UVA CS 4774: Machine Learning

S3: Lecture 18:

Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

Module I

Today: Neural Network Models on 1D Grid / Language Data

Data: X

Task: y

Representation: : x, f()

Score Function: L()

Search/Optimization : argmin()

Models, Parameters, metrics

X: 1D Grid Text

Y: Many possible Kinds,

e.g., 1D Grid Text

Word2Vec/ Recurrent / Transformer

Many self-supervised kinds, e.g., skip-gram / mask LM

SGD / Backprop

weights, bias / hyperparameters / Accuracy / F1 / more metris

Today: Neural Network Models on 1D Grid / Language Data

Data: X

↓

Task: y

↓

Representation: :   x, f()

↓

Score Function: L()

↓

Search/Optimization : argmin()

↓

Models, Parameters, metrics

X: 1D Grid Text

↓

Y: Many possible Kinds,

e.g., 1D Grid Text

↓

Word2Vec/ Recurrent / Transformer

↓

Many self-supervised kinds, e.g., skip-gram / mask LM

↓

SGD / Backprop

↓

weights, bias / hyperparameters / Accuracy / F1 / more metris

What is NLP

- **Wiki: Natural language processing** (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.

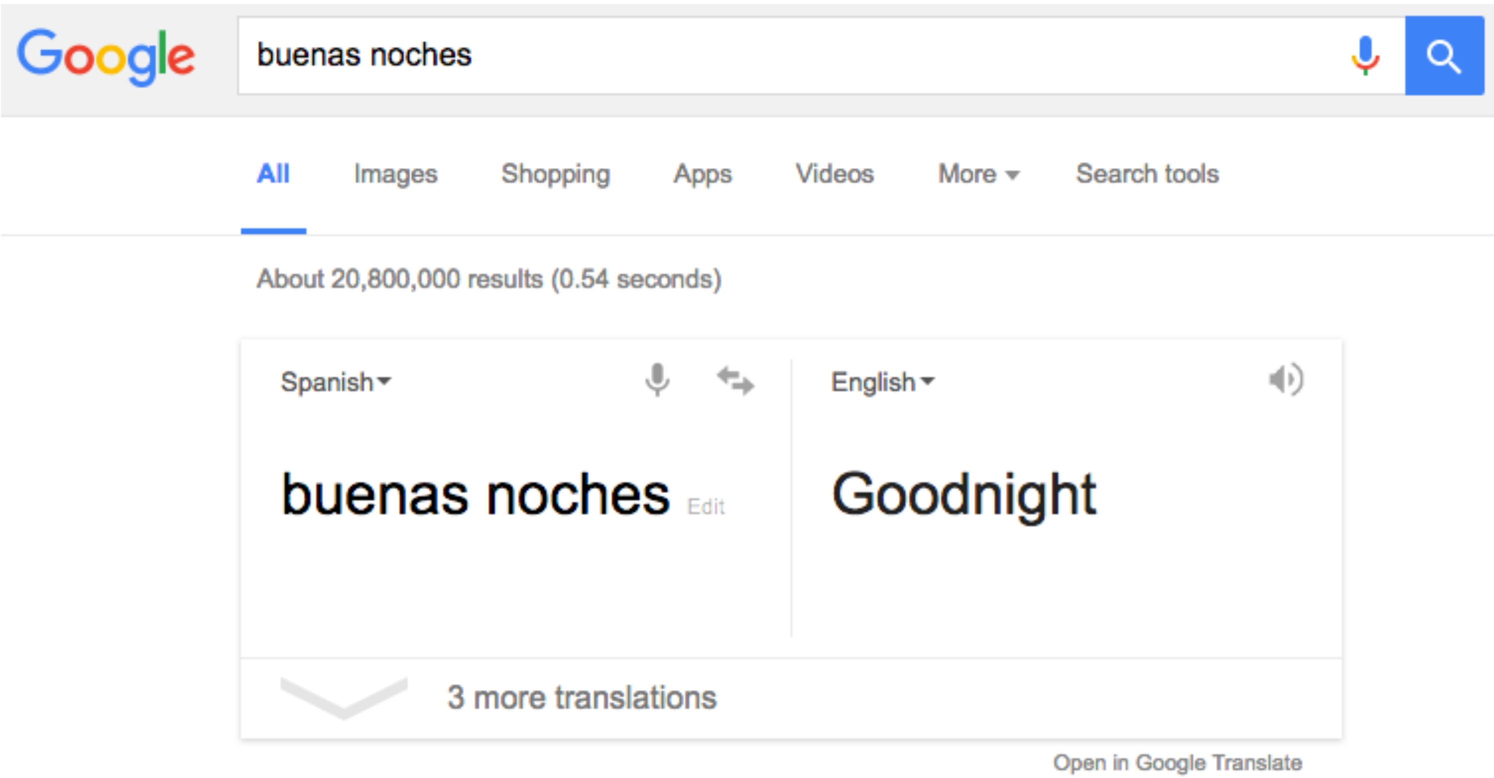
Credit: kaiwai Chang

Go beyond the keyword matching

- Identify the structure and meaning of words, sentences, texts and conversations
- Deep understanding of broad language
- NLP is all around us

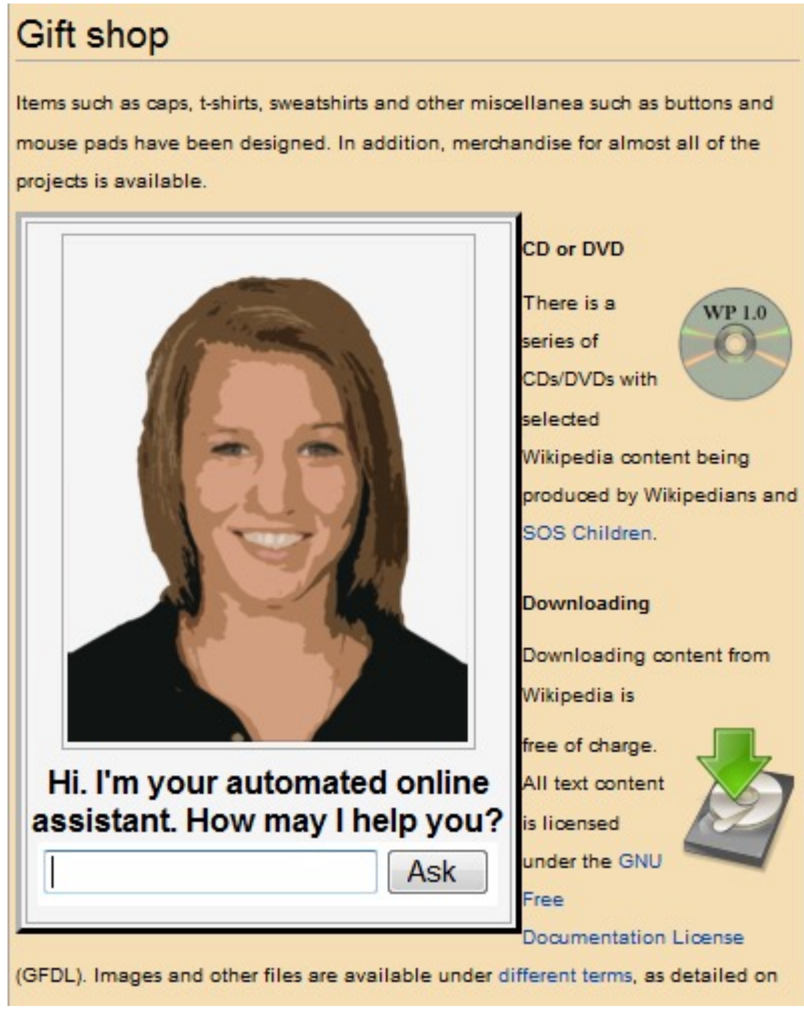
Credit: kaiwai Chang

Machine translation



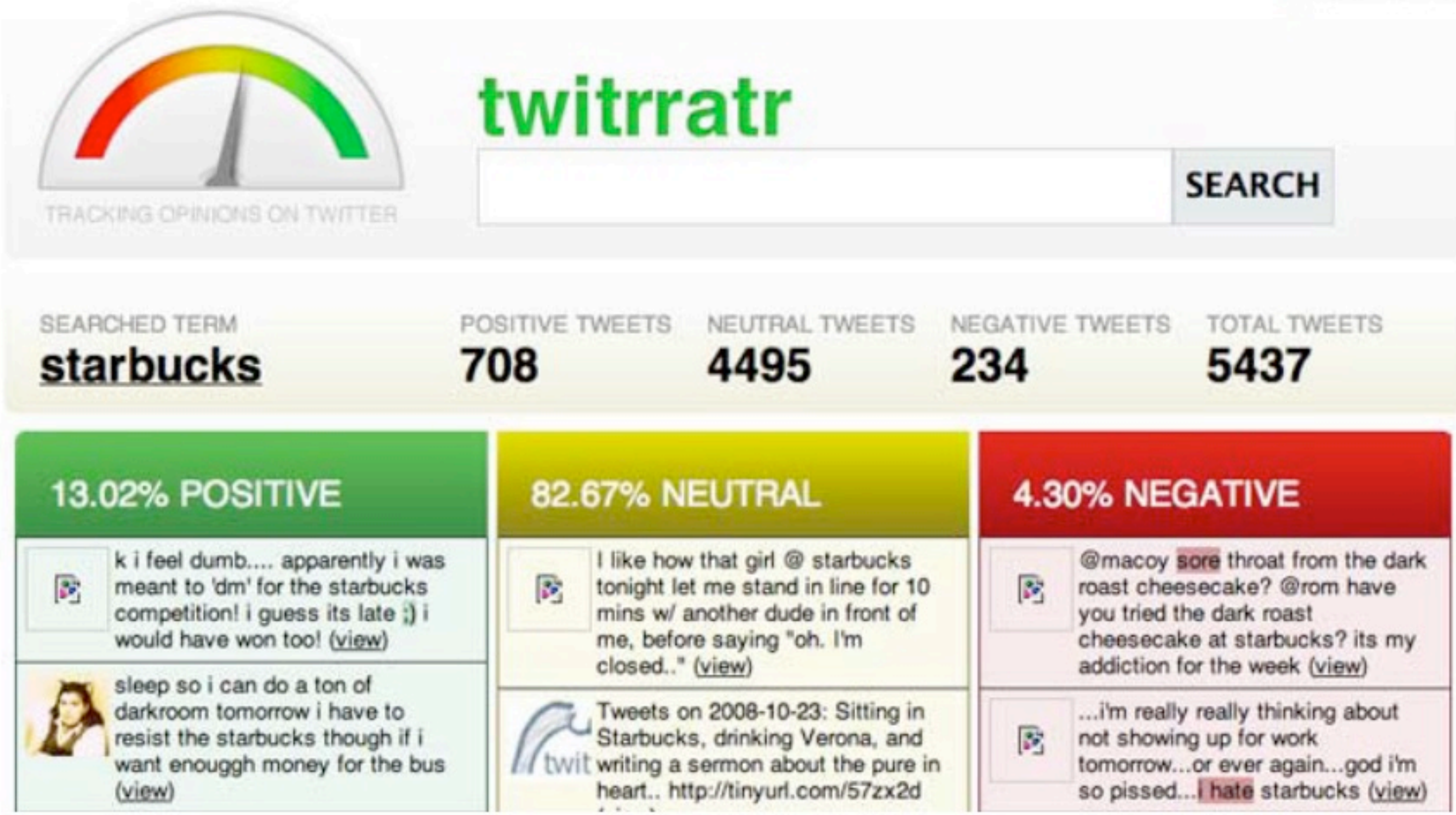
Credit: kaiwai Chang

Dialog Systems



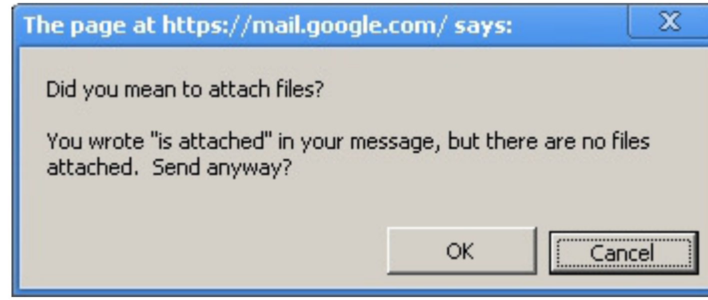
Credit: kaiwai Chang

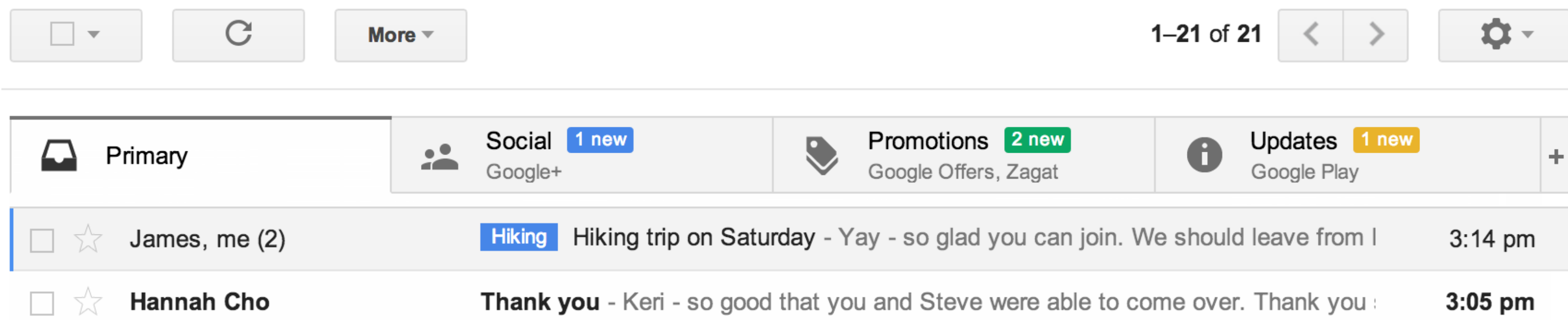
Sentiment/Opinion Analysis



Credit: kaiwai Chang

Text Classification



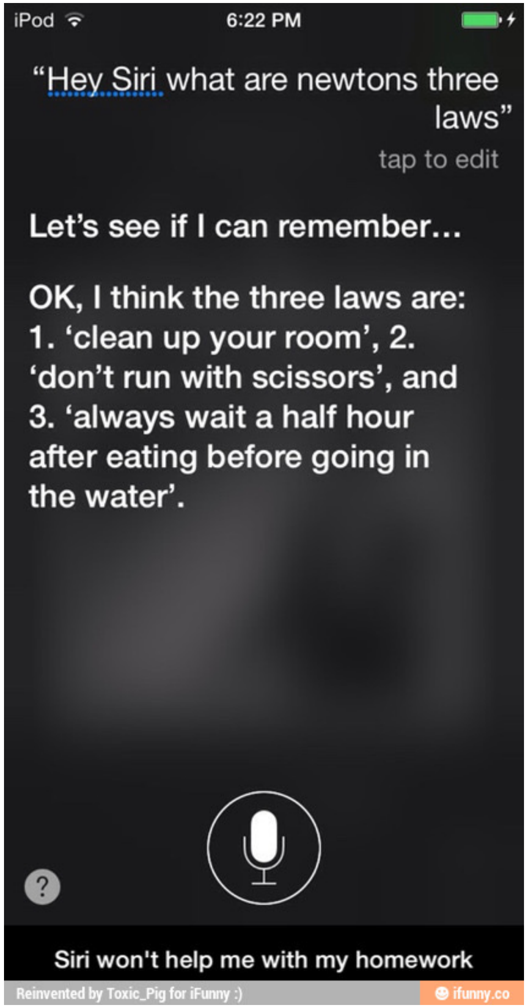


www.wired.com

Credit: kaiwai Chang

Question answering

['Watson' computer wins at 'Jeopardy']



credit: ifunny.com

Credit: kaiwai Chang

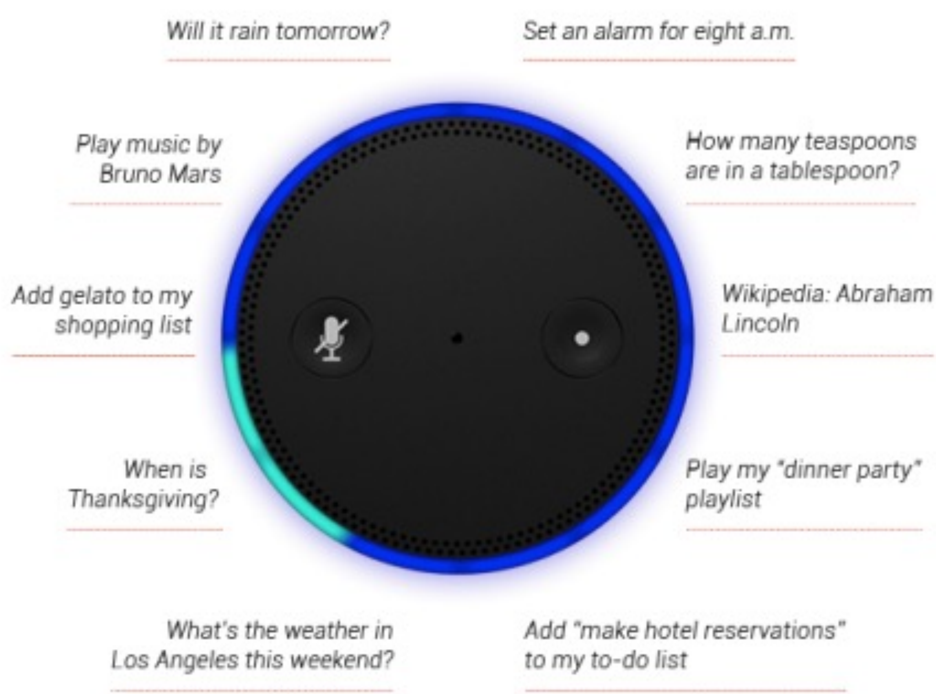
Language Comprehension

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh.** As **a boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

- Q: who wrote Winnie the Pooh?
- Q: where is Chris lived?

Credit: kaiwai Chang

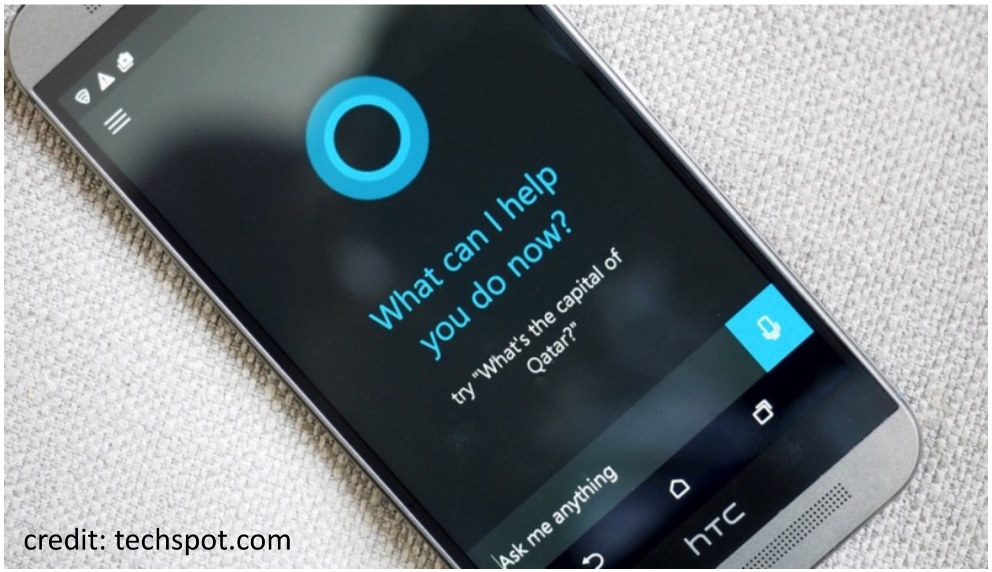
Natural language instruction



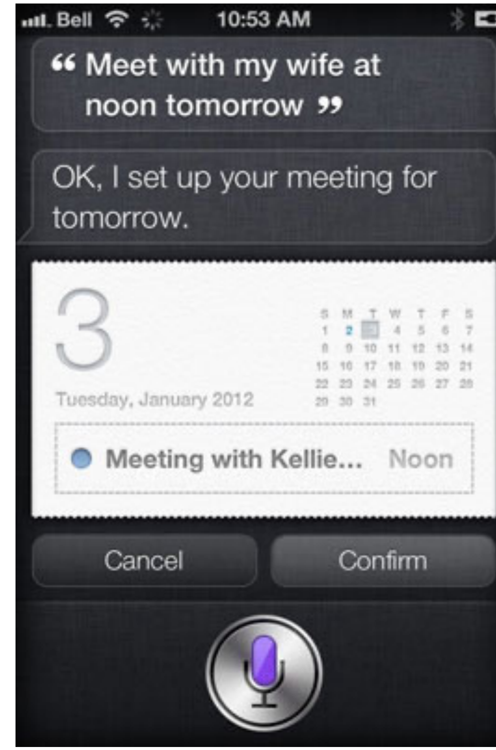
https://youtu.be/KkOCeAtKHIc?t=1m28s

Credit: kaiwai Chang

Digital personal assistant



credit: techspot.com



- Semantic parsing – understand tasks
- Entity linking – “my wife” = “Kellie” in the phone book

Credit: kaiwai Chang

Challenges – ambiguity

- Pronoun reference ambiguity

Credit: http://www.printwand.com/blog/8-catastrophic-examples-of-word-choice-mistakes

Credit: kaiwai Chang

Challenges – language is not static

- Language grows and changes
  - e.g., cyber lingo

LOL	
G2G	
BFN	
B4N	
Idk	
FWIW	
LUWAMH	

Credit: kaiwai Chang

Challenges – scale

- Examples:
  - Bible (King James version): ~700K
  - Penn Tree bank ~1M from Wall street journal
  - Newswire collection: 500M+
  - Wikipedia: 2.9 billion word (English)
  - Web: several billions of words

Credit: kaiwai Chang

Today: Neural Network Models on 1D Grid / Language Data

Data: X

Task: y

Representation: :   x, f()

Score Function: L()

Search/Optimization : argmin()

Models, Parameters, metrics

X: 1D Grid Text

Y: Many possible Kinds,

e.g., 1D Grid Text

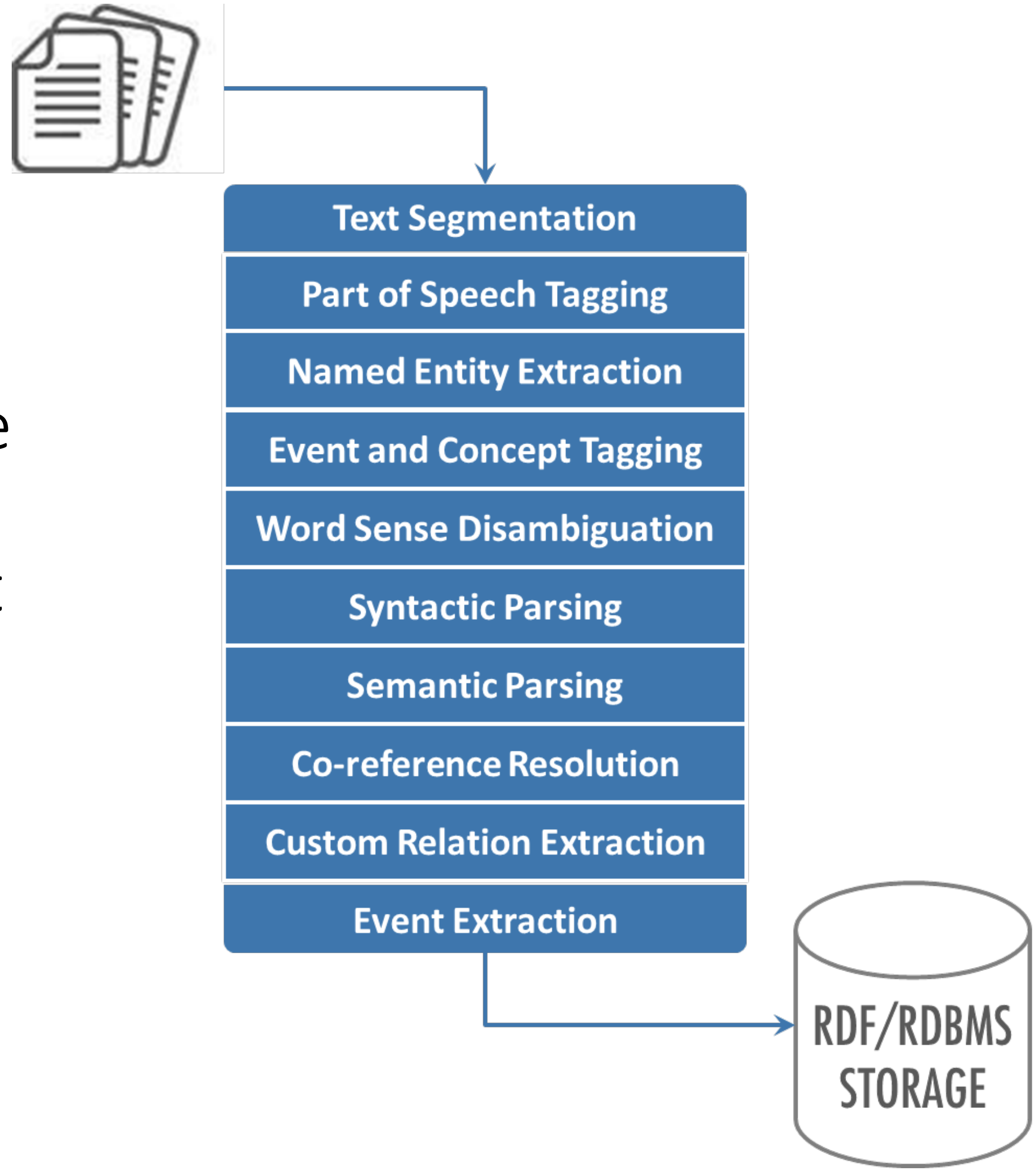
Word2Vec/ Recurrent / Transformer

Many self-supervised kinds,
e.g., skip-gram / mask LM

SGD / Backprop

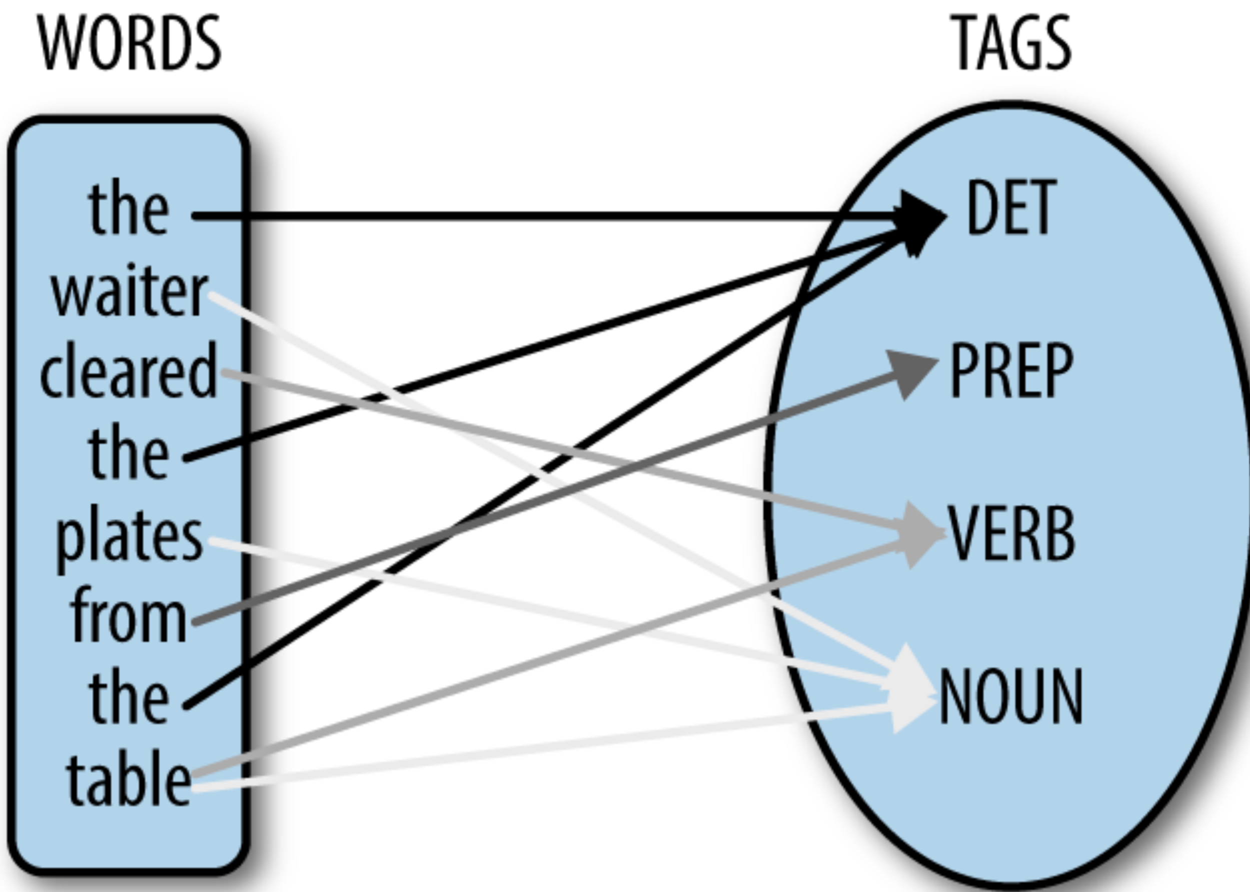
weights, bias /
hyperparameters / Accuracy /
F1 / more metris

Classic NLP Pipeline Components for Understanding Text



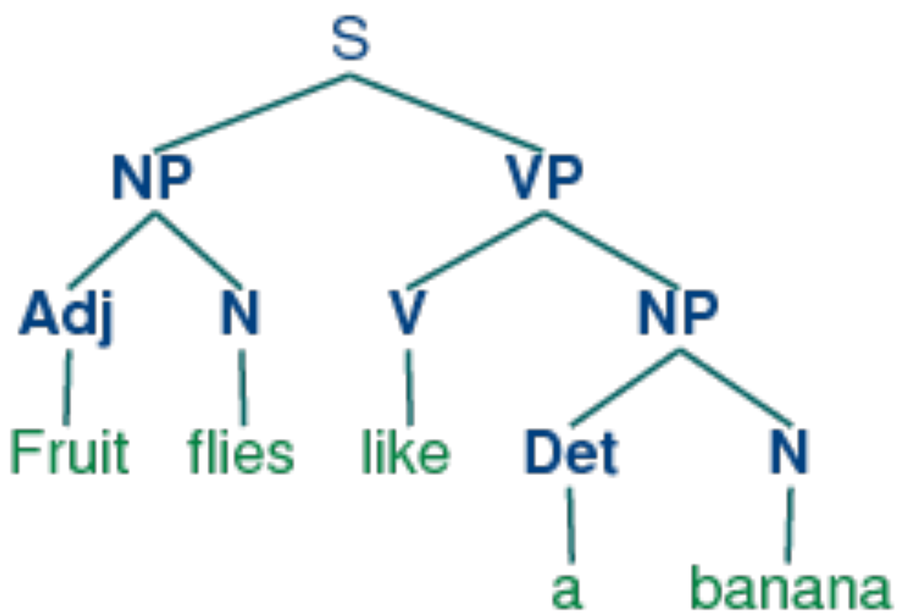
Credit: kaiwai Chang

Part of speech tagging



Credit: kaiwai Chang

Syntactic (Constituency) parsing



Credit: kaiwai Chang

Syntactic structure => meaning

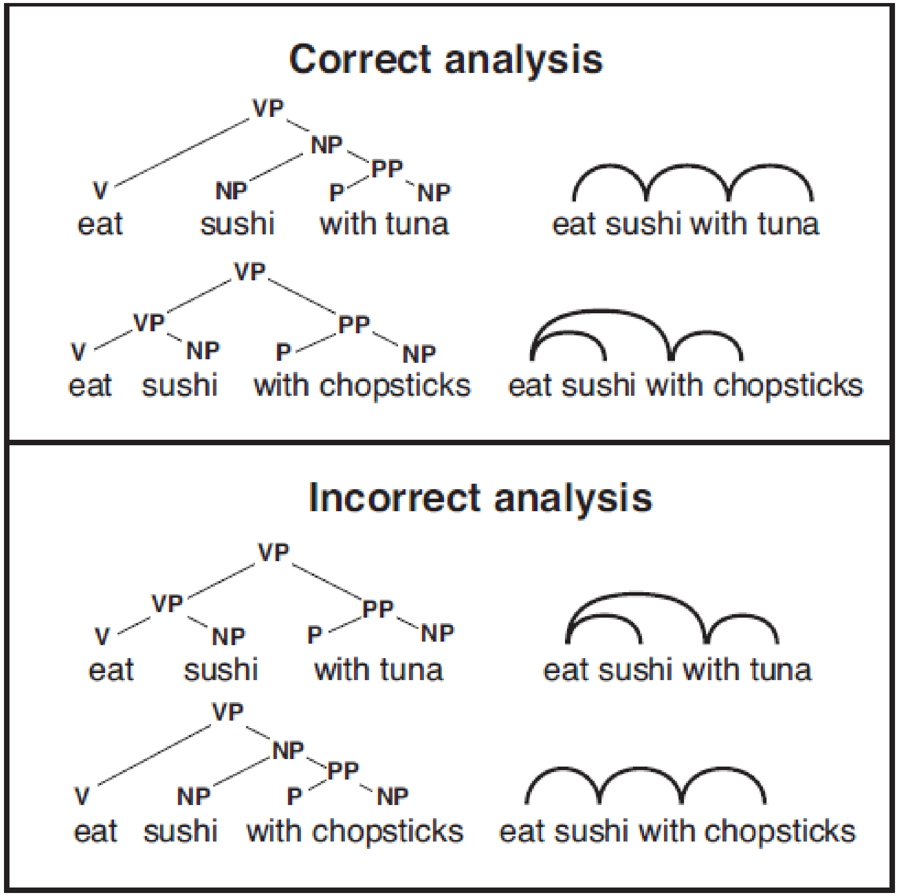
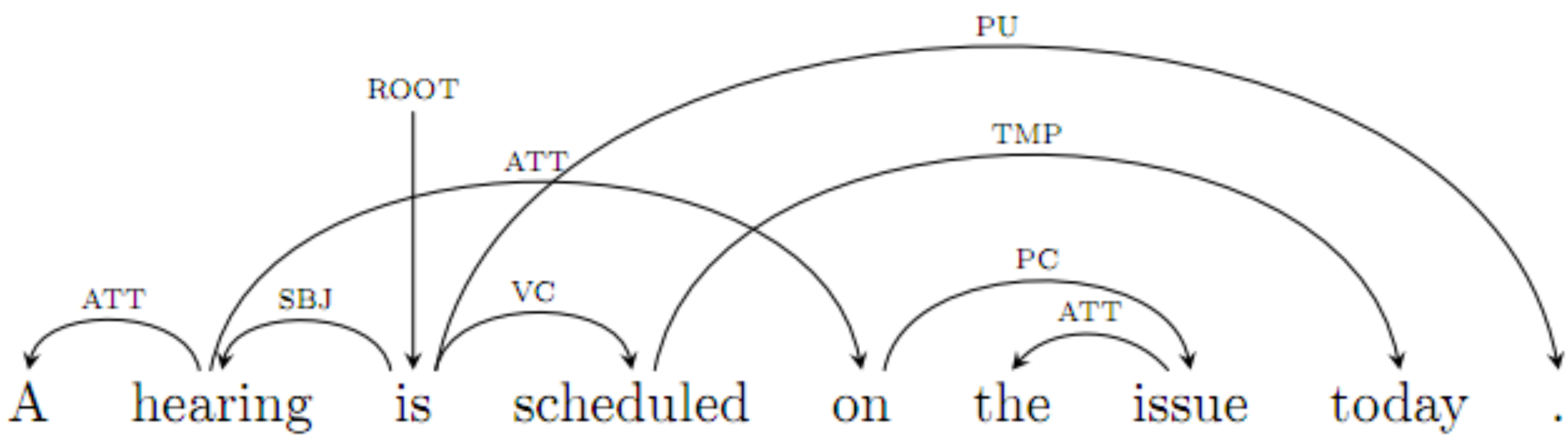


Image credit: Julia Hockenmaier, Intro to NLP

Credit: kaiwai Chang

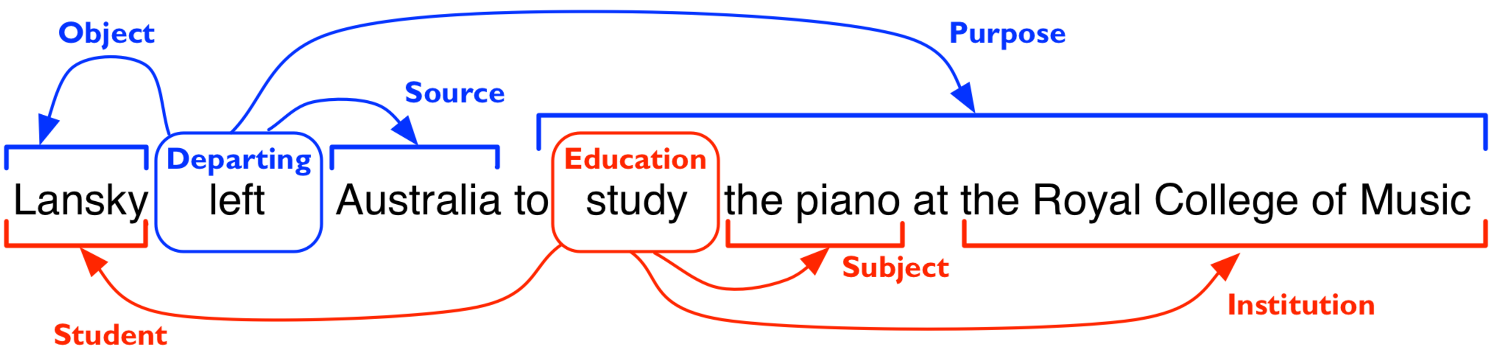
Dependency Parsing



Credit: kaiwai Chang

Semantic analysis

- Word sense disambiguation
- Semantic role labeling



Credit: Ivan Titov

Credit: kaiwai Chang

Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

Credit: kaiwai Chang

Q: [Chris] = [Mr. Robin] ?

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh.** As **a boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

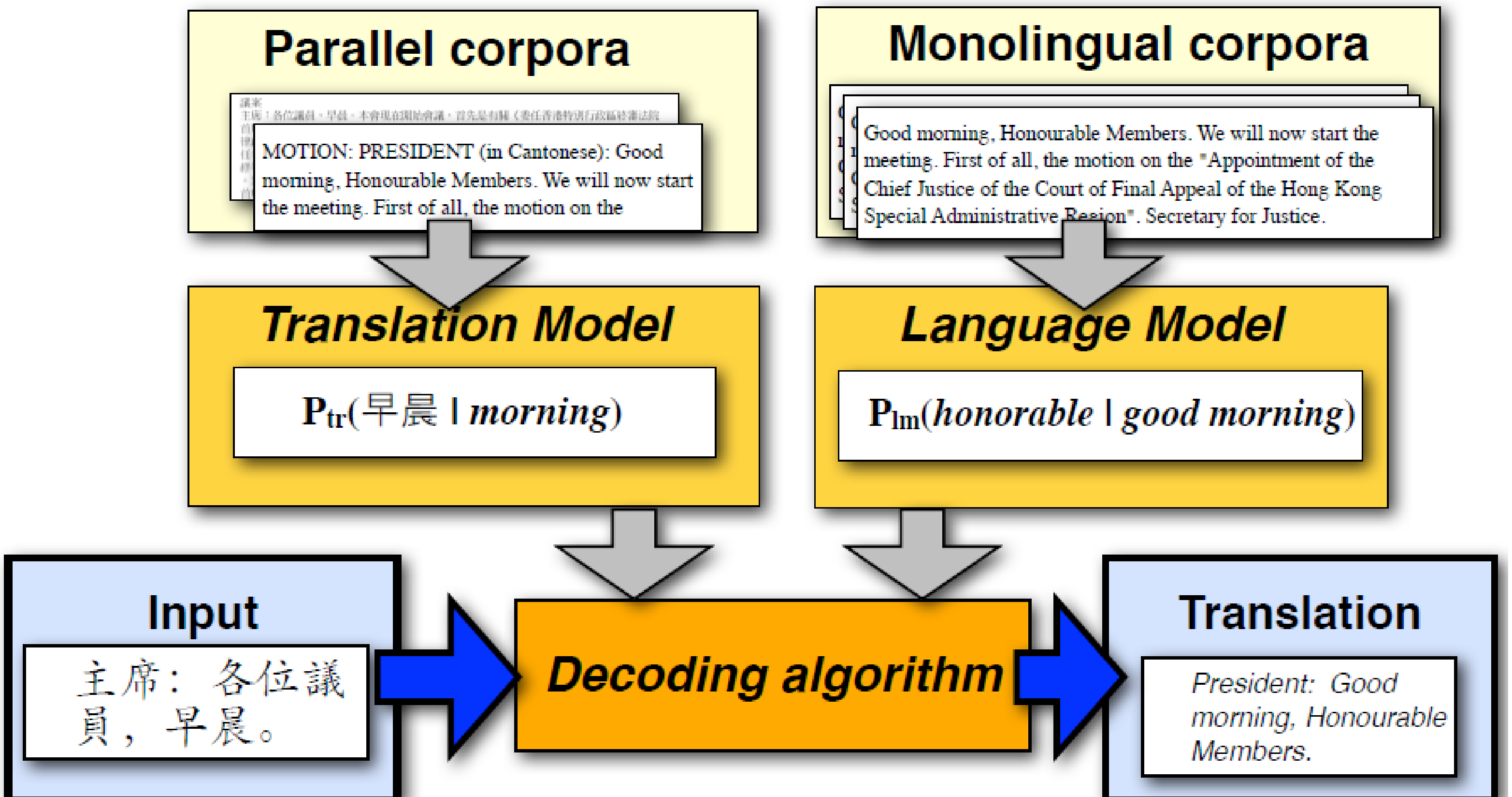
Slide from Dan Roth

Co-reference Resolution

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Slide from Dan Roth

Statistical machine translation



Credit: Kai-Wei Chang

UVA CS 4774: Machine Learning

S3: Lecture 18:

Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Module II

Today: Neural Network Models on 1D Grid / Language Data

Data: X

Task: y

Representation: : x, f()

Score Function: L()

Search/Optimization : argmin()

Models, Parameters, metrics

X: 1D Grid Text

Y: Many possible Kinds,

e.g., 1D Grid Text

Word2Vec/ Recurrent / Transformer

Many self-supervised kinds, e.g., skip-gram / mask LM

SGD / Backprop

weights, bias / hyperparameters / Accuracy / F1 / more metris

Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

Variable Length Issue in Natural Language Data:

X

This Food is not good.

Y

This wonderful book is a pleasure to read.

Recap: The bag of words representation

$f($ I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet. $)=c$

Recap: The bag of words representation

$$f\left(\begin{array}{|l|l|}\hline \text{great} & 2 \\ \hline \text{love} & 2 \\ \hline \text{recommend} & 1 \\ \hline \text{laugh} & 1 \\ \hline \text{happy} & 1 \\ \hline \dots & \dots \\ \hline \end{array}\right) = c$$

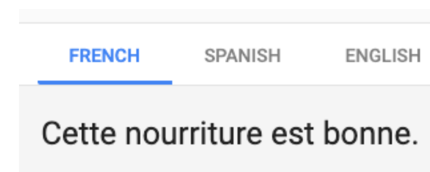
BOW NOT Applicable to many NLP tasks:

- removes position information and can not (or hard to) represent word compositions

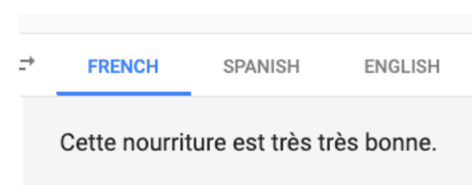
Y: French Translation

X

This Food is good.



This Food is very very good.

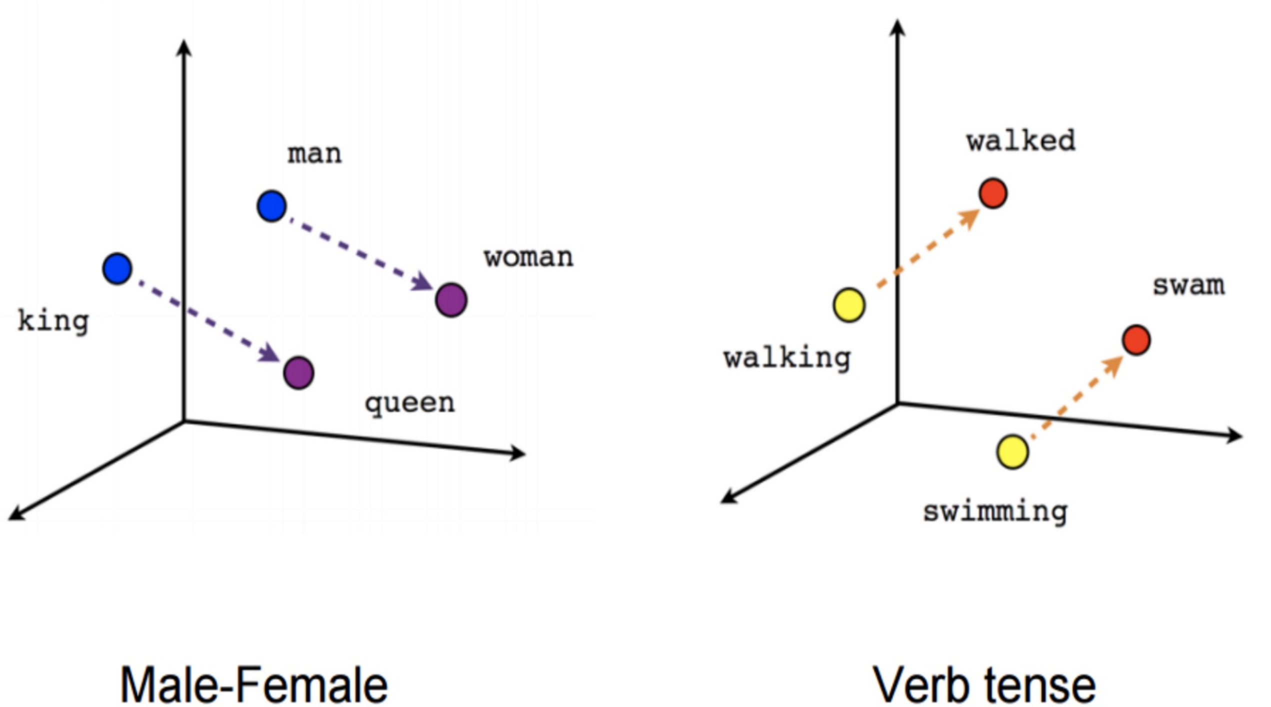


Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

How to Represent A Word in DNN

- Basic approach – “one hot vector”
  - Binary vector
  - Length = | vocab |
  - 1 in the position of the word id, the rest are 0
  - However, does not represent word meaning
  - Extremely high dimensional (there are over 200K words in the English language)
  - Extremely sparse



- Solution: Distributional Word Embedding Vectors

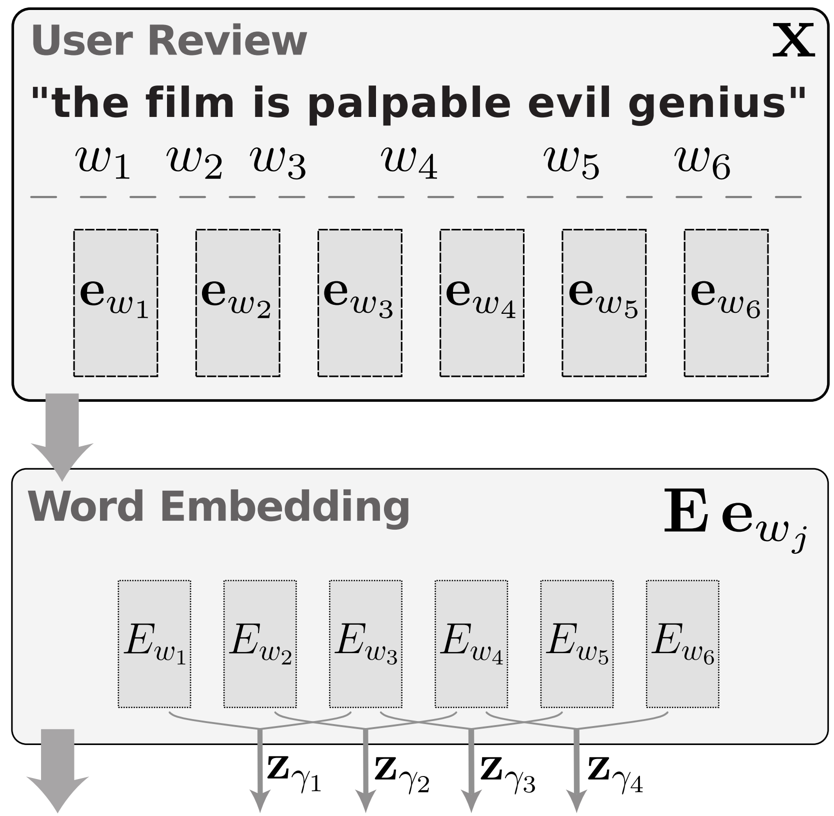
Popular word embeddings

- GloVe (Global Vectors)
  - Pennington et al., 2014
- fasttext
  - Bojanowski et al., 2017

However, Natural language is

- Variable-length
- Composition of multiple words
- Word meaning is contextual

- Elmo
  - Peters, 2018
- BERT
  - Devlin et al., 2018

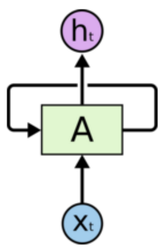


Roadmap : f() on natural language

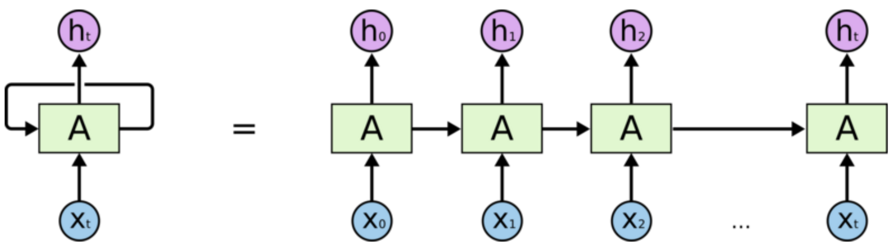
- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

Recurrent Neural Networks

- Allow us to operate over sequences of vectors (with variable length)
- Allow Sequences in the input, as the output, or in the most general case both



Recurrent Neural Networks have loops.



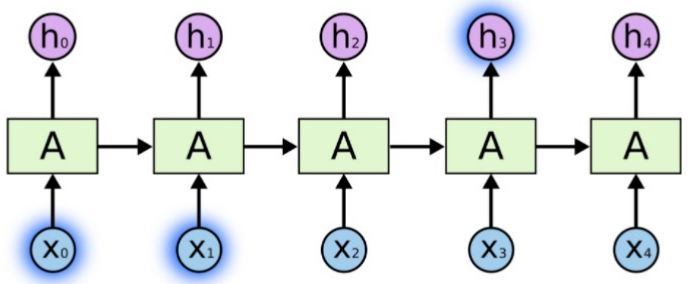
An unrolled recurrent neural network.

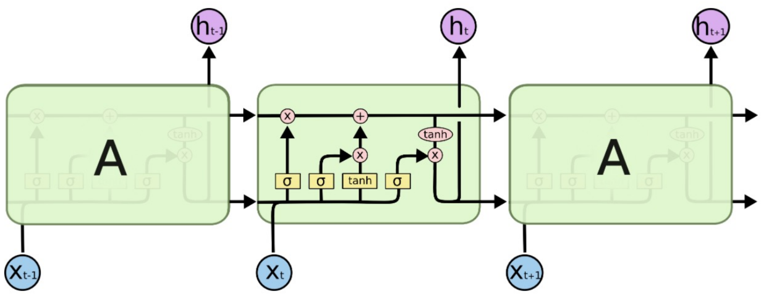
Recurrent Neural Networks are networks with loops in them, allowing information to persist.

Image Credits from Christopher Olah

Deep RNN in the 90's

- Prof. Schmidhuber invented "Long short-term memory" – Recurrent NN (LSTM-RNN) model in 1997





The repeating module in an LSTM contains four interacting layers.

Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780.

Image Credits from Christopher Olah

Recurrent Neural Networks Got Popular

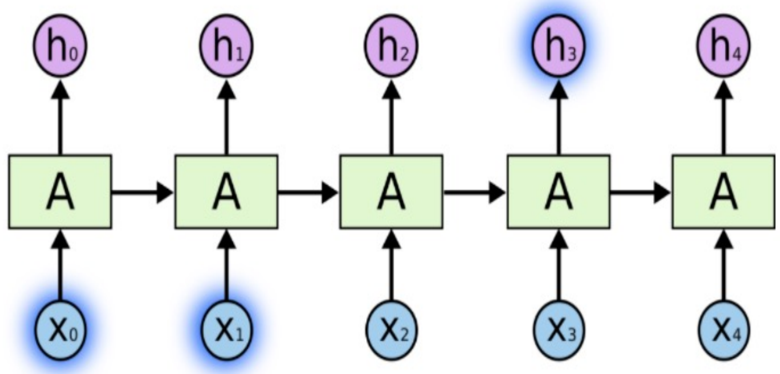
- Incredible success applying RNNs to language modeling and sequence learning problems

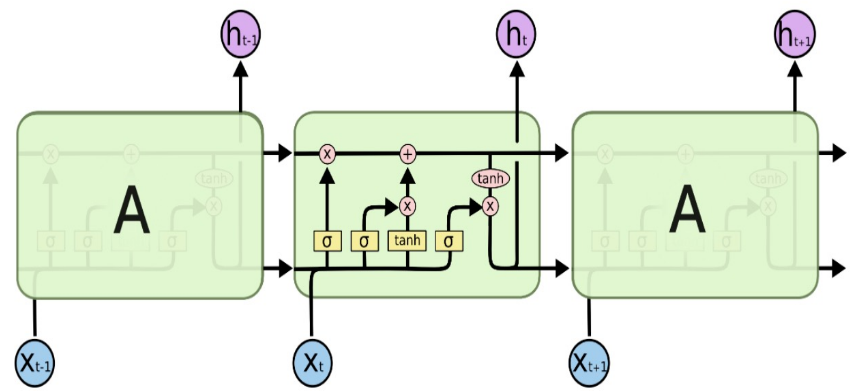
Task	**Input Sequence**	**Output Sequence**
Machine translation (Sutskever et al. 2014)	English	French
Question answering (Bordes et al. 2014)	Question	Answer
Speech recognition (Graves et al. 2013)	Voice	Text
Handwriting prediction (Graves 2013)	Handwriting	Text
Opinion mining (Irsoy et al. 2014)	Text	Opinion expression

LSTM

- "Long short-term memory" – Recurrent NN (LSTM-RNN)

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}) = \overrightarrow{LSTM}(\mathbf{x}_t)$$





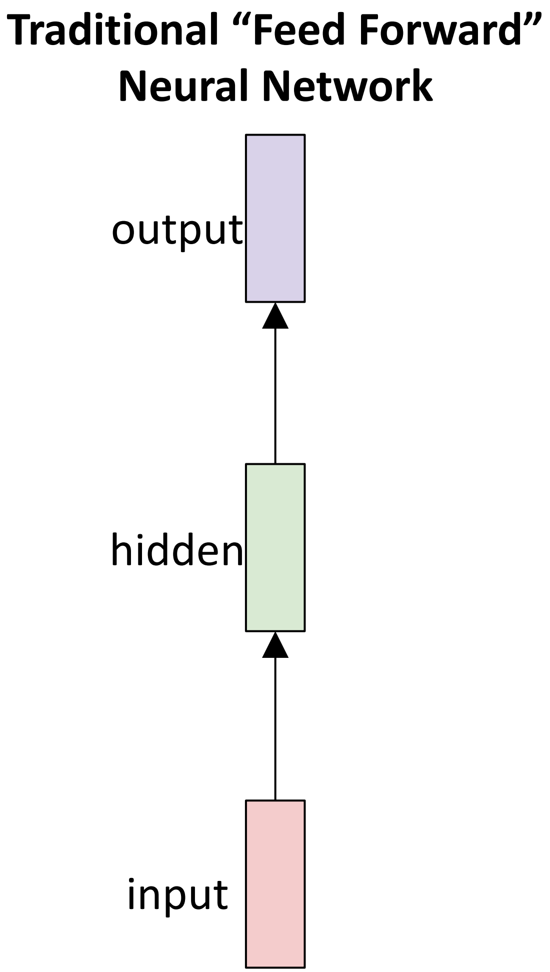
The repeating module in an LSTM contains four interacting layers.

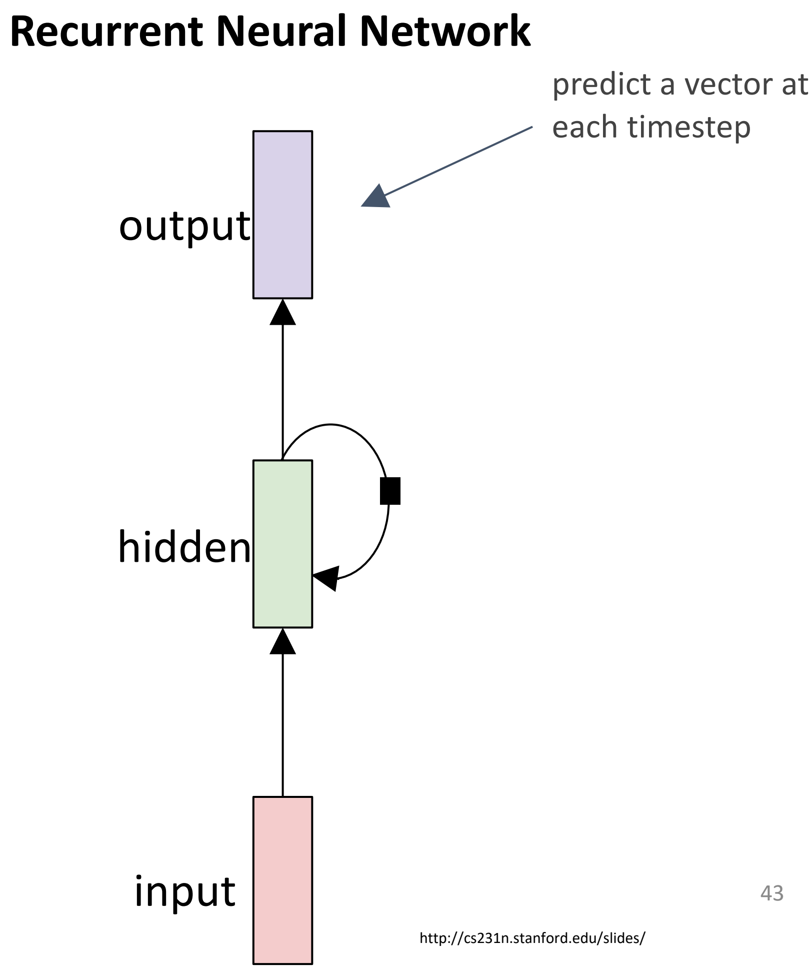
Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780.

Image Credits from Christoph

RNN models dynamic temporal dependency

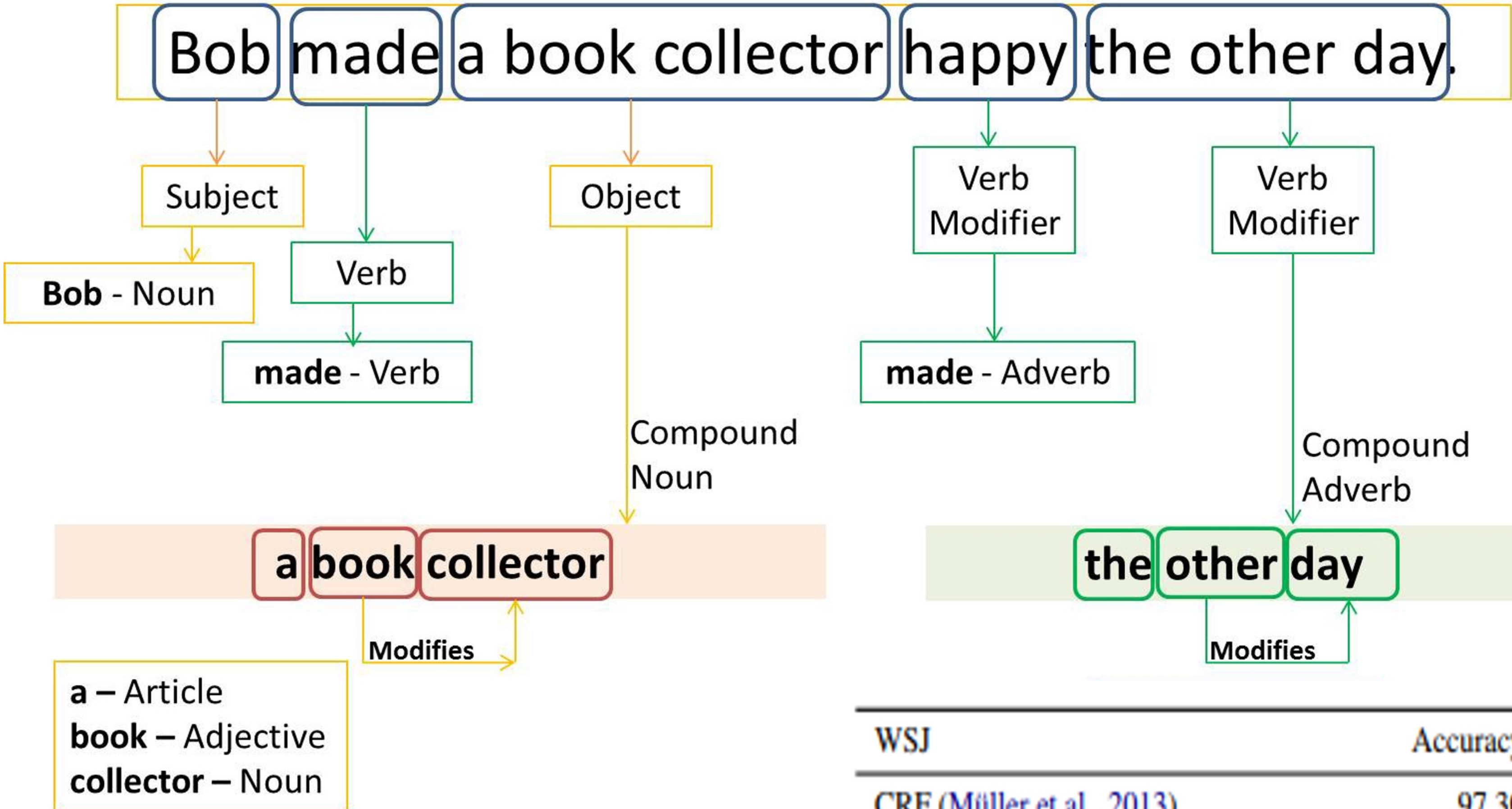
- Make fully-connected layer model each unit recurrently
- Units form a directed chain graph along a sequence
- Each unit uses recent history and current input in modeling





http://cs231n.stanford.edu/slides/

POS tagging (solved by CovNet or RNN-LSTM)



https://nlp.stanford.edu/software/tagger.shtml
https://www.nltk.org/book/ch05.html

WSJ	Accuracy
CRF (Müller et al., 2013)	97.30
Convnet (dos Santos and Zadrozny, 2014)	97.32
bi-LSTM (Ling et al., 2015)	97.36
bi-LSTM (Plank et al., 2016)	97.22
CNN (this work)	97.30

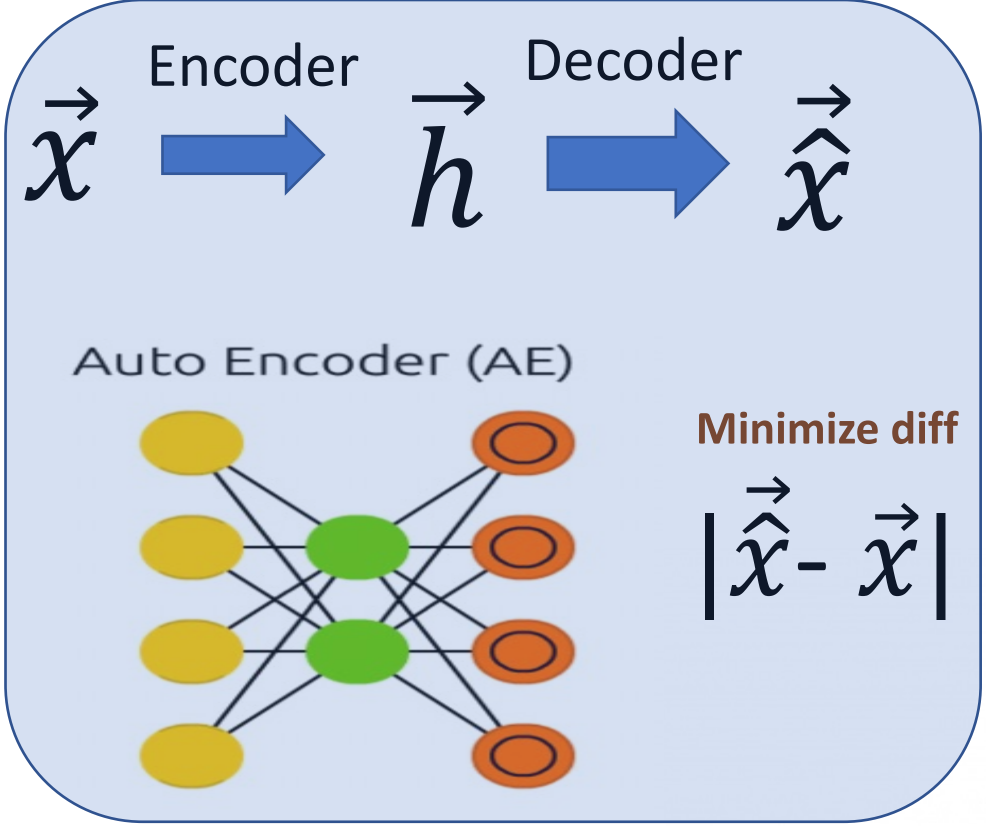
Table 2: Tagging accuracy on the WSJ test set.

http://cs231n.stanford.edu/slides/

RNN can models variable-length input / output

Anything requiring long-range patterns	Anything generative
• Question detection • Natural language context understanding • Entity disambiguation • Sentence embedding	• Machine translation • Natural language generation • Question answering • Skip-thoughts

http://cs231n.stanford.edu/slides/



RNN can models variable-length input / output

Seq2Seq

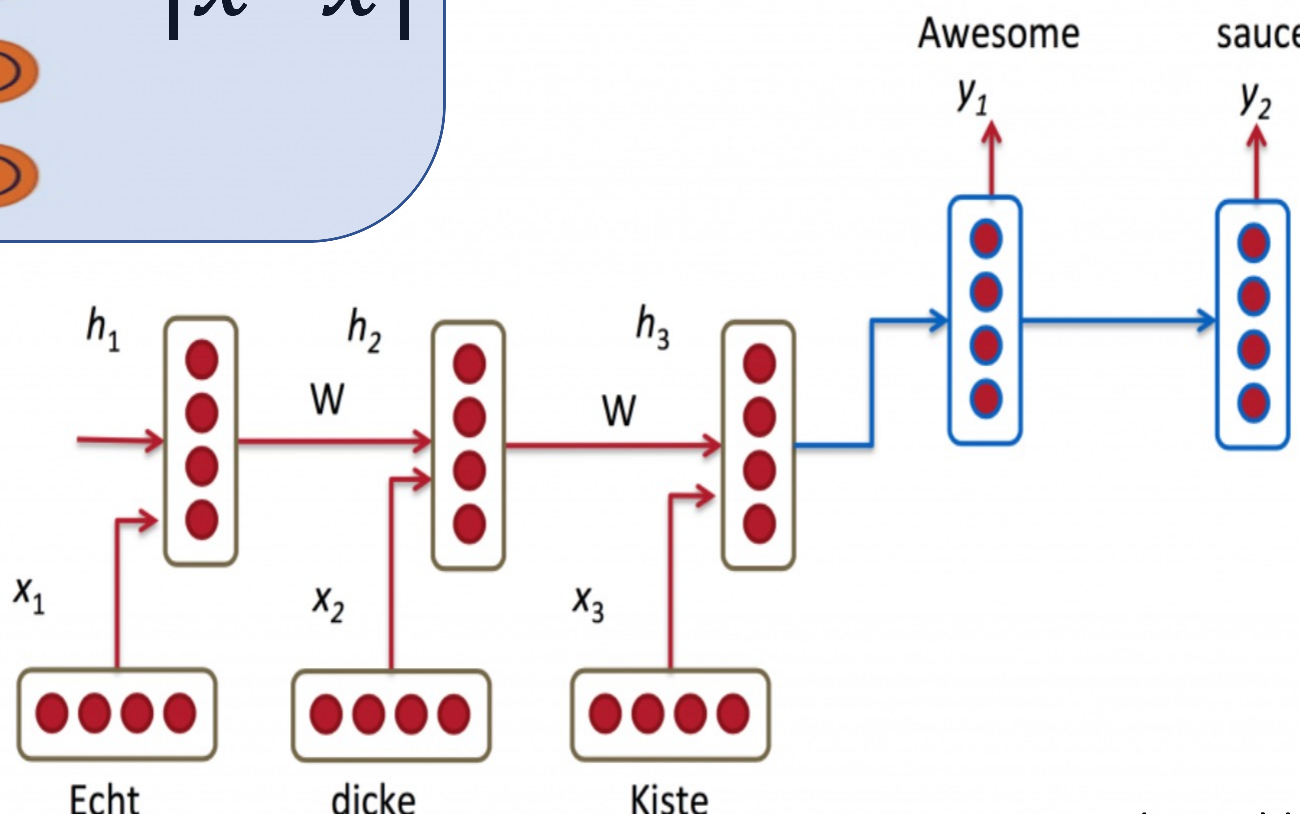


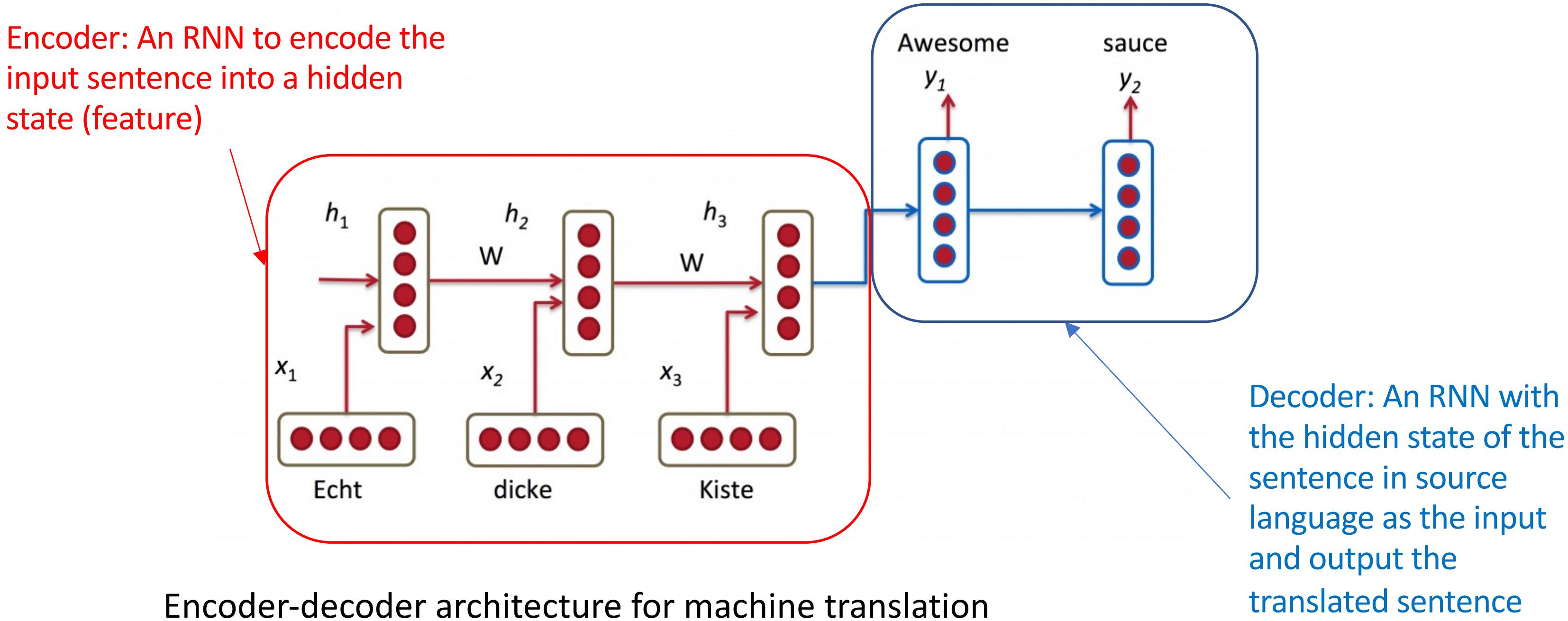
Image credit : wildML

Seq2Seq for Machine Translation

In machine translation, the input is a sequence of words in source language, and the output is a sequence of words in target language.

Two LSTMs for Machine Translation (German to English)

- Encoder LSTM (on Germany)
- Decoder LSTM (on English)



Encoder-decoder architecture for machine translation

Adapt from Professor Qiang Yang of HK UST

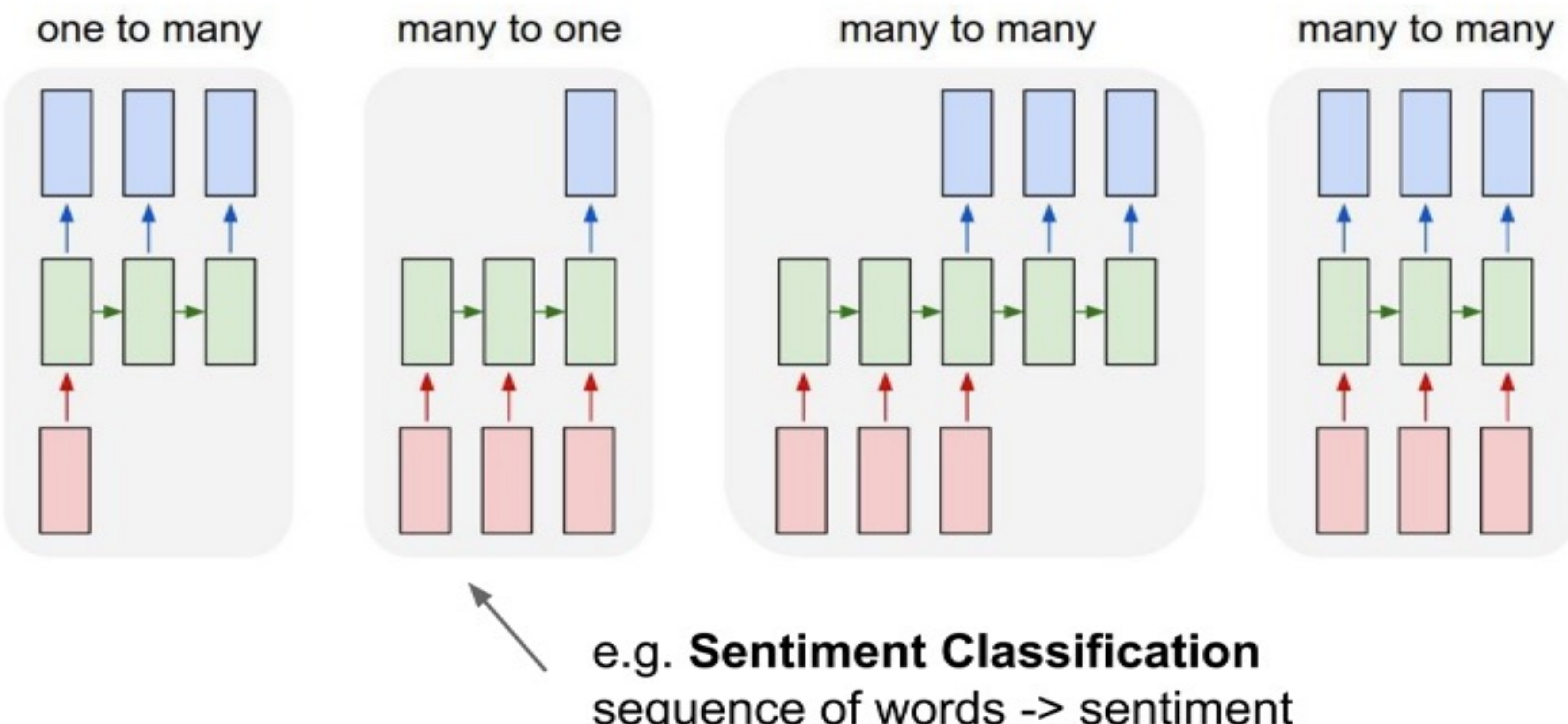
Seq2Seq for more Sequence-to-Sequence Generation Tasks

Given source sentences, learn an optimal model to automatically generate <u>accurate</u> and <u>diversified</u> target sentences that look like human generated sentences.

Source Sentence → Seq2Seq Model → Target Sentence

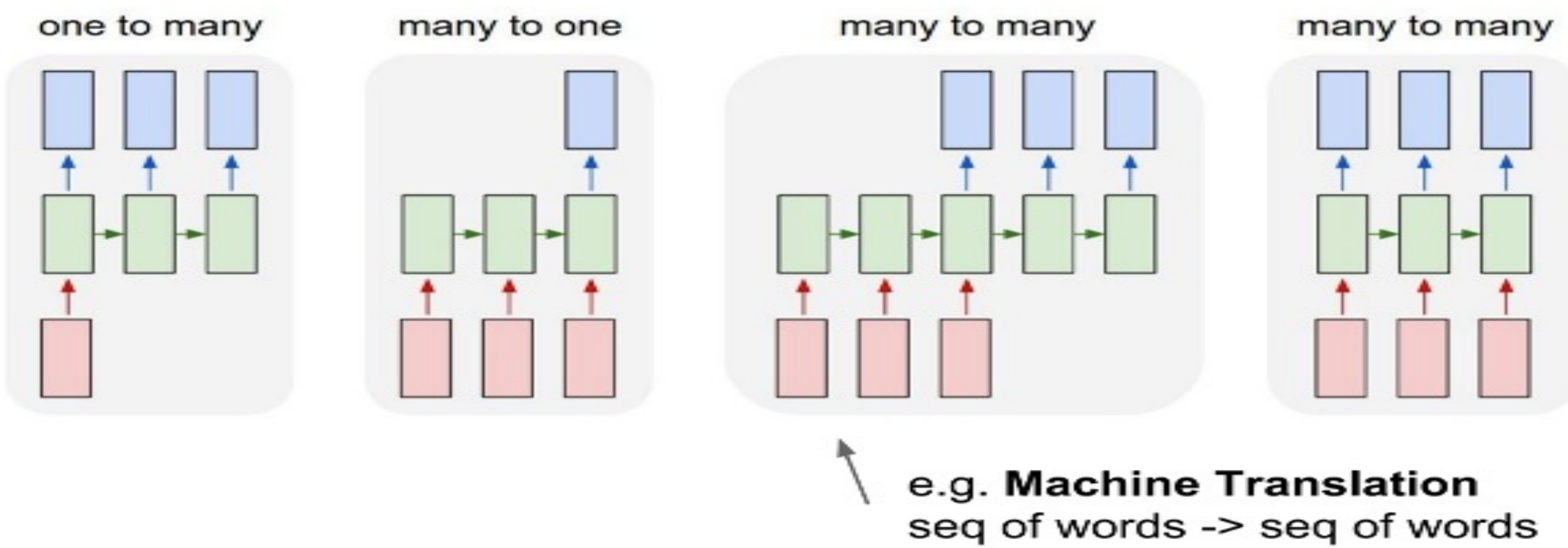
- Paraphrase generation: "How did Trump win the election?" → "How did Trump become president?"
- Dialogue generation: "You know French?" → "Sure do ... my Mom's from Canada"
- Question answering: "What was the name of the 1937 treaty?" → "Bald Eagle Protection Act"
- Style Transfer: "Just a dum funny question hahahaha" → "Just a senseless , funny question."

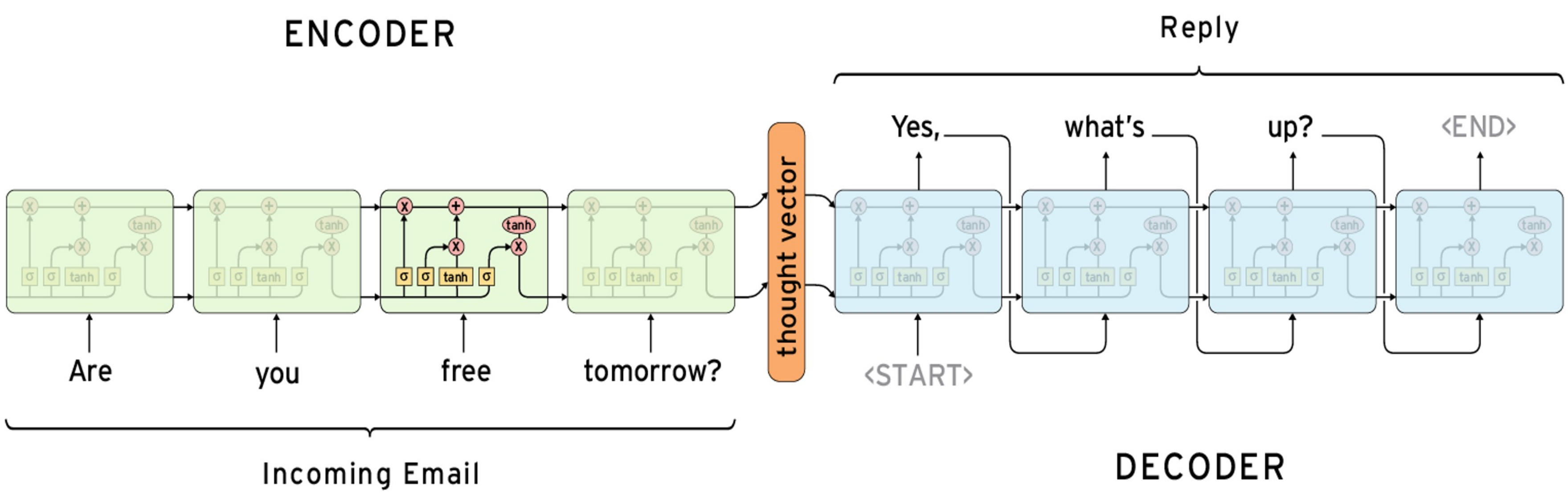
Recurrent Neural Networks (RNNs) can handle



http://cs231n.stanford.edu/slides/

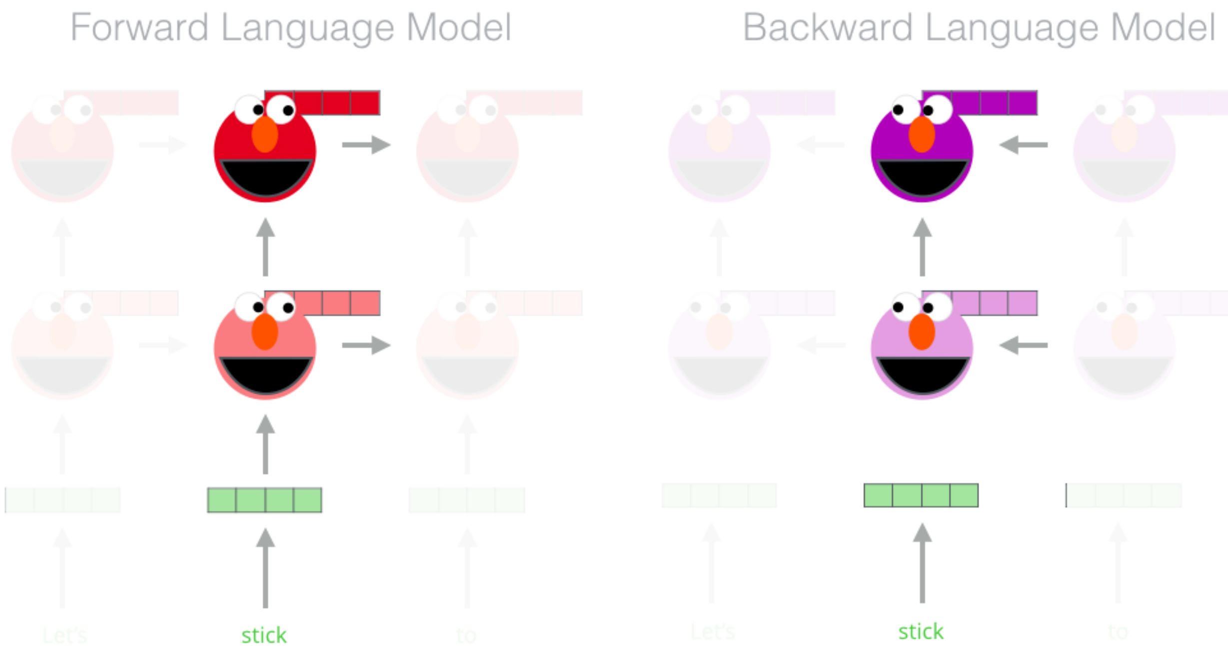
Recurrent Neural Networks (RNNs) can handle





Embedding of "stick" in "Let's stick to" - Step #2

1- Concatenate hidden layers



2- Multiply each vector by a weight based on the task

x s_2

x s_1

x s_0

3- Sum the (now weighted) vectors

ELMo embedding of "stick" for this task in this context

contextual embedding

ELMo's embedding of a word given the sentence is the concatenation of its biLSTM's hidden states for the word.

UVA CS 4774: Machine Learning

S3: Lecture 18:

Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

University of Virginia

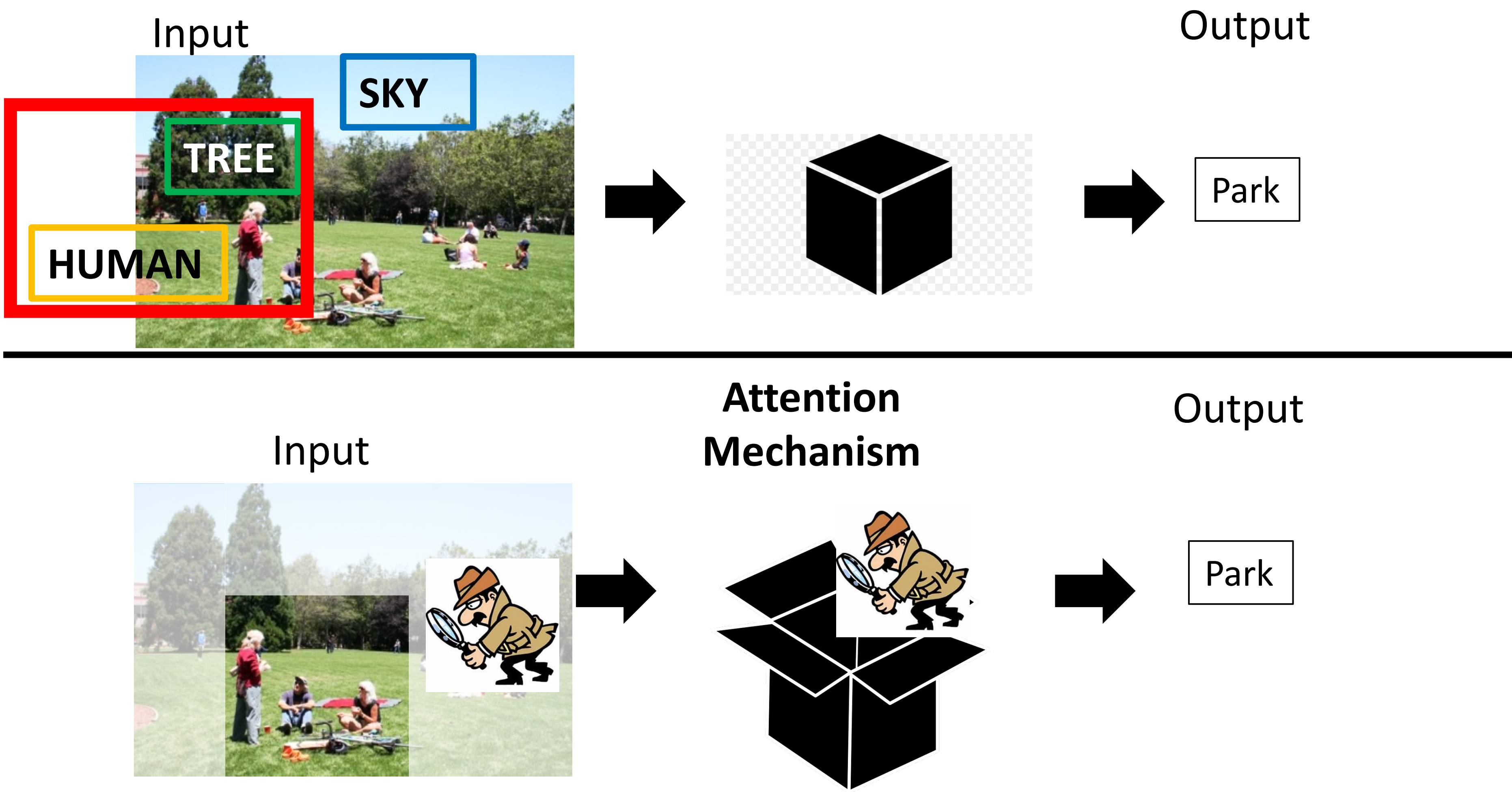
Department of Computer Science

Module III

Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

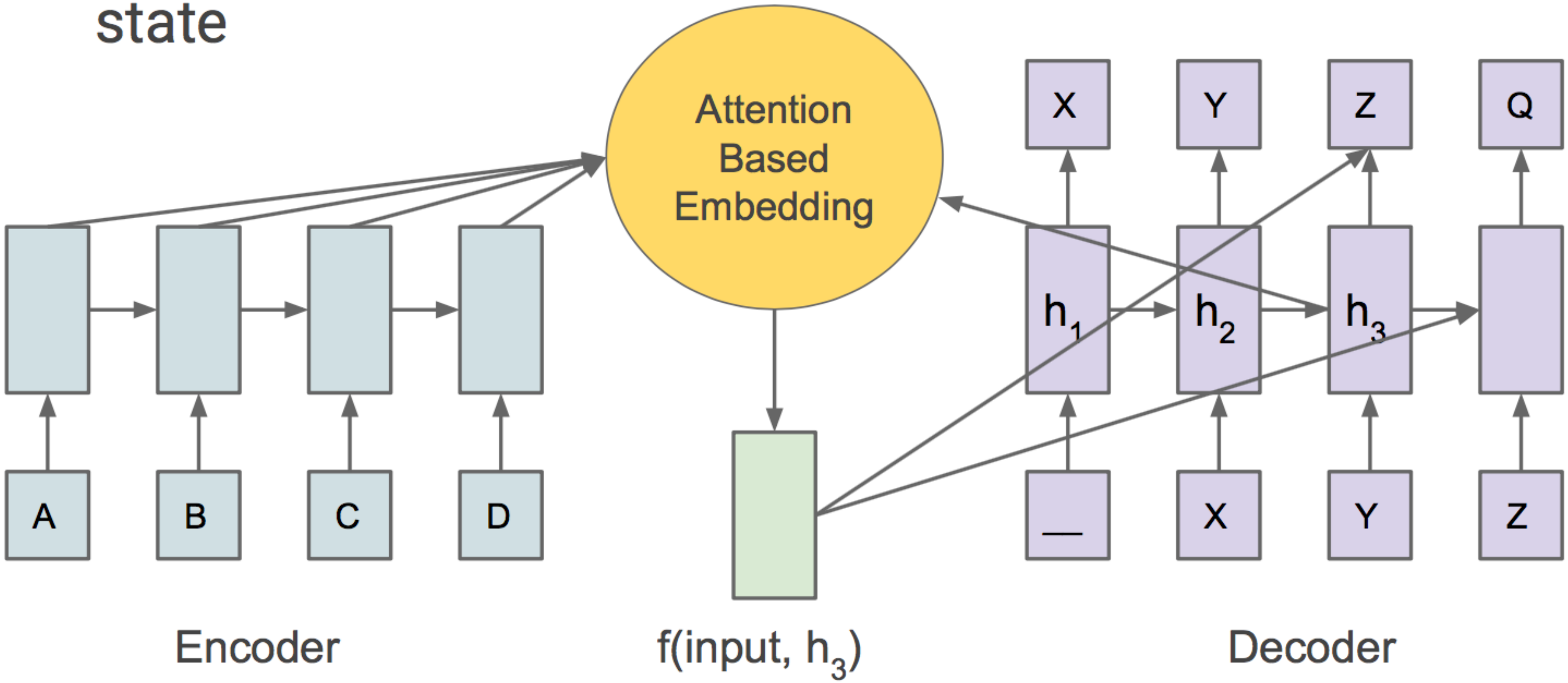
Attention Trick



Attention Trick:

Seq2Seq with Attention

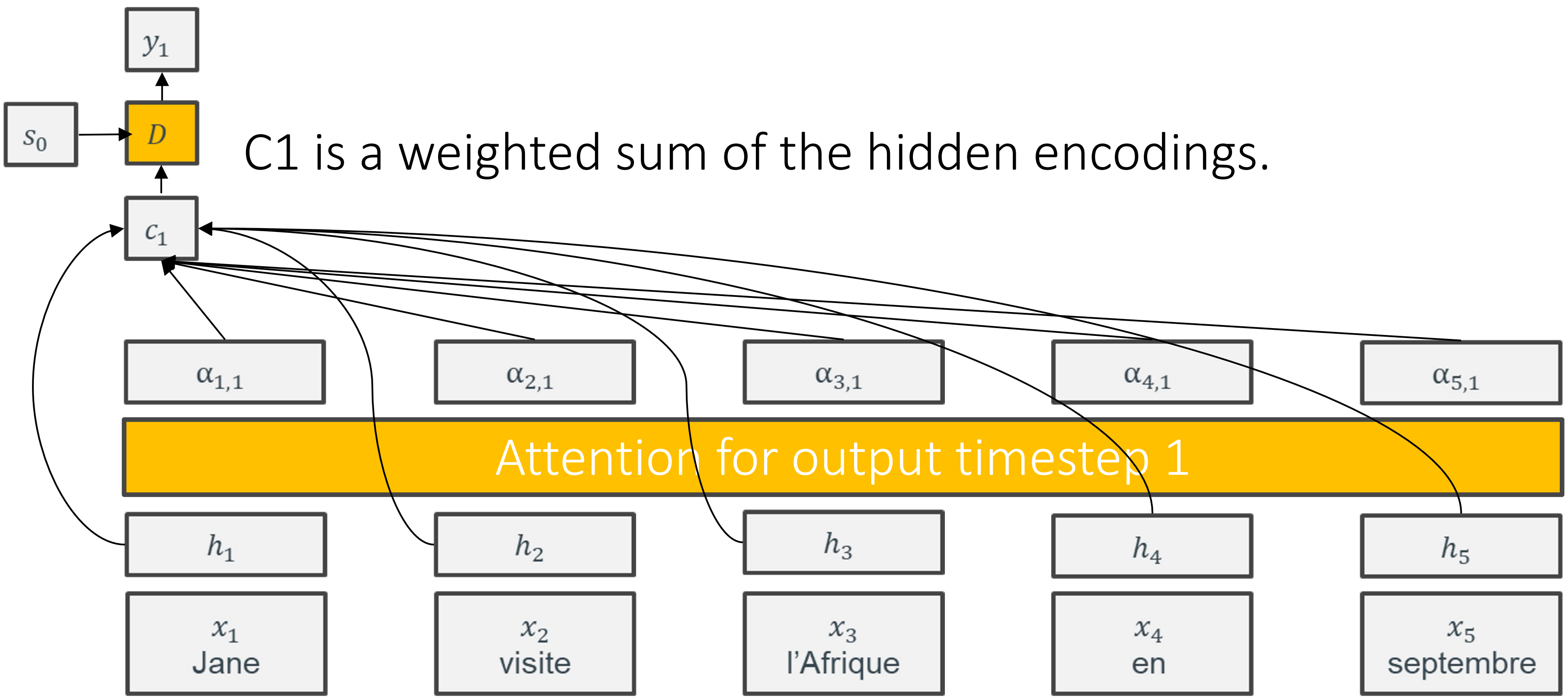
- Embedding used to predict output, and compute next hidden state



Adapt from From NIPS 2017 DL Trend Tutorial

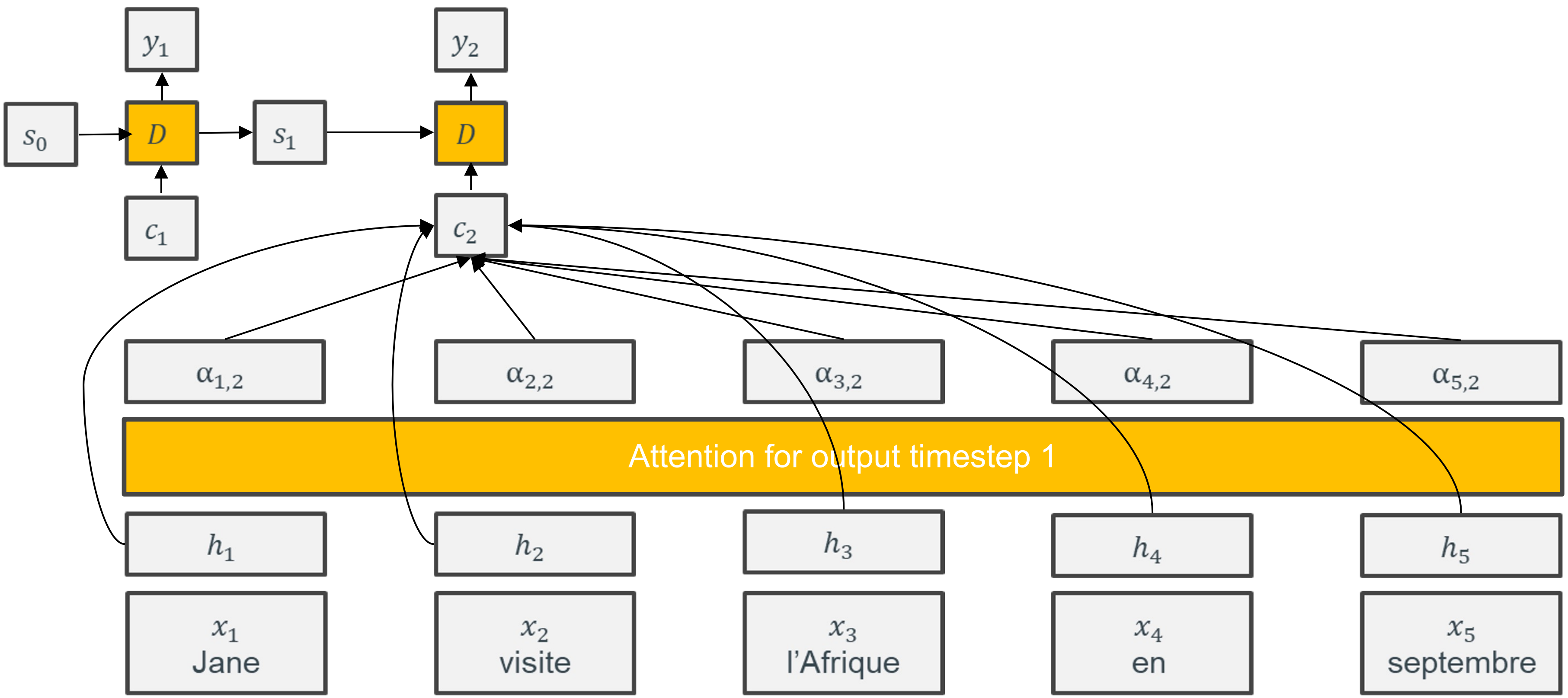
The attention module gives us a weight for each input.

C1 is a weighted sum of the hidden encodings.



Based: Dr. Yangqiu Song's slides

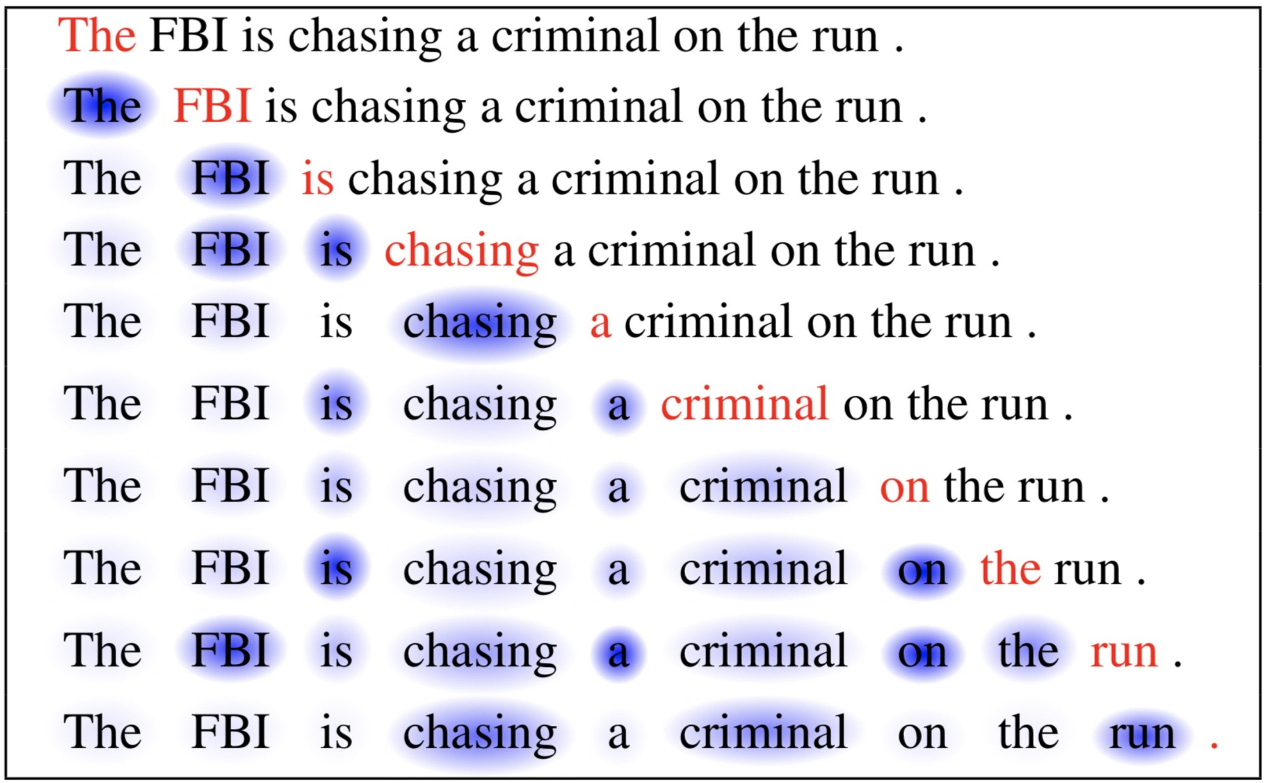
We then repeat for future timesteps.



Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

Self-attention creates attention layers mapping from a sequence to itself.



Transformer: Exploiting Self Attentions

- A Google Brain model.
  - Variable-length input
  - Fixed-length output (but typically extended to a variable-length output)
  - **No recurrence**
  - Surprisingly not patented.

- Uses 3 kinds of attention
  - Encoder self-attention.
  - Decoder self-attention.
  - Encoder-decoder multi-head attention.

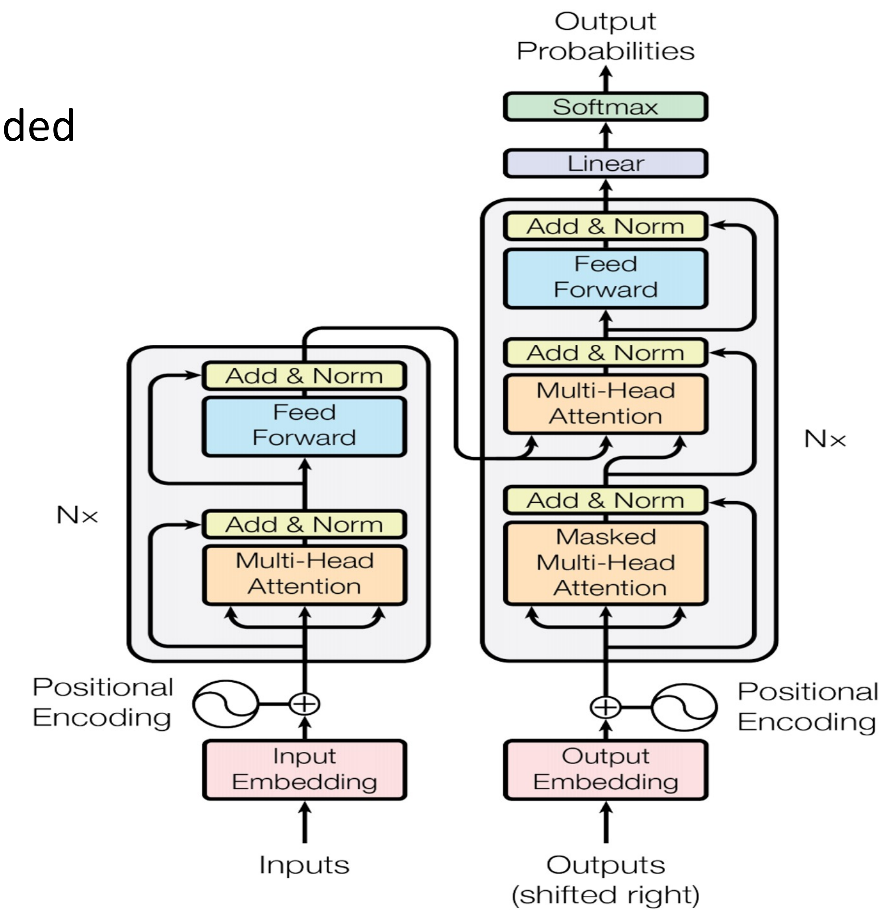
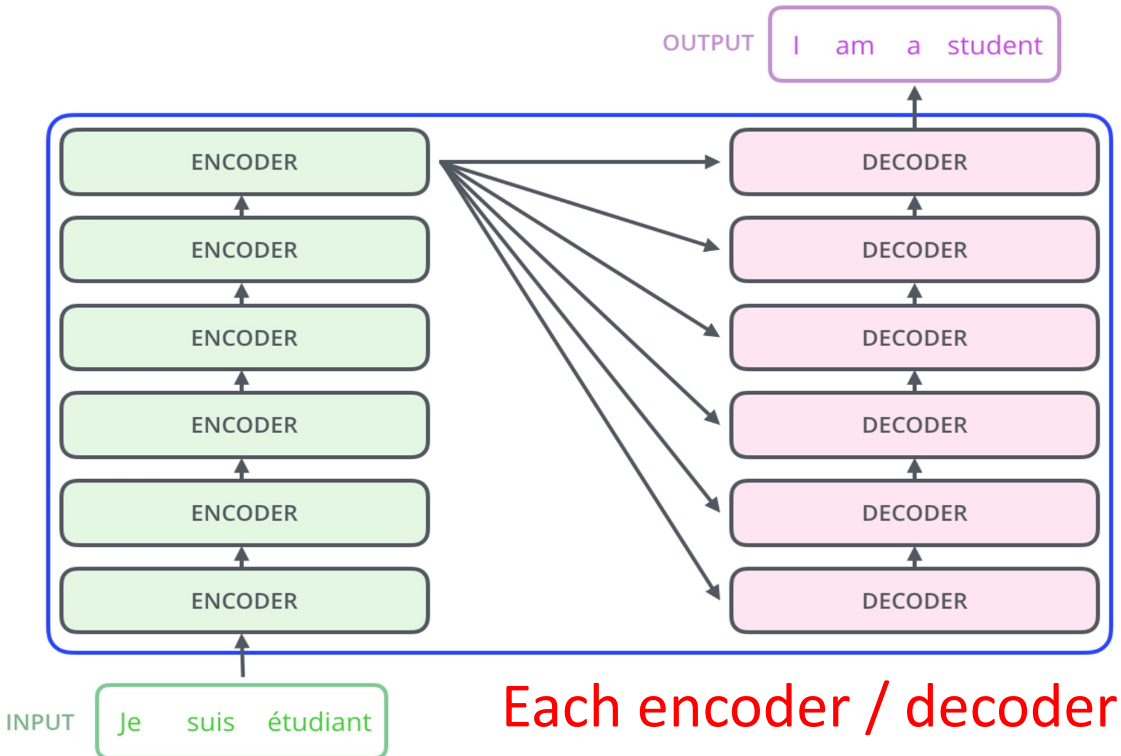


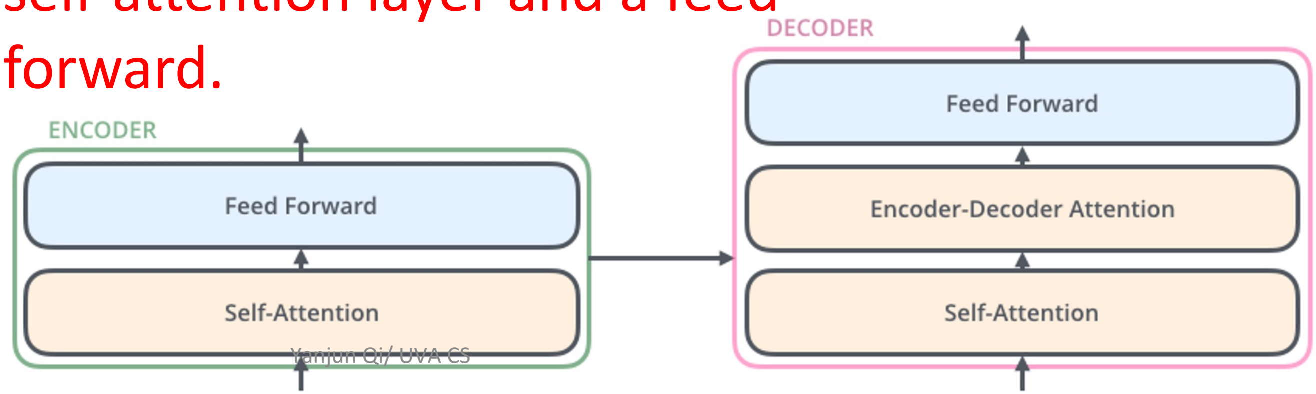
Figure 1: The Transformer - model architecture.

Based: Dr. Yangqiu Song's slides

Transformer is Seq2Seq model



Each encoder / decoder layer has a self-attention layer and a feed forward.

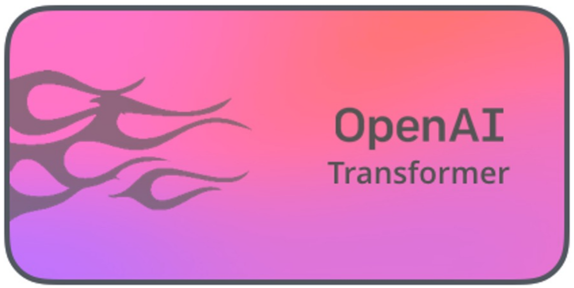


Based: Dr. Yangqiu Song's slides

Roadmap : f() on natural language

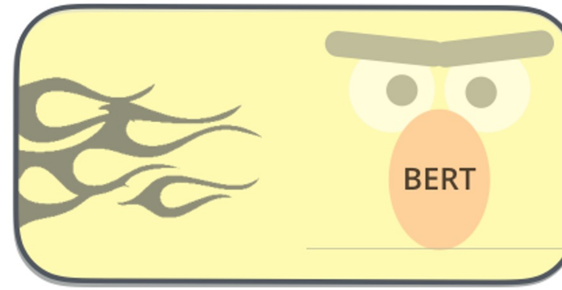
- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA)
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now)
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 …

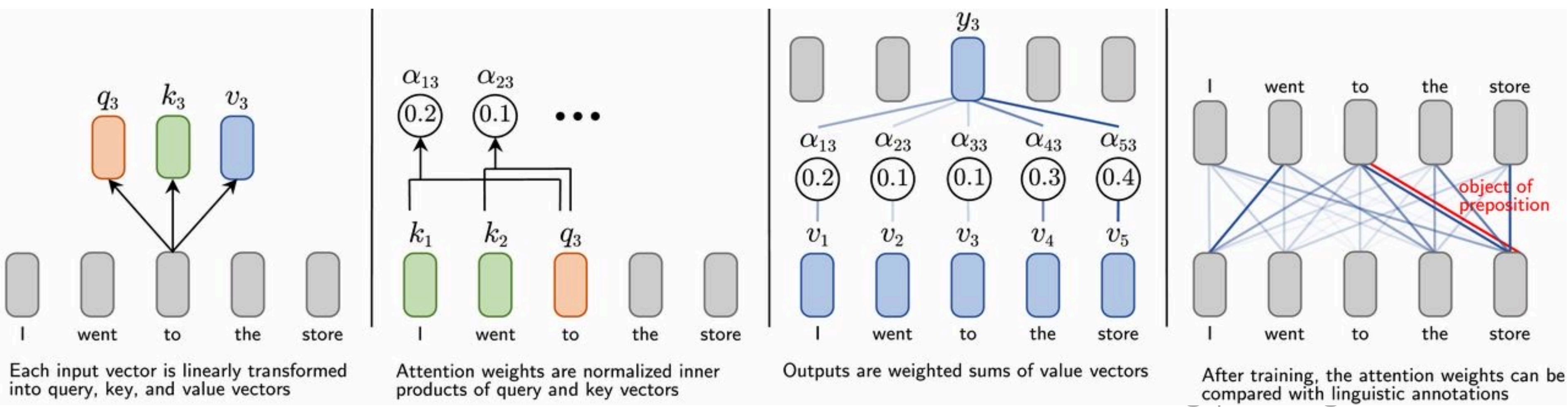
Notable pre-trained NLP models



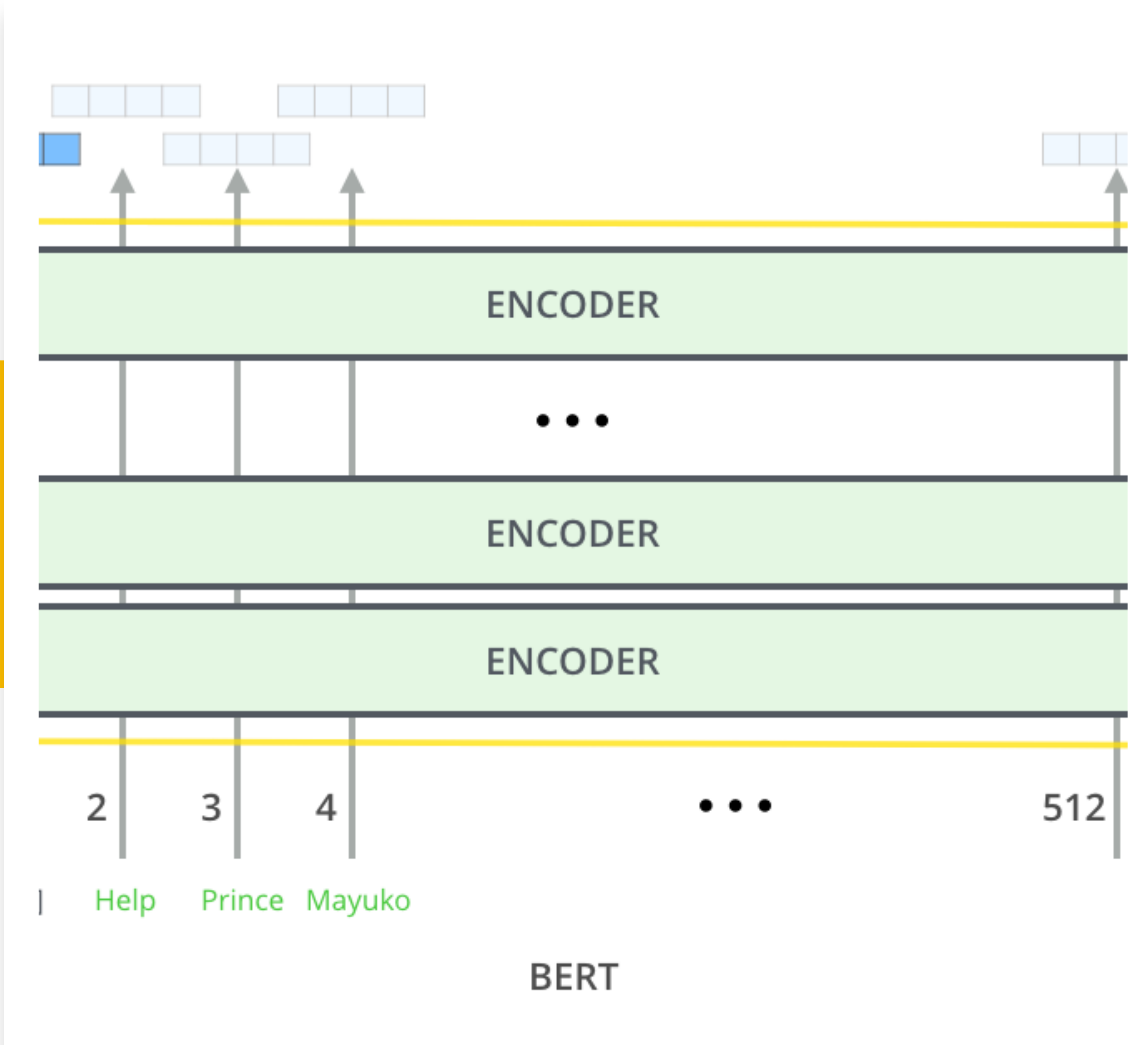
BERT: Bidirectional Encoder Representations from Transformers

Pre-trained transformer encoder for sentence embedding





BERT: **B**idirectional **Encoder R**epresentations from **Transformers**.

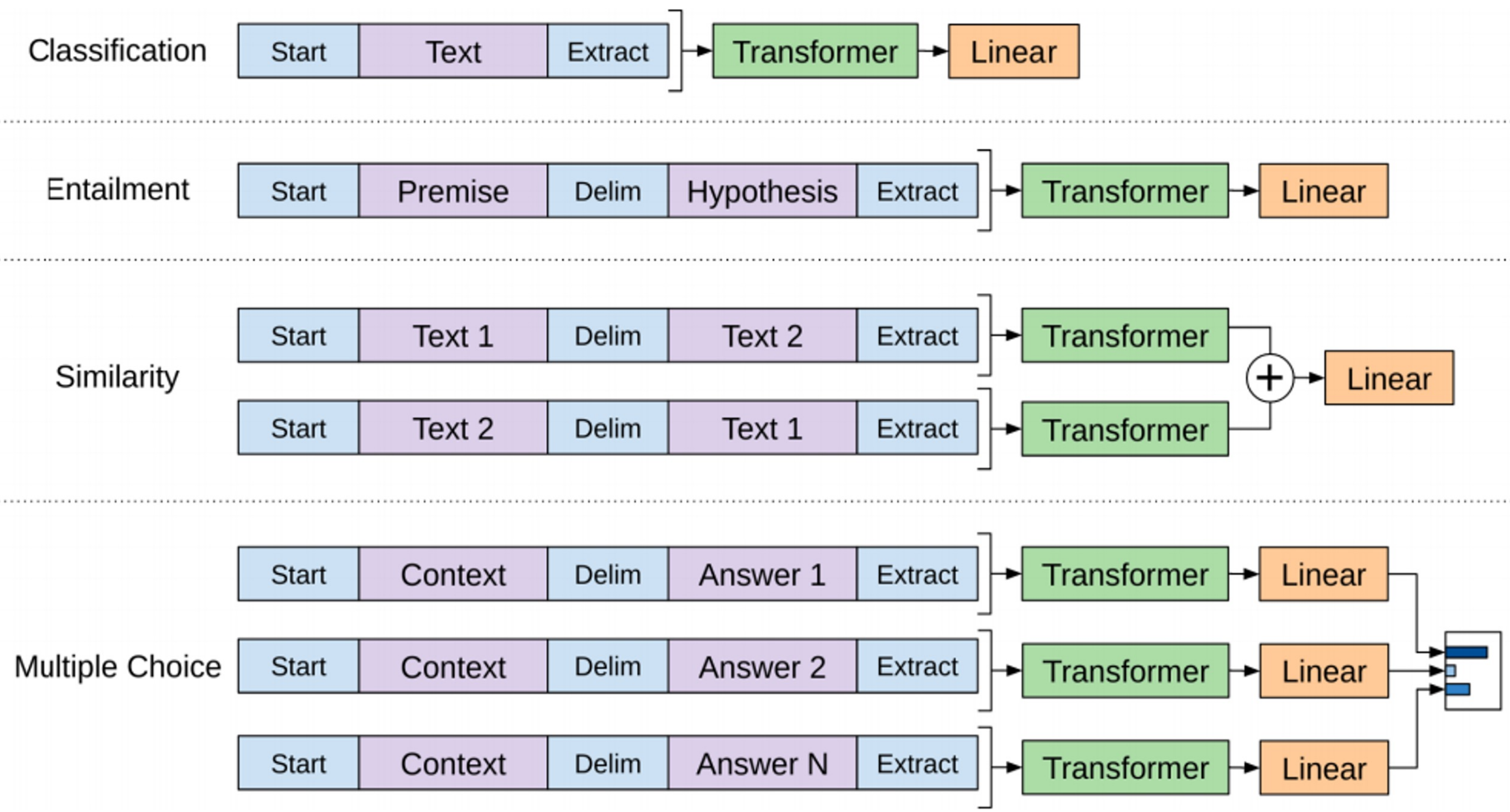


BERT's architecture is just a transformer's encoder stack.

Based: Dr. Yangqiu Song's slides

Open AI's GPT-2: 1.5 billion parameters! Trained on 8M pages from reddit

As with BERT, you can use the pretrained GPT models for any task. Different tasks use the OpenAI transformer in different ways.



GPT: generative pre-training,

GPT 's architecture is just a transformer's decoder stack.

https://colab.research.google.com/drive/18TfLvJ3ITNOeZLeFS3Zf27Fo-PcaEylb?usp=sharing

Today: Neural Network Models on 1D Grid / Language Data

Data: X

Task: y

Representation: : x, f()

Score Function: L()

Search/Optimization : argmin()

Models, Parameters, metrics

X: 1D Grid Text

Y: Many possible Kinds,

e.g., 1D Grid Text

Word2Vec/ Recurrent / Transformer

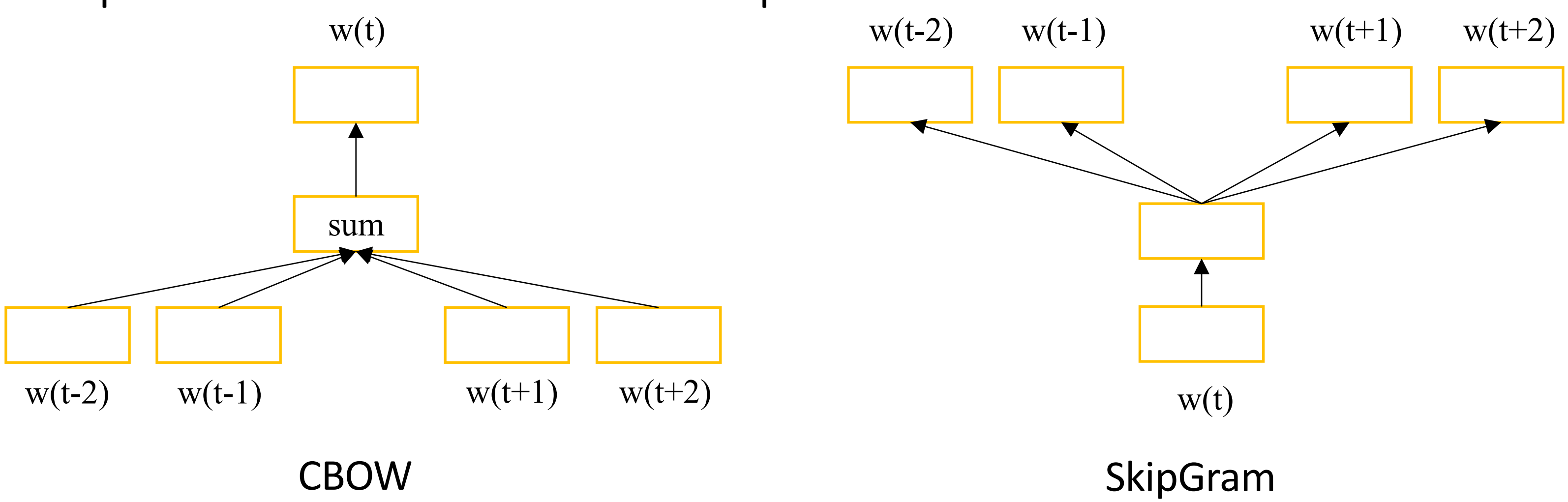
Many self-supervised kinds, e.g., skip-gram / mask LM

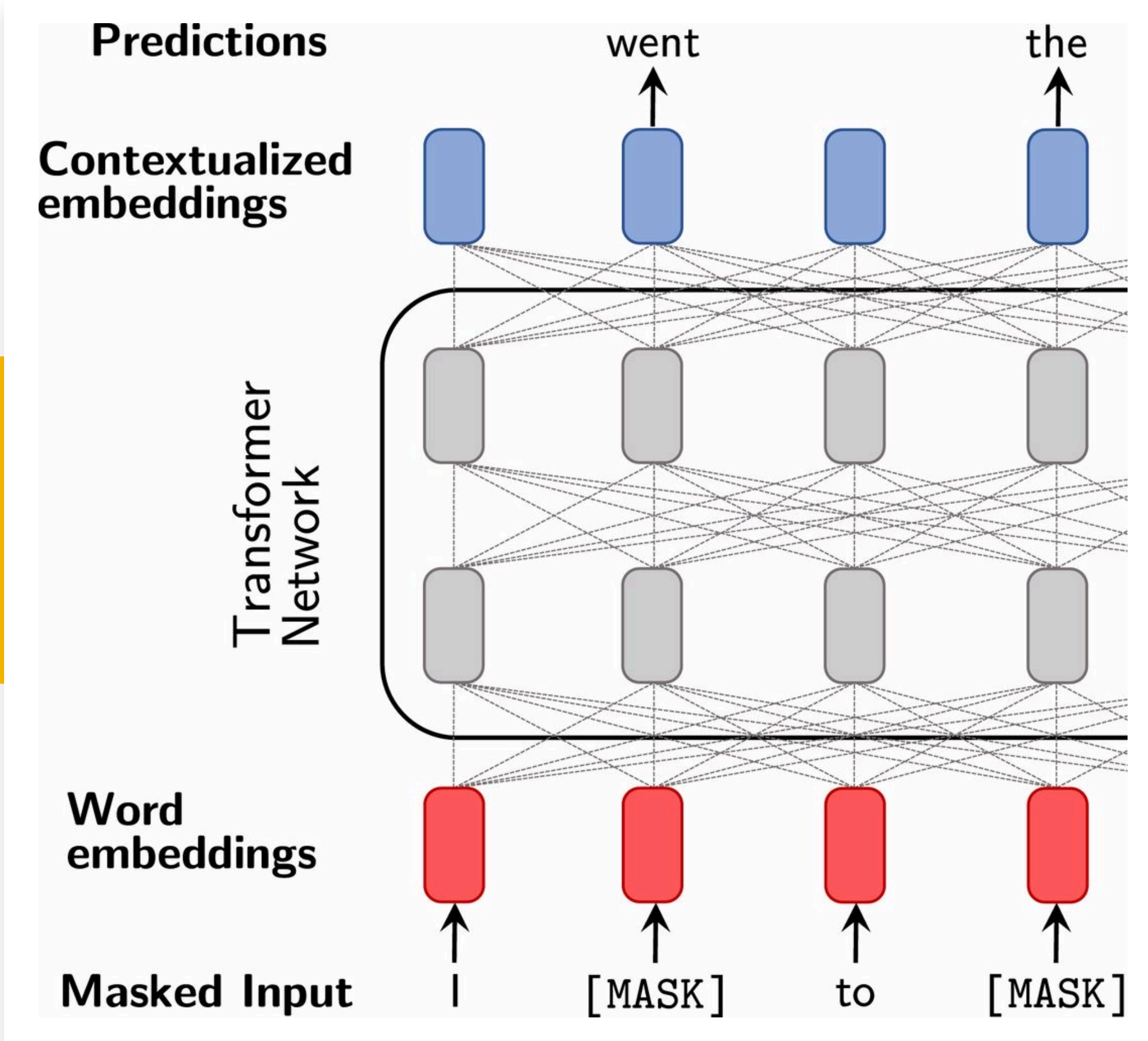
SGD / Backprop

weights, bias / hyperparameters / Accuracy / F1 / more metris

Word2vec: CBOW / SkipGram (Basic Word2Vec)

- Distributed representations of words and phrases and their compositionality (NIPS 2013, Mikolov et al.)
- CBOW
  - predict the input tokens based on context tokens
- SkipGram
  - predict context tokens based on input tokens





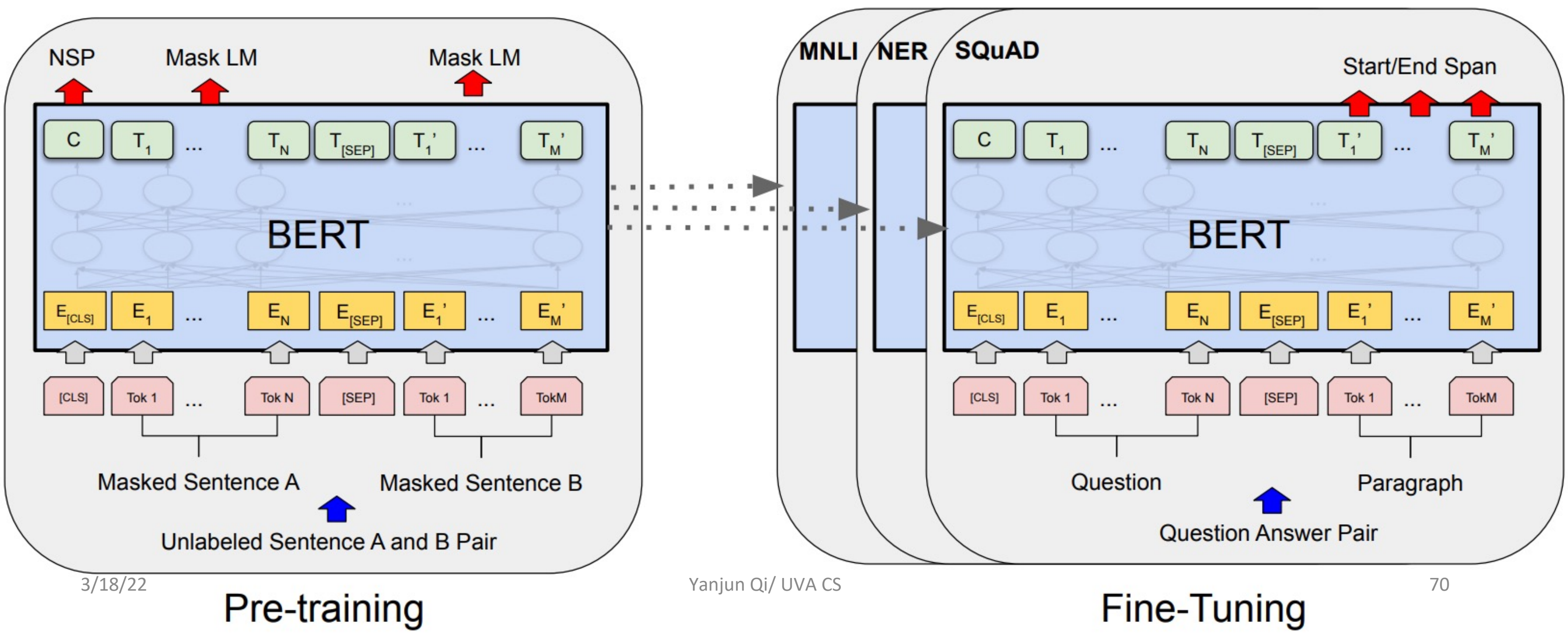
BERT is trained just like a skip-gram model.

Based: PNAS 19

BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL 2019, Devlin et al.)

- Denoising Auto Encoder
- [MASK]: a unique token introduced in the training process to mask some tokens
- Predict masked tokens based on their context information,
- Pre-train and fine-tune

- Intuition: representation should be robust to the introduction of noise
  - Masked Language Model (MLM)



google-research

ALBERT: A lite BERT (2019, Lan et al.)

- proposes Sentence Order Prediction (SOP) task to replace Next Sentence Prediction (NSP)
- in NSP, the negative next sentence is sampled from other passages that may have different topics with the current one, turning the NSP into a far easier topic model problem.
- in SOP, two sentences that exchange their position are regarded as a negative sample, making the model concentrate on the coherence of the semantic meaning.

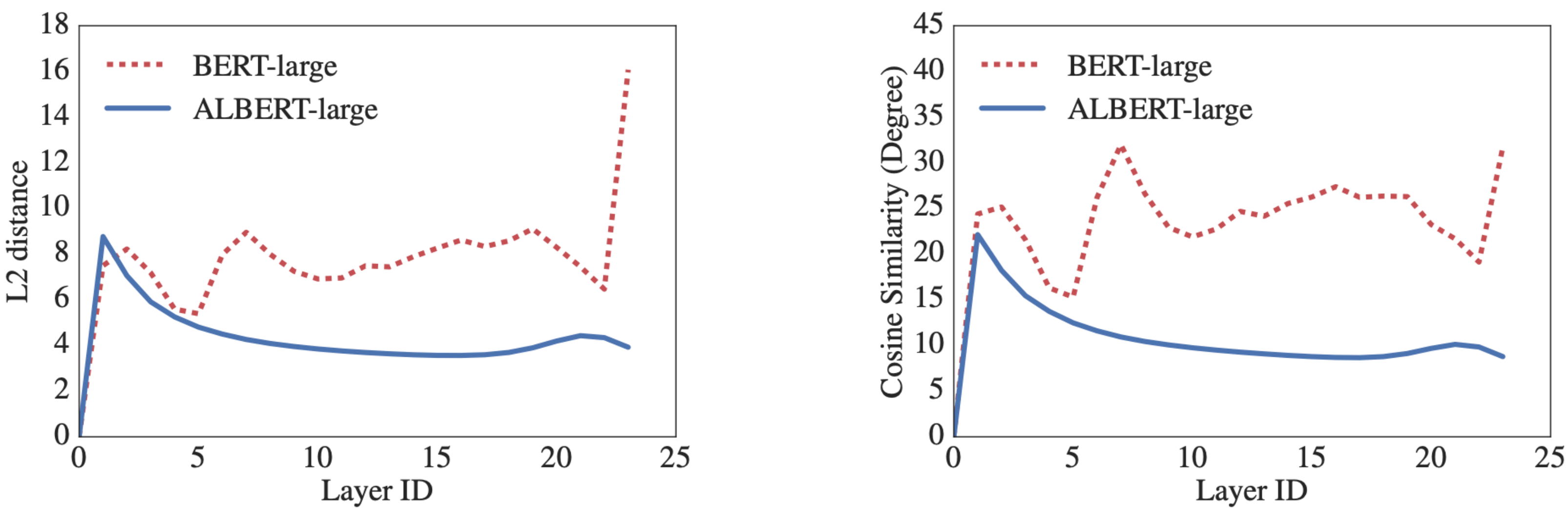
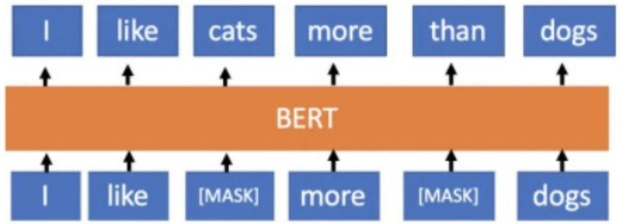
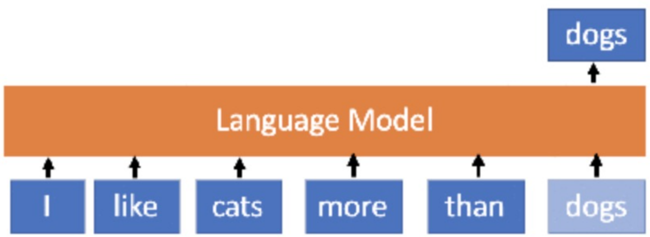
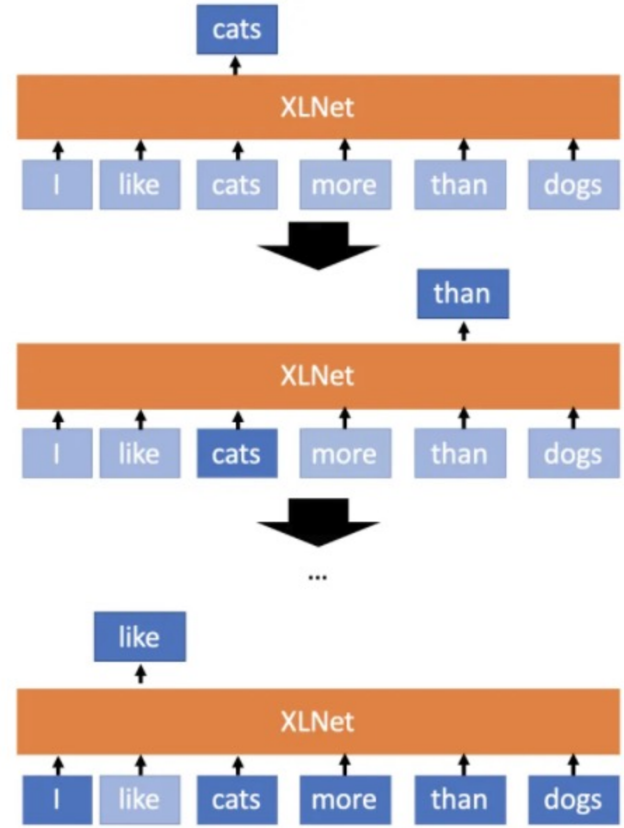


Figure 1: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.





The prediction scheme for a traditional language model. Shaded words are provided as input to the model while unshaded words are masked out.

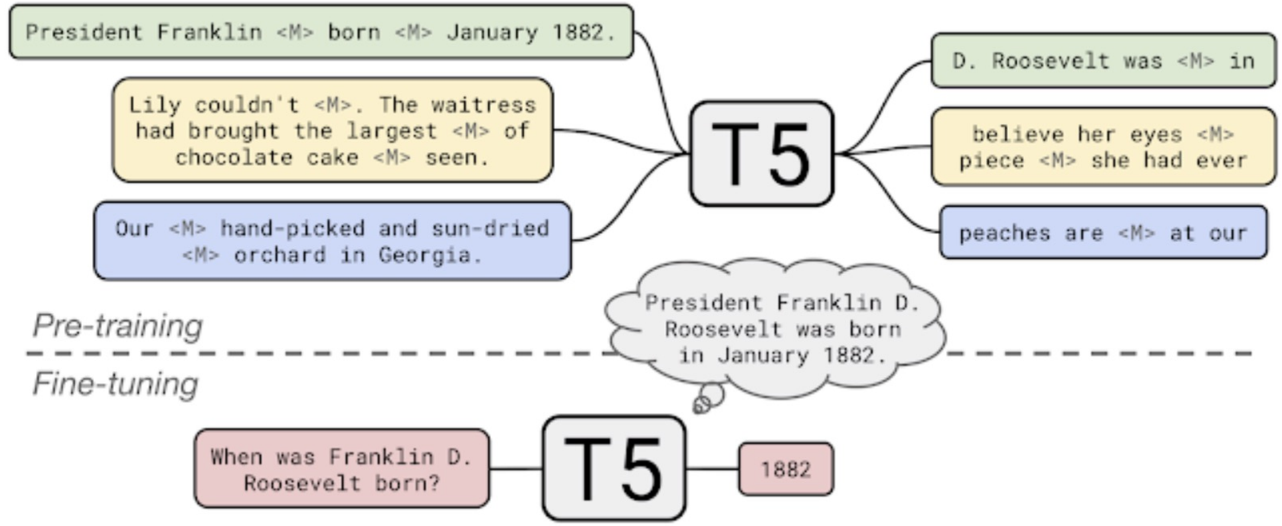


XLNet (Generalized autoregressive pretraining for language understanding(NeurIPS 2019, Yang et al.)

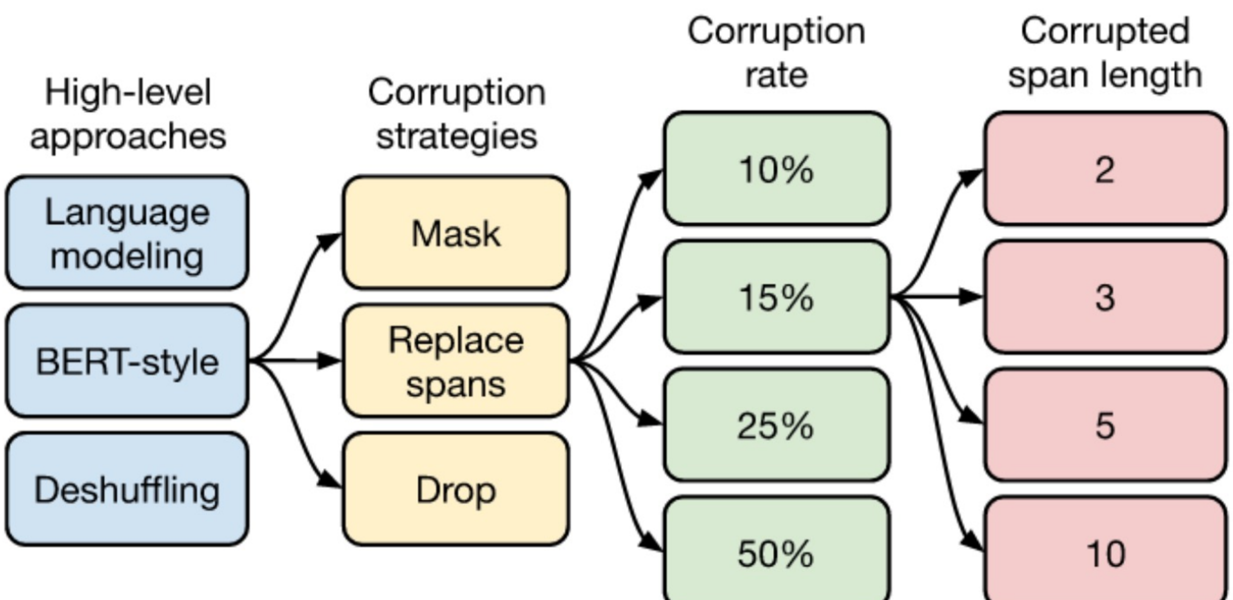
- Transformer-XL: Extra Long Transformer
  - Transformer uses fix length. So can not be too long range
  - So adding recurrence mechanism among segments + relative encoding scheme
- XLNetPLM: Permutation Language Model
  - learning bidirectional contexts by permutation

google-research

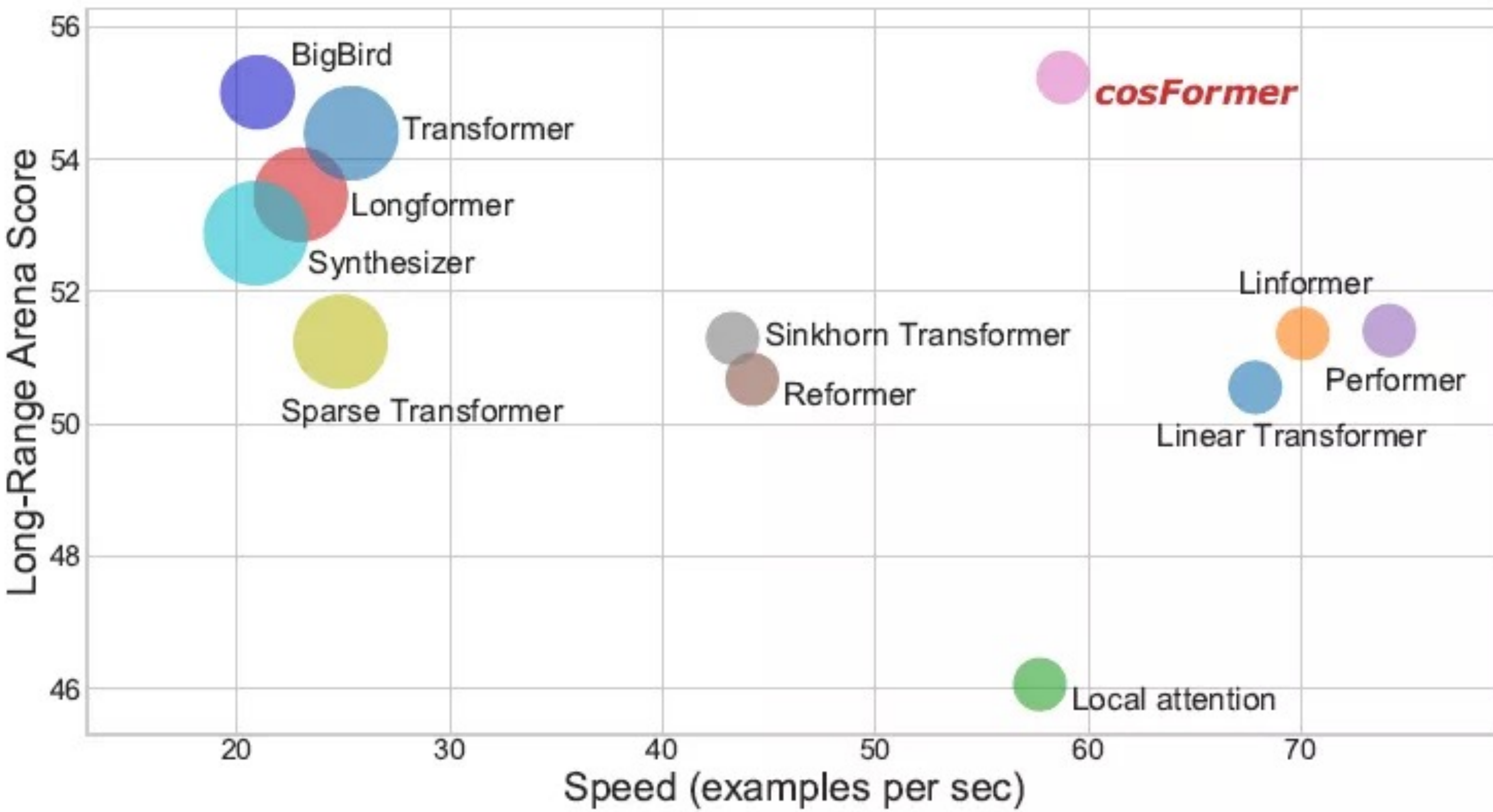
T5: Even more noise

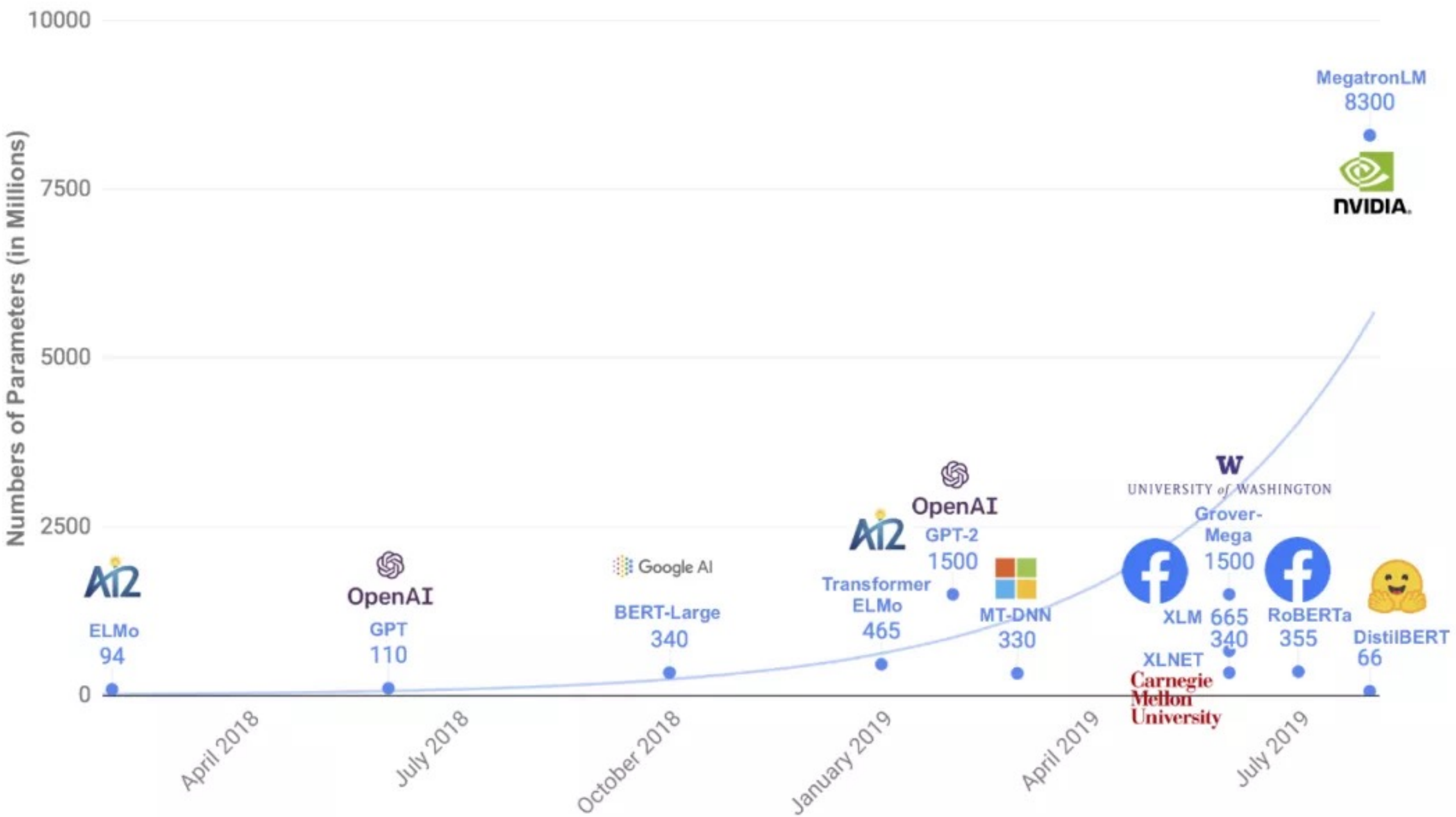


During pre-training, T5 learns to fill in dropped-out spans of text (denoted by <M>) from documents in C4. To apply T5 to closed-book question answer, we fine-tuned it to answer questions without inputting any additional information or context. This forces T5 to answer questions based on “knowledge” that it internalized during pre-training.



Various new transformer models




Numbers of Parameters (in Millions)
10000
7500
5000
2500
0
ELMo
94
GPT
110
BERT-Large
340
Transformer ELMo
465
GPT-2
1500
MT-DNN
330
XLM 665
XLNET
340
Grover-Mega
1500
RoBERTa
355
DistilBERT
66
MegatronLM
8300
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019

Today Recap: Neural Network Models on 1D Grid / Language Data

Data: X

Task: y

Representation: : x, f()

Score Function: L()

Search/Optimization : argmin()

Models, Parameters, metrics

X: 1D Grid Text

Y: Many possible Kinds,

e.g., 1D Grid Text

Word2Vec/ Recurrent / Transformer

Many kinds, e.g.,
Cross Entropy Loss

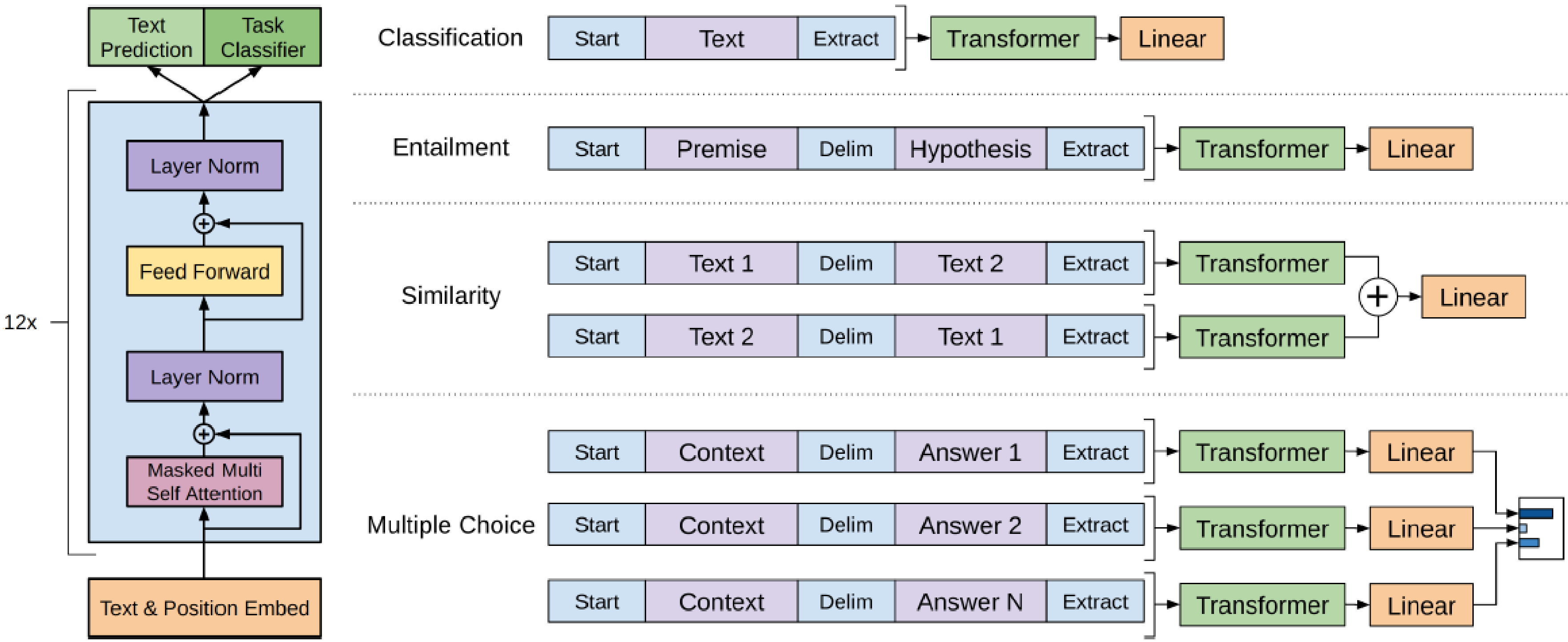
SGD / Backprop

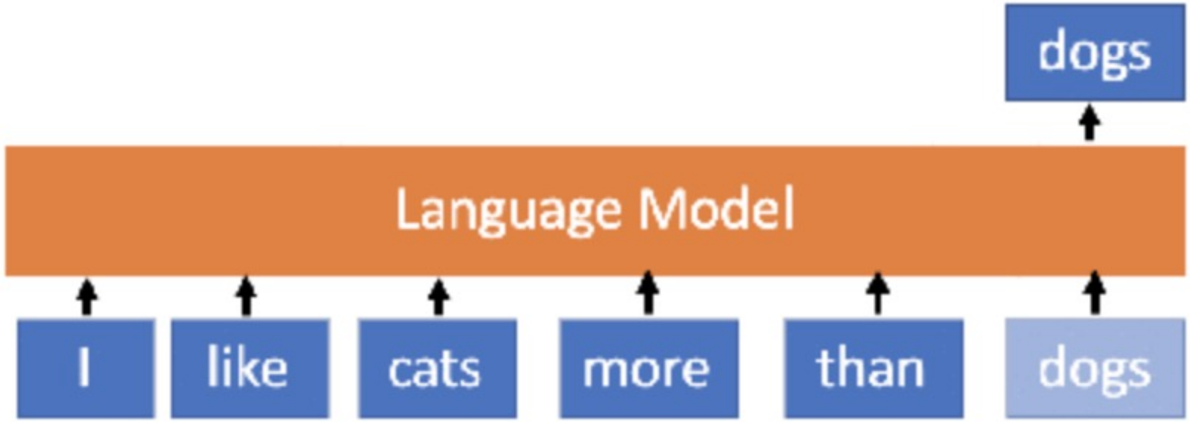
weights, bias /
hyperparameters / Accuracy /
F1 / more metris

References

- ❑ Dr. Yann Lecun's deep learning tutorials
- ❑ Dr. Li Deng's ICML 2014 Deep Learning Tutorial
- ❑ Dr. Kai Yu's deep learning tutorial
- ❑ Dr. Rob Fergus' deep learning tutorial
- ❑ Prof. Nando de Freitas' slides
- ❑ Olivier Grisel's talk at Paris Data Geeks / Open World Forum
- ❑ Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.
- ❑ Dr. Hung-yi Lee's CNN slides
- ❑ NIPS 2017 DL Trend Tutorial

GPT1 - Improving Language Understanding by Generative Pre-Training (Radford et al. 2018)





The prediction scheme for a traditional language model. Shaded words are provided as input to the model while unshaded words are masked out.

Autoregressive Models

$$P(x;\theta) = \prod_{n=1}^{N} P(x_n | x_{<n};\theta)$$

- Each factor can be parametrized by θ, which can be shared.
- The variables can be arbitrarily ordered and grouped, as long as the ordering and grouping is consistent.

Adapt from From NIPS 2017 DL Trend Tutorial