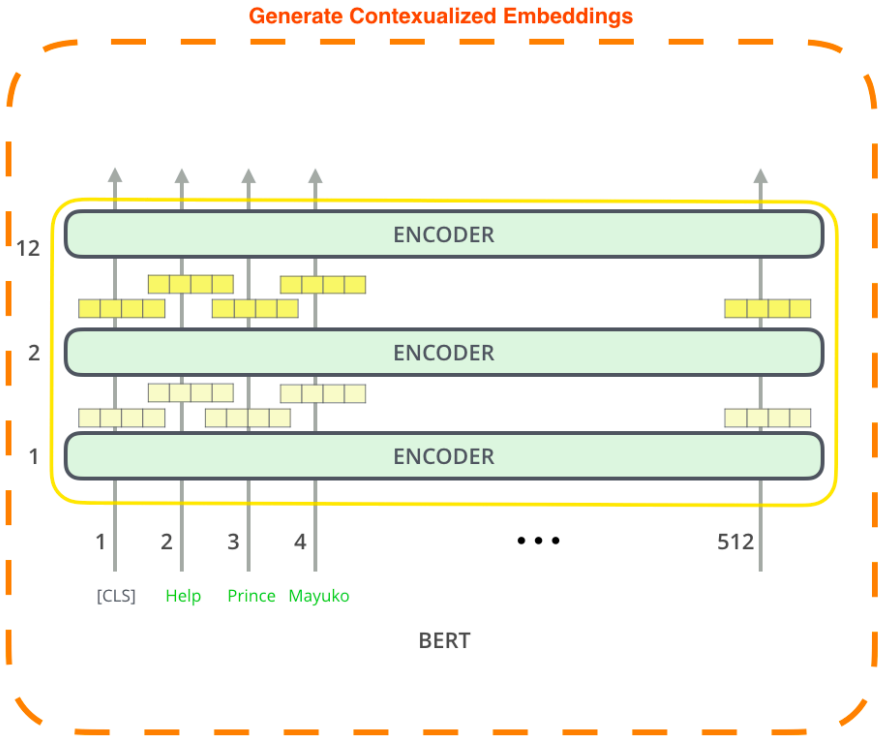
http://nlp.seas.harvard.edu/2018/04/03/attention.html

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)

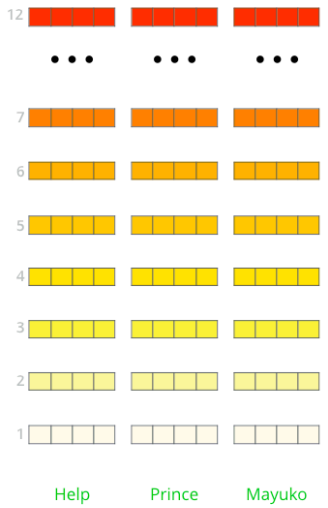
- Stack of many Transformer Encoders







The output of each encoder layer along each token's path can be used as a feature representing that token.

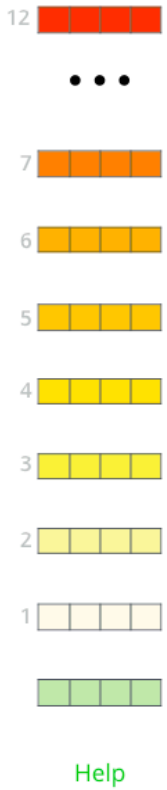


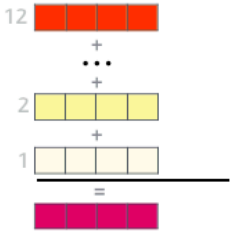
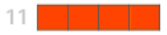
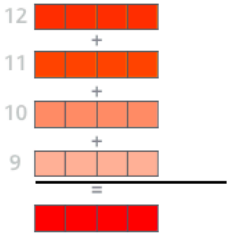
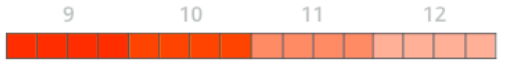
But which one should we use?

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

Contextualized Word Embedding

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER



		Dev F1 Score
First Layer		91.0
Last Hidden Layer		94.9
Sum All 12 Layers		95.5
Second-to-Last Hidden Layer		95.6
Sum Last Four Hidden		95.9
Concat Last Four Hidden		96.1

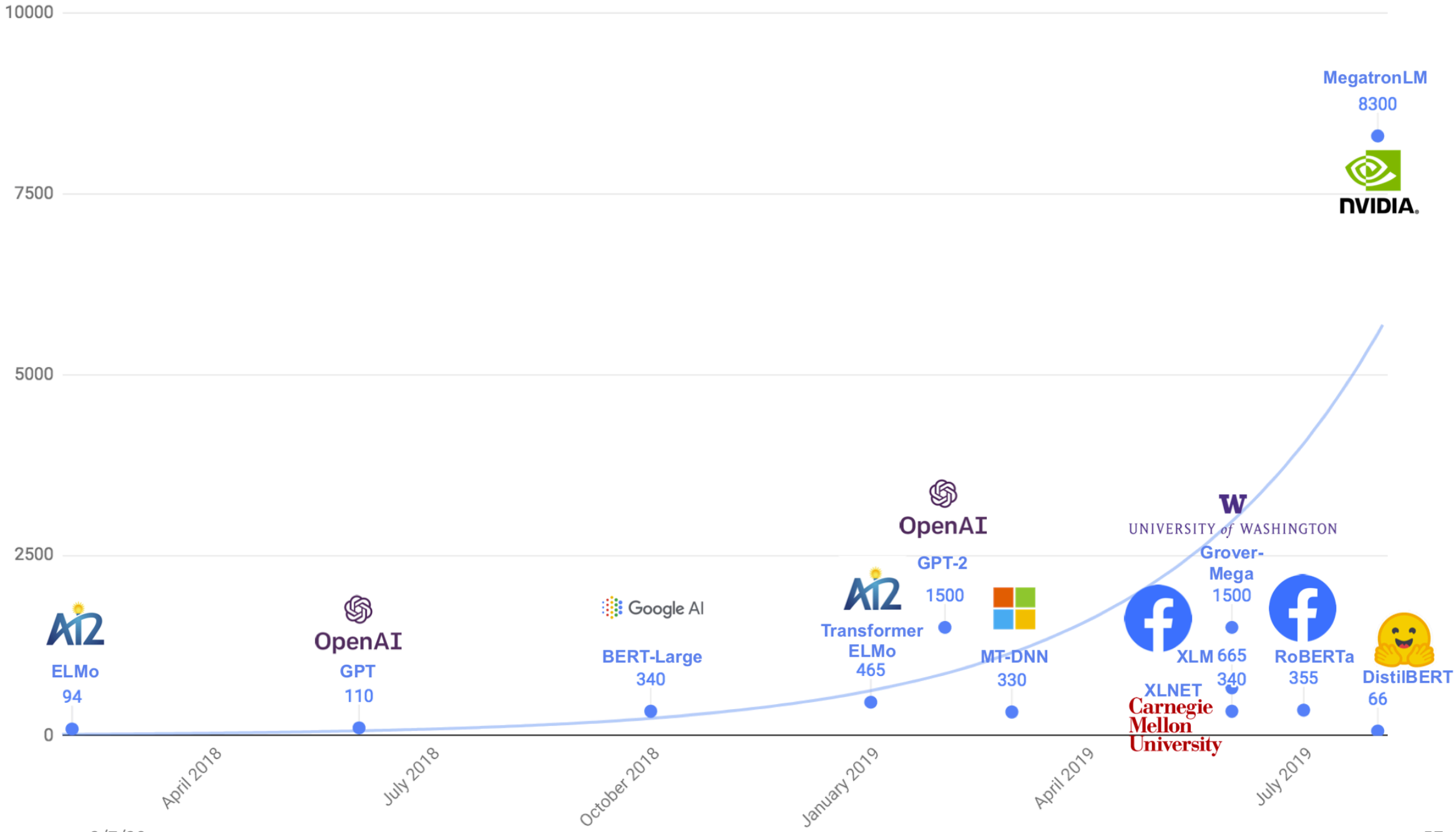
http://jalammar.github.io/illustrated-bert/

Contextual Word Embeddings

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

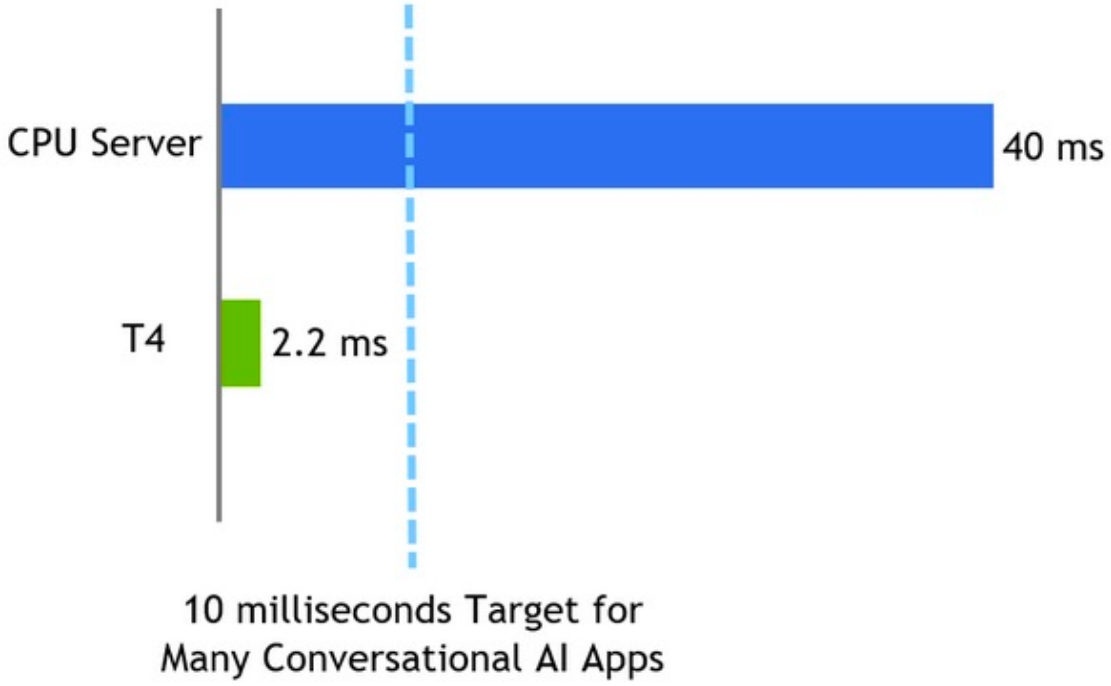
Cost to Train XLNet
$245,000

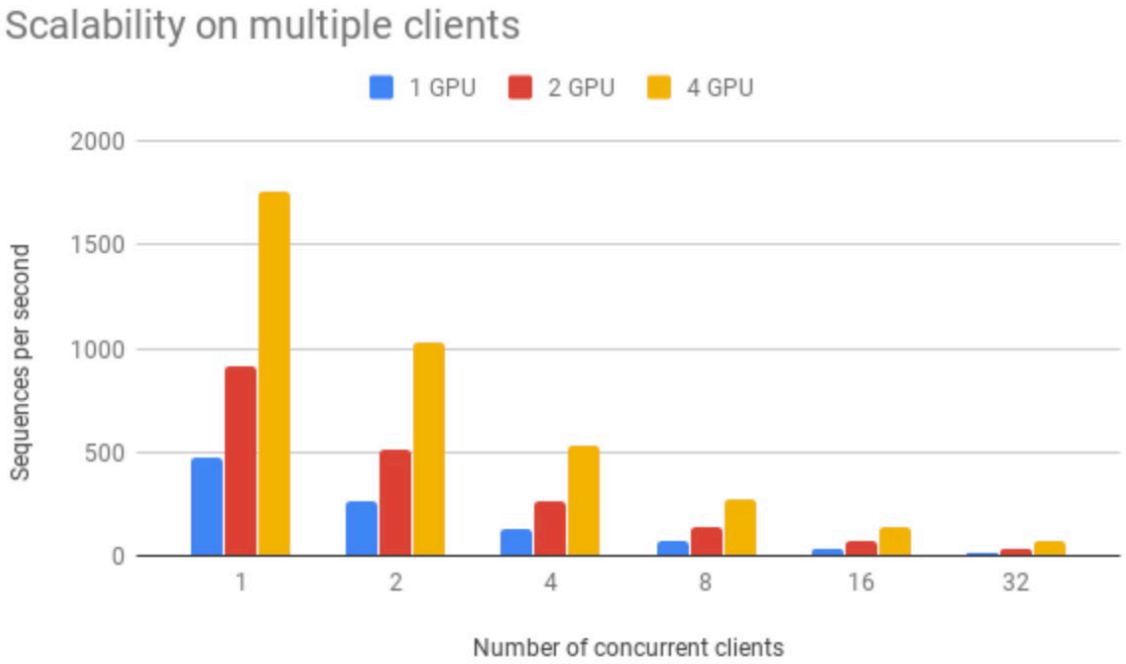
https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8



https://medium.com/@Moscow25/the-best-deep-natural-language-papers-you-should-read-bert-gpt-2-and-looking-forward-1647f4438797

Bert Inference/Forward Time





https://devblogs.nvidia.com/nlu-with-tensorrt-bert/

https://github.com/hanxiao/bert-as-service

Abstraction

FBLearner · UBER Michelangelo · Amazon SageMaker · Determined AI · Neptune Machine Learning Lab · H2O · FLOYD · DOMINO DATA LAB

"All-in-one"

Data

- Versioning: Quilt, DVC, Pachyderm
- Labeling: figure eight, scale, prodigy
- Workflow: Luigi, TRIFACTA
- Database
- Storage: NFS, S3
- or

Development

- Distributed Training: Distributed TensorFlow, Horovod
- Frameworks: PyTorch, Keras, fast.ai
- Software Engineering: jupyter, git, vim
- Hyperparam Optimization: SIGOPT
- Experiment Management: comet, Weights & Biases, TensorBoard, Losswise
- Resource Management: RiseML, slurm, PY

Training/Evaluation

- or

Deployment

- Monitoring: ?
- Hardware / Mobile: ML, NVIDIA TensorRT, TensorFlow Lite, ?
- Web: ALGORITHMIA
- Interchange: ONNX
- CI / Testing: Buildkite, ?
- or · or

https://fullstackdeeplearning.com/

Deep Learning State of The Art 2020

- https://docs.google.com/spreadsheets/d/1fG-K0MwoQ4y0p_9bpMsF1ERlBMbmgtqW06nbMgtIX20/edit?folder=0ABm40ig4NWpDUk9PVA#gid=0
- https://nlpprogress.com/
- https://fullstackdeeplearning.com/march2019

Advice untuk TA

Coba mengumpulkan sendiri data Bahasa Indonesia

- Banyak pelajaran berharga (berguna di Industri)
- Kontribusi untuk data Bahasa

Pahami Data

- Data Collection
- Data Exploration
- Data Preparation

Advice untuk TA

Buat Baseline

- Coba gunakan model sederhana dari library (quick prototyping)
- Apakah datanya menantang? Apakah ada yang salah dengan format data?
- Evaluasi hasil, lalu tentukan langkah selanjutnya berdasarkan evaluasi

Simpan semua Intermediate Result

- Hasil cleaning, preprocessing
- Feature extraction
- Hasil prediksi, dll

Advice untuk TA

Pahami Proses

- Jumlah Parameter
- Ukuran Matrix
- Tipe Input/Output

Terima Kasih