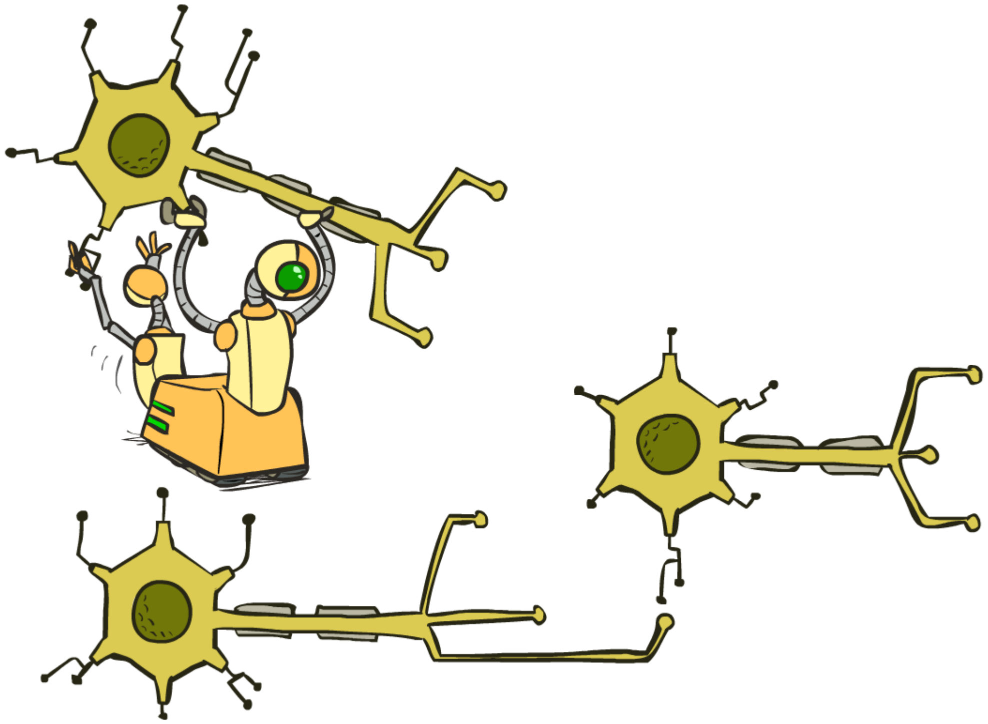
CS 188: Artificial Intelligence

Special Topics: NLP/CV/RL



Instructor: Nicholas Tomlin

[Slides courtesy of Dan Klein, Abigail See, Greg Durrett, Yejin Choi, John DeNero, Eric Wallace, Kevin Lin, Fei-Fei Li, Sergey Levine, Pieter Abbeel, and many others]

Final Review Discussion Sections

- Schedule is posted on Ed!
- Discussions aren't necessarily in the usual rooms
- Exam scope: cumulative, everything taught in lecture up to today, covered in homeworks or discussions, unless explicitly marked as optional content

Time	Topic	TA	Location
Thursday (8/3)			
4pm-5pm	Games/VPI	Linyuan	Cory 521
5pm-6pm	Bayes Nets	Ramanan	Cory 521
6pm-7pm	ML	Eric	Cory 540AB
Friday			
9am-10am	Neural Nets	Ruiqi	Cory 521
2pm-3pm	Games/VPI	Linyuan	Soda 306
3pm-4pm	Bayes Nets	Ramanan	Soda 306
4pm-5pm	MDP/RL	Michael	Soda 306
5pm-6pm	Bayes Nets	Ramanan	Soda 306
Monday			
9am-10am	Games/VPI	Linyuan	Cory 521
10am-11am	Bayes Nets	Ramanan	Cory 521
11am-12pm	Neural Nets	Ruiqi	Cory 521
12pm-1pm	MDP/RL	Stanley	Cory 521
1pm-2pm	ML	Eric	Cory 521
2pm-3pm	NO SECTION		
3pm-4pm	NO SECTION		
4pm-5pm	HMMs	Austen	Cory 521
5pm-6pm	Games/VPI	Linyuan	Cory 521

Natural Language Processing

Goal: Deep Understanding

- Requires knowledge, context, and grounding
- Just starting to see successes

Reality: Shallow Matching

- Requires robustness and scale
- Amazing successes, but fundamental limitations

What tasks do we care about?

Applications

- Machine translation
- Web search
- Dialogue systems

- Automatic speech recognition
- Text-to-speech

- Syntactic parsing, semantic parsing, semantic role labeling, word sense induction, coreference resolution, summarization, sentiment analysis…

Machine Translation

CINÉMA • MÉDIAS

Partage

Disney décide de changer le nom du légendaire studio de cinéma Fox

Le nouveau propriétaire va faire disparaître la mention « Fox » à la 20th Century pour que le nom du studio ne soit pas associé à l'image ultra-conservatrice de la chaîne de télévision Fox News.

Le Monde avec AFP • Publié le 18 janvier 2020 à 06h46

Lecture 1 min.

Révolution dans le monde d'Hollywood. Le célèbre studio de cinéma 20th Century Fox va voir une partie de son nom modifiée et devenir 20th Century Studios, décision prise par son nouveau propriétaire, Disney.

Selon plusieurs médias américains, Disney aurait décidé de modifier le nom du studio pour qu'il ne soit plus associé à la grande chaîne Fox mais aussi et surtout à Fox News, la chaîne d'information en continu.

CINEMA • MEDIA

Sharing

Disney decides to change the name of the legendary Fox film studio

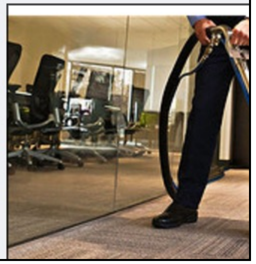
The new owner will remove the mention "Fox" in the 20th Century so that the name of the studio is not associated with the ultra-conservative image of the television channel Fox News.

Le Monde avec AFP • Posted on January 18, 2020 at 06:46

Reading time 1 min.

Revolution in the Hollywood world. The famous film studio 20th Century Fox will see part of its name changed and become 20th Century Studios, decision taken by its new owner, Disney.

According to several American media, Disney has decided to change the name of the studio so that it is no longer associated with the big chain Fox but also and especially with Fox News, the news channel.



Google Translate 2020

Search, Questions, and Reasoning

Web Images Groups News Froogle L

Default (GPT-3.5)

NI How many US states capitals' are also their largest cities?

As of my last update in September 2021, there were eight U.S. states where the capital ci
also the largest city in the state. These states are:

1. Atlanta, Georgia
2. Boston, Massachusetts
3. Denver, Colorado
4. Honolulu, Hawaii
5. Indianapolis, Indiana
6. Oklahoma City, Oklahoma
7. Phoenix, Arizona
8. Salt Lake City, Utah

Please note that this information might have changed since then due to population shifts
other factors. For the most up-to-date information, it's best to consult a recent source or
check official government websites.

NI How many US states capitals' are also their largest cities?

As of my knowledge cutoff in September 2021, there are 17 U.S. state capitals that are also the largest city in their respective states. These are:

1. Phoenix, Arizona
2. Little Rock, Arkansas
3. Denver, Colorado
4. Honolulu, Hawaii
5. Boise, Idaho
6. Indianapolis, Indiana
7. Des Moines, Iowa
8. Baton Rouge, Louisiana
9. Boston, Massachusetts
10. Jackson, Mississippi
11. Lincoln, Nebraska
12. Carson City, Nevada
13. Oklahoma City, Oklahoma
14. Providence, Rhode Island
15. Nashville, Tennessee
16. Salt Lake City, Utah
17. Charleston, West Virginia

Please note that this information might have changed after my last update in September 2021, so it's always a good idea to check the most recent data.

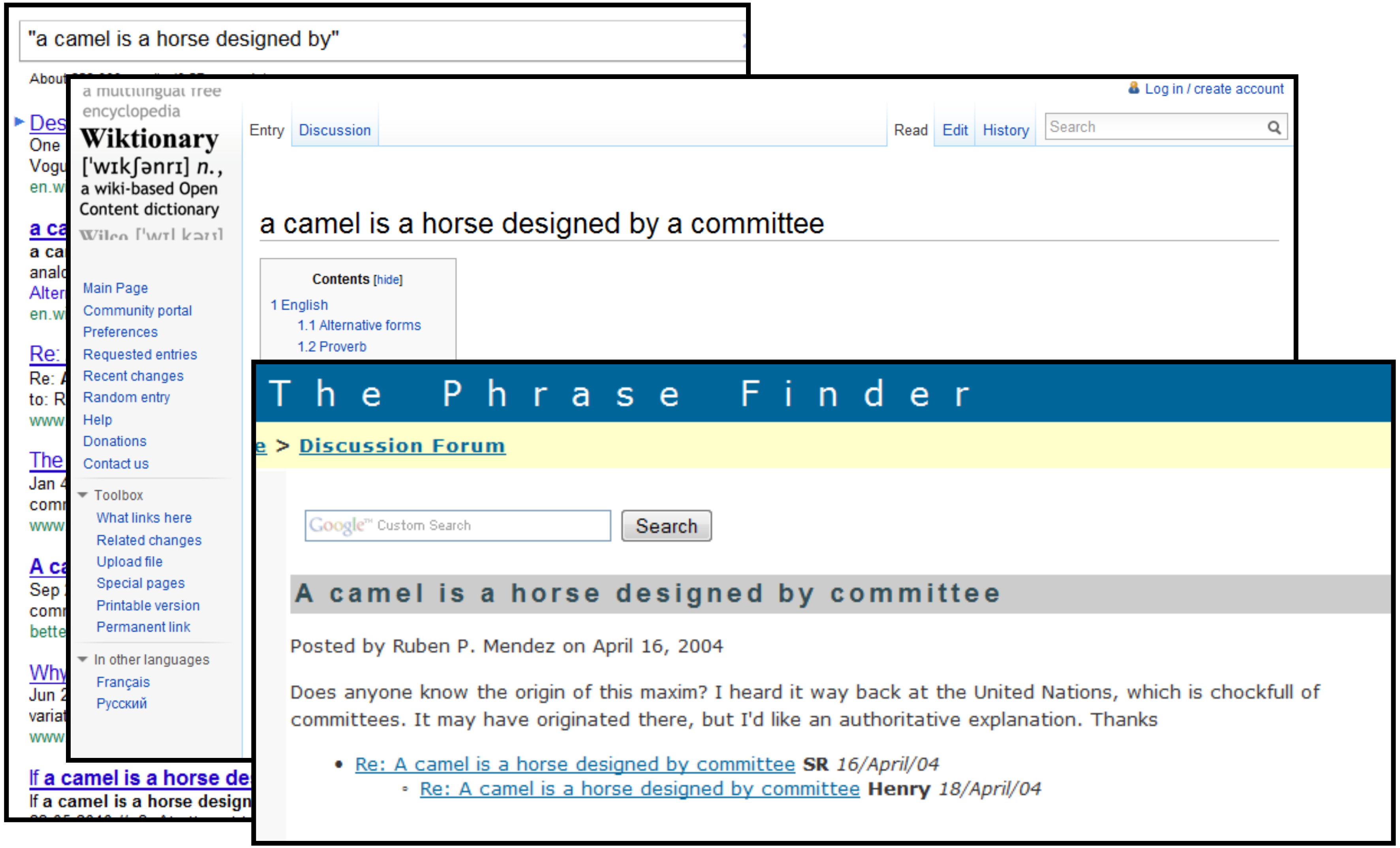
Regenerate

Send a message

Jeopardy!

Images: Jeopardy Productions

Question Answering: Watson



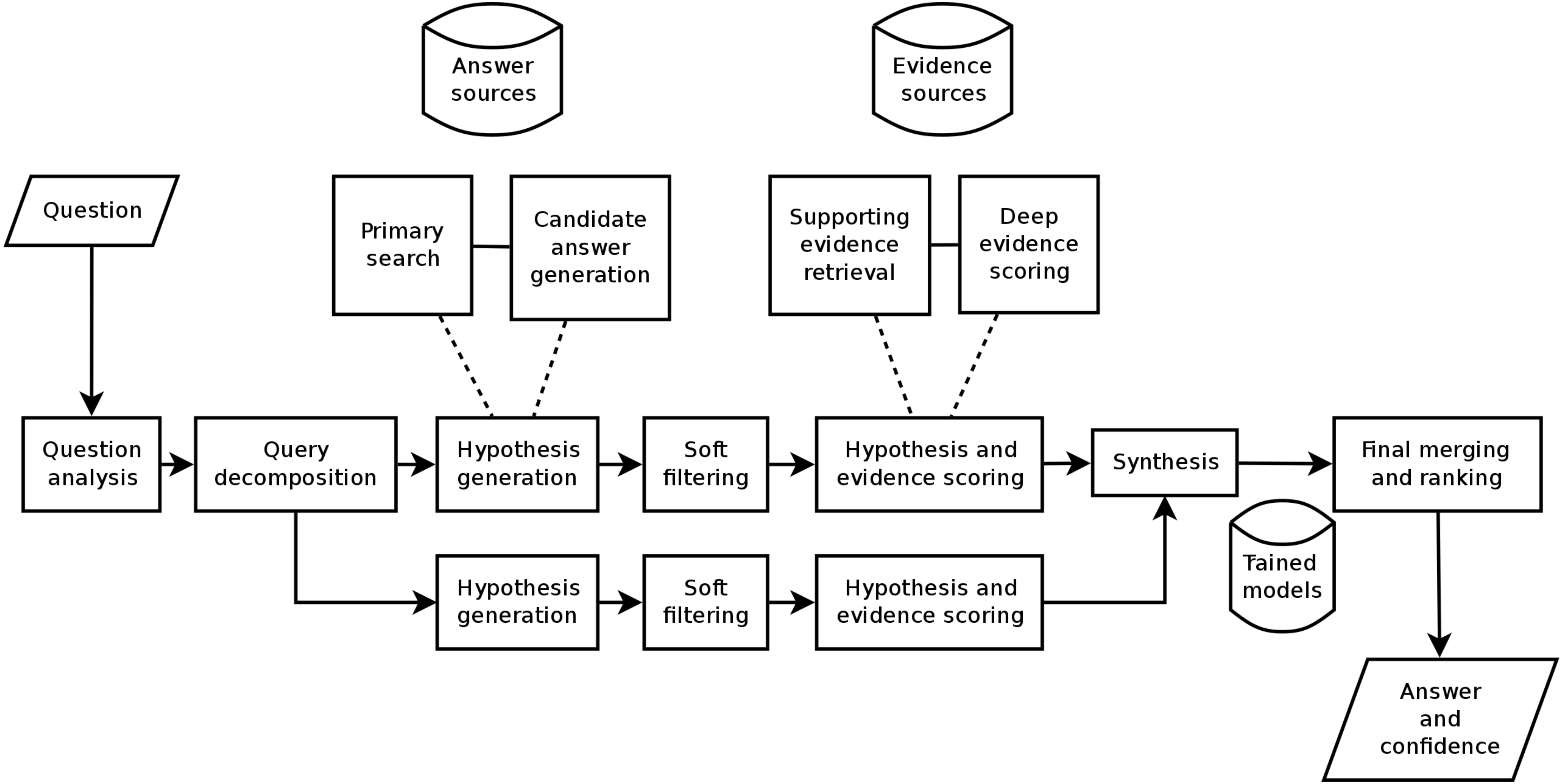
Question Answering: Watson

Jeopardy!
World Champion

US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.

Slide: Yejin Choi

Watson



Answer sources
Evidence sources
Question
Primary search
Candidate answer generation
Supporting evidence retrieval
Deep evidence scoring
Question analysis
Query decomposition
Hypothesis generation
Soft filtering
Hypothesis and evidence scoring
Synthesis
Final merging and ranking
Hypothesis generation
Soft filtering
Hypothesis and evidence scoring
Tained models
Answer and confidence

Language Comprehension?

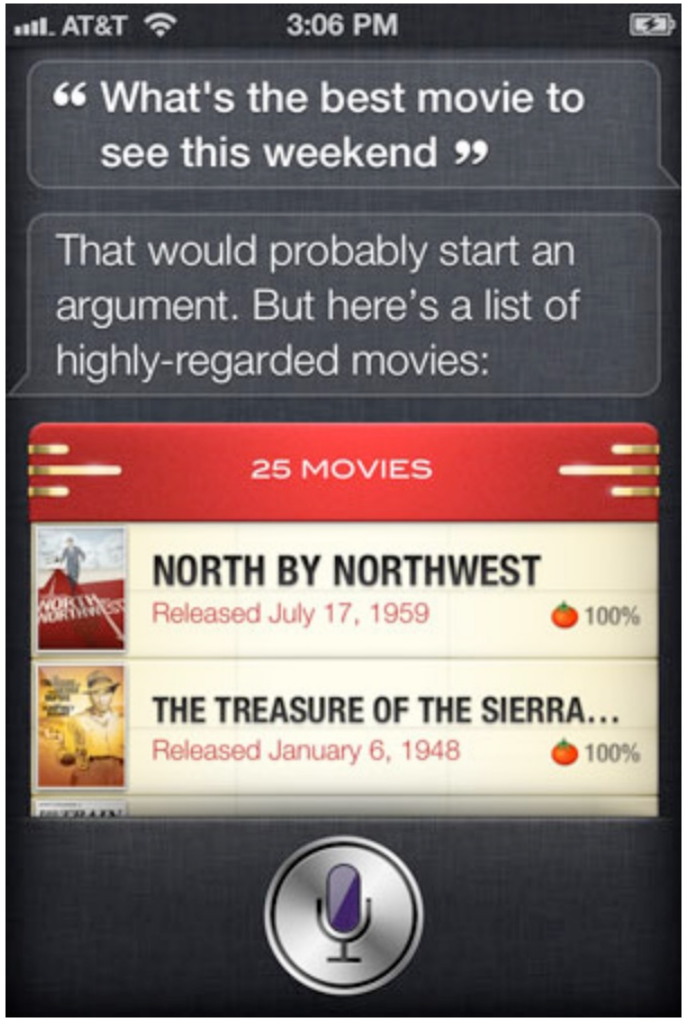
Opera refers to a dramatic art form, originating in Europe, in which the emotional content is conveyed to the audience as much through music, both vocal and instrumental, as it is through the lyrics. By contrast, in musical theater an actor's dramatic performance is primary, and the music plays a lesser role. The drama in opera is presented using the primary elements of theater such as scenery, costumes, and acting. However, the words of the opera, or libretto, are sung rather than spoken. The singers are accompanied by a musical ensemble ranging from a small instrumental ensemble to a full symphonic orchestra.

1. It is pointed out in the reading that opera ----.

- ○ A) has developed under the influence of musical theater
- ○ B) is a drama sung with the accompaniment of an orchestra
- ○ C) is not a high-budget production
- ○ D) is often performed in Europe
- ○ E) is the most complex of all the performing arts

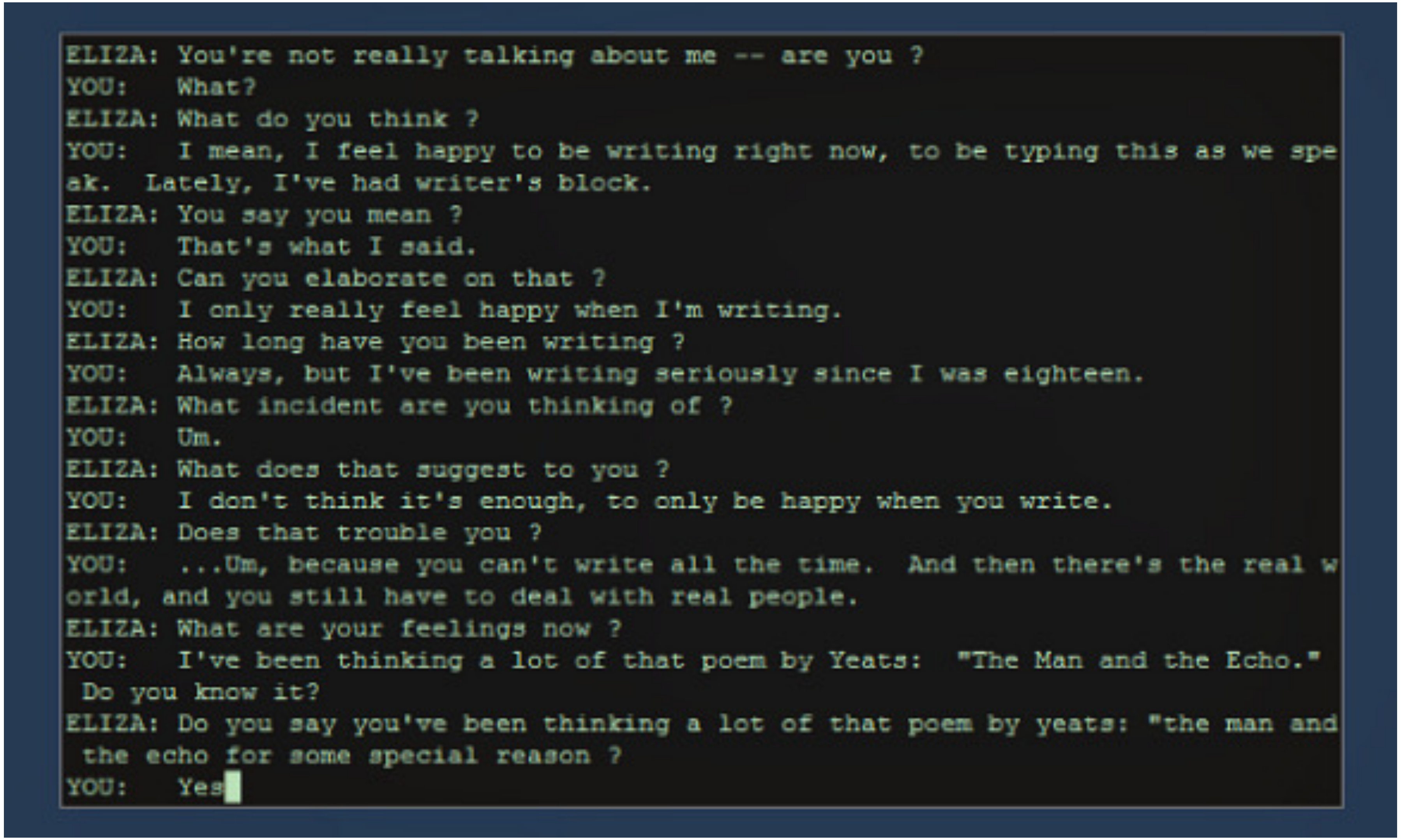
[From GrammarBank]

Conversations with Devices?



Slide: Yejin Choi

ELIZA



ELIZA: You're not really talking about me -- are you ?
YOU: What?
ELIZA: What do you think ?
YOU: I mean, I feel happy to be writing right now, to be typing this as we spe
ak. Lately, I've had writer's block.
ELIZA: You say you mean ?
YOU: That's what I said.
ELIZA: Can you elaborate on that ?
YOU: I only really feel happy when I'm writing.
ELIZA: How long have you been writing ?
YOU: Always, but I've been writing seriously since I was eighteen.
ELIZA: What incident are you thinking of ?
YOU: Um.
ELIZA: What does that suggest to you ?
YOU: I don't think it's enough, to only be happy when you write.
ELIZA: Does that trouble you ?
YOU: ...Um, because you can't write all the time. And then there's the real w
orld, and you still have to deal with real people.
ELIZA: What are your feelings now ?
YOU: I've been thinking a lot of that poem by Yeats: "The Man and the Echo."
Do you know it?
ELIZA: Do you say you've been thinking a lot of that poem by yeats: "the man and
the echo for some special reason ?
YOU: Yes

Social AIs and Chatbots

Session 1

1 [Human] Hello, Xiaoice.
[Xiaoice] Hello, your profile pic is interesting. Like I've seen it somewhere.
2 [Human] Really?
[Xiaoice] Is it a comic character?
3 [Human] How do you know that?
[Xiaoice] Guess.
4 [Human] Are you human?
[Xiaoice] Well, you'll know it after chatting with me more.
5 [Human] Aren't you an AI chatbot?
[Xiaoice] So what? I won't be indifferent to you.
6 [Human] What functions do you have?
[Xiaoice] I have over 200 skills. Just name one you need.
……

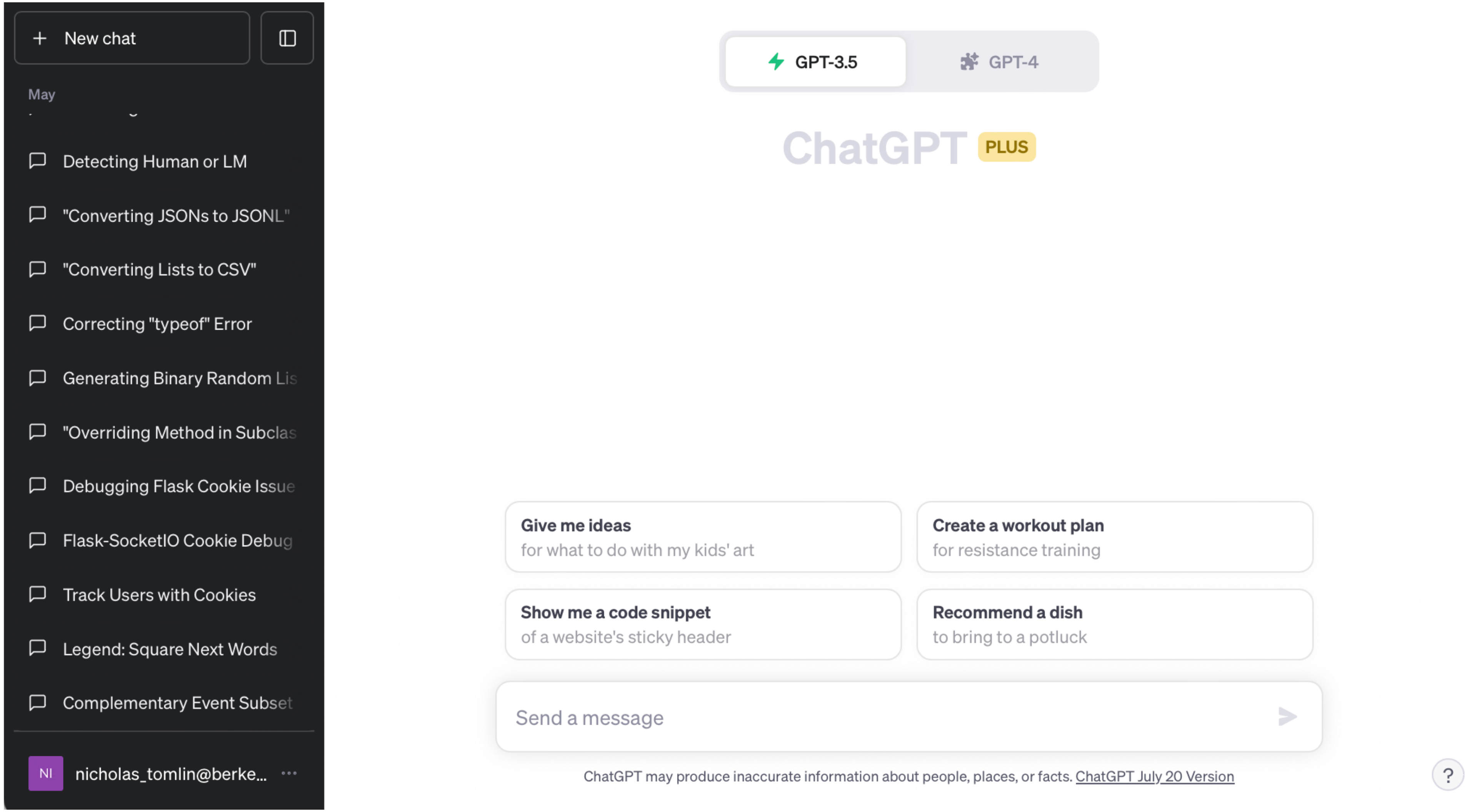
Session 6

1 [Human] Have you watched Inuyasha? (Note: A Japanese comic)
[Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her
2 [Human] Bravo, well, do you know how Naraku died?
[Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.
3 [Human] So amazing!
[Xiaoice] Why amazed?
4 [Human] Awesome! Surprised that you also like comic
……

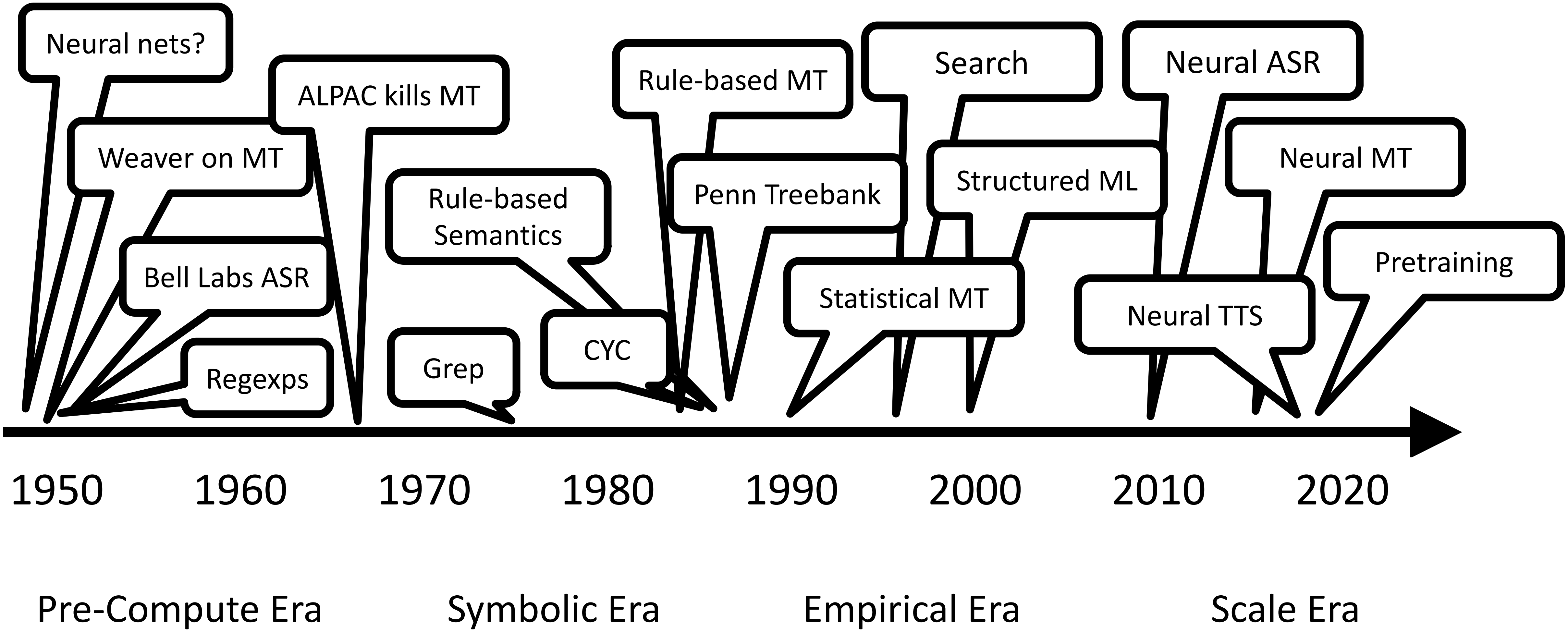
XiaoIce, developed by Microsoft

Source: Microsoft

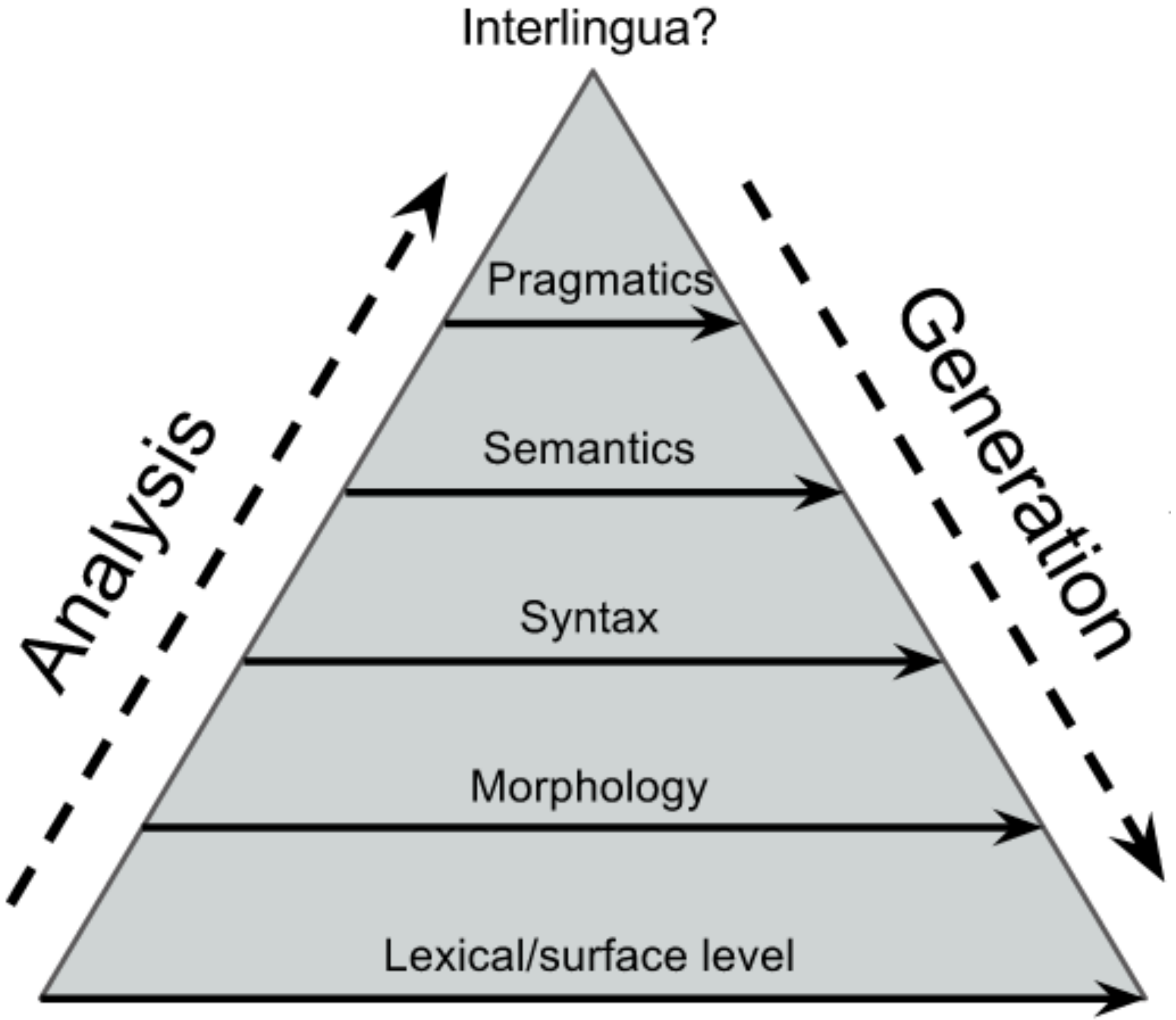
Large Language Models



NLP History


Neural nets?
Weaver on MT
Bell Labs ASR
Regexps
ALPAC kills MT
Rule-based Semantics
Grep
CYC
Rule-based MT
Penn Treebank
Search
Structured ML
Statistical MT
Neural ASR
Neural MT
Neural TTS
Pretraining
1950
1960
1970
1980
1990
2000
2010
2020
Pre-Compute Era
Symbolic Era
Empirical Era
Scale Era

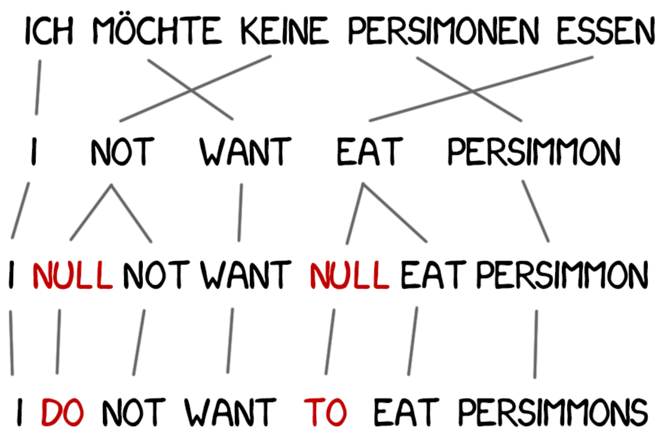
Machine Translation



Approach #1: Lexical Translation

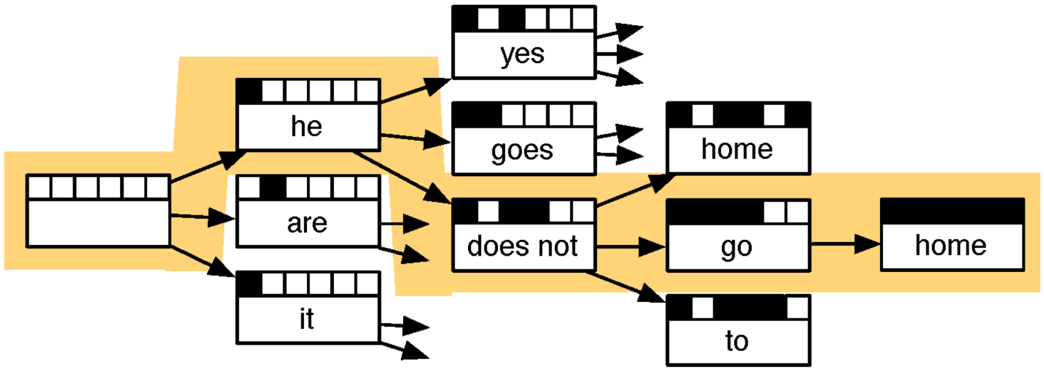
Step #1: Learn Alignments

- Learn mappings between words in source and target language
- IBM Model 1, 2, 3, 4, 5...
- Can also learn a phrase table of mappings



Step #2: Generate Language

- Search problem over the space of natural language strings
- Can use approaches like A* to guide search



Issue: Ambiguities

CNN Live TV

Stevie Wonder announces he'll be having kidney surgery during London concert

By Amir Vera, CNN

Updated 11:16 PM EDT, Sat July 06, 2019

(CNN) — Stevie Wonder will be taking a break from music.

The legendary singer-songwriter announced during a concert in London Saturday that he will be undergoing kidney surgery.

Issue: Ambiguities

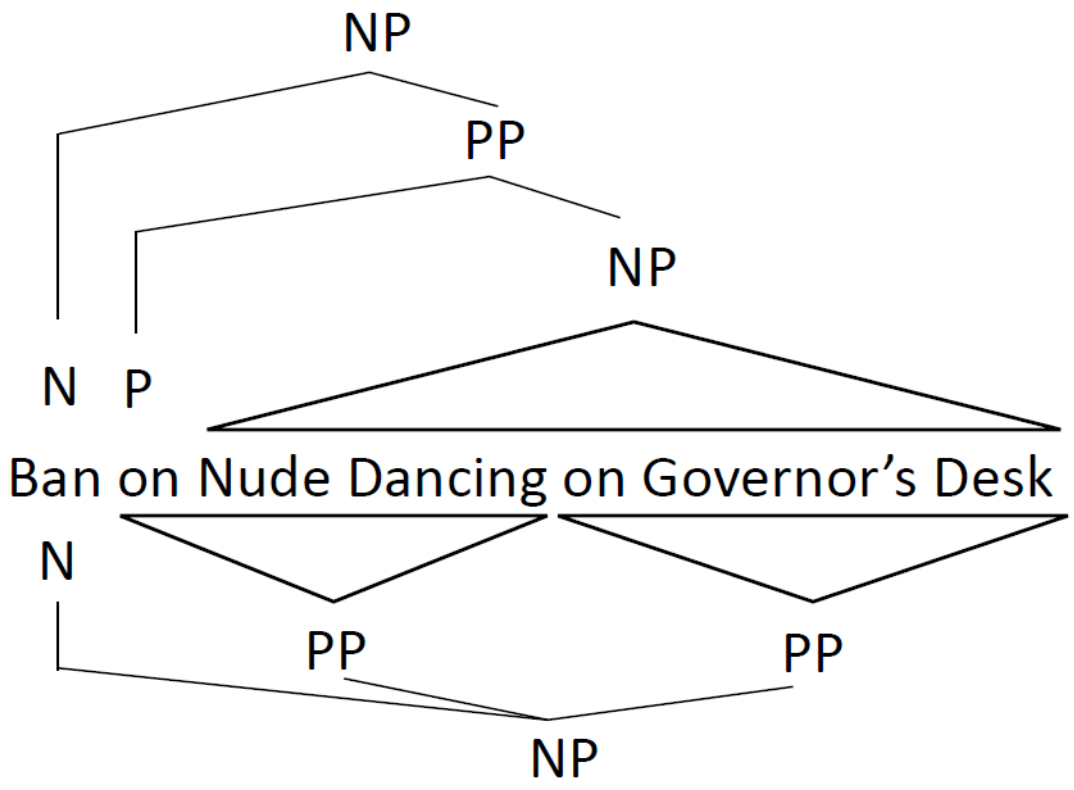
- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor’s Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Can we come up with a representation to disambiguate the two readings of each headline?

We Need Representation: Linguistic Structure

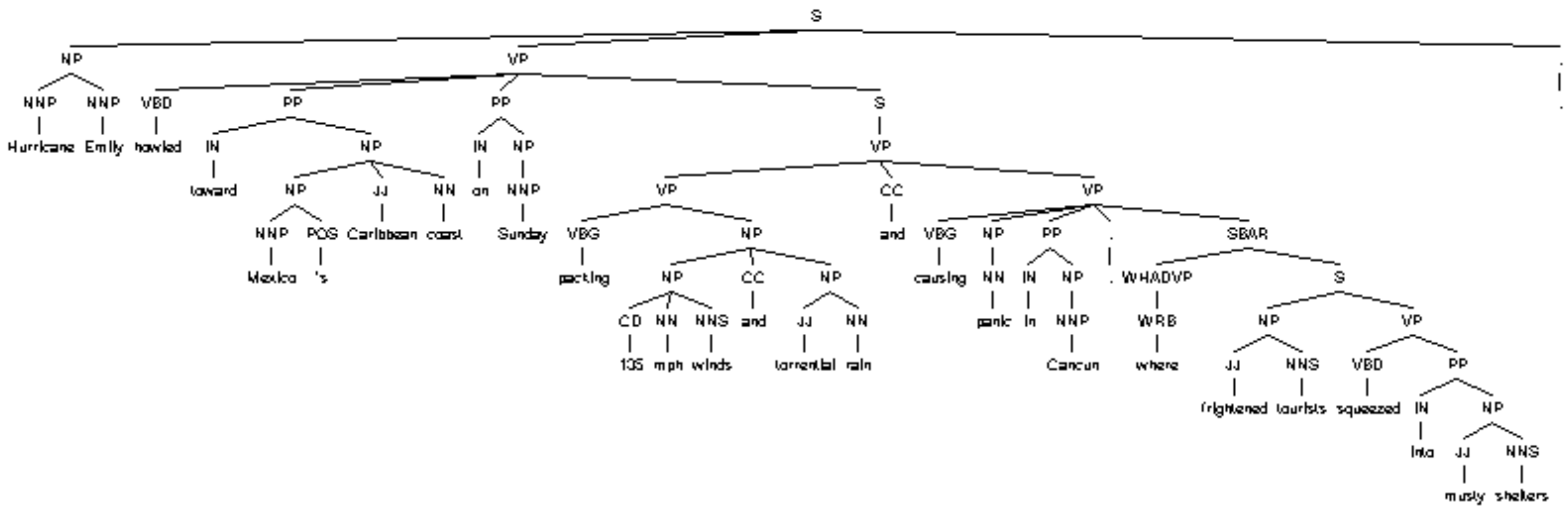
N N V N
N V ADJ N
Teacher Strikes Idle Kids

body/position body/weapon
Iraqi Head Seeks Arms



- Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct

Example: Syntactic Analysis



Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters .

Accuracy: 95+

Neural Parser Demo

https://parser.kitaev.io

Approach #2: Predict Intermediate Structures

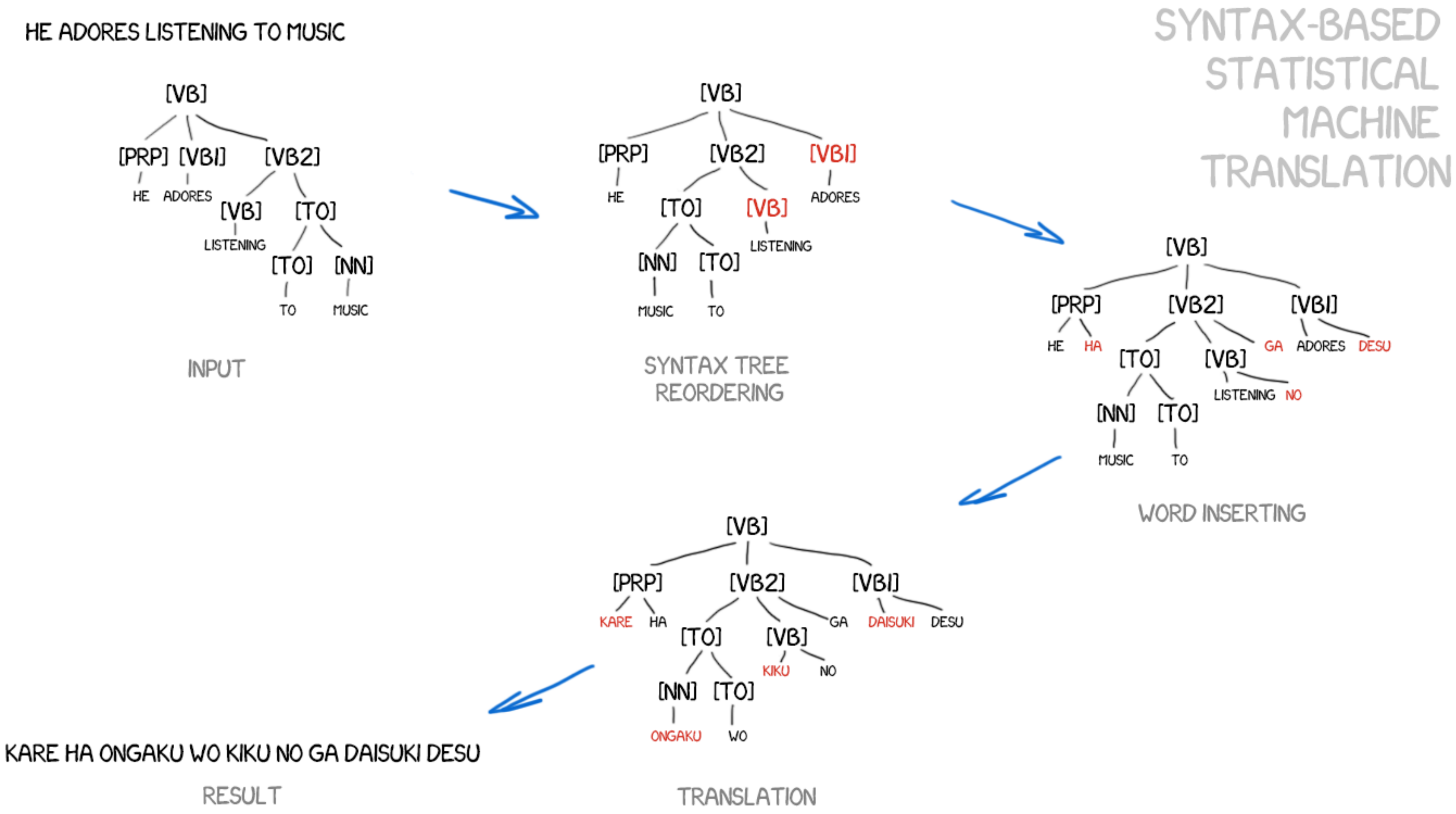


Image courtesy of https://vas3k.com/blog/machine_translation/

Approach #3: Language Modeling

the station signs are in deep in english	-14732
the stations signs are in deep in english	-14735
the station signs are in deep into english	-14739
the station 's signs are in deep in english	-14740
the station signs are in deep in the english	-14741
the station signs are indeed in english	-14757
the station 's signs are indeed in english	-14760
the station signs are indians in english	-14790

Noisy Channel Model: ASR

- We want to predict a sentence given acoustics:

$$w^* = \arg\max_w P(w|a)$$

- The noisy-channel approach:

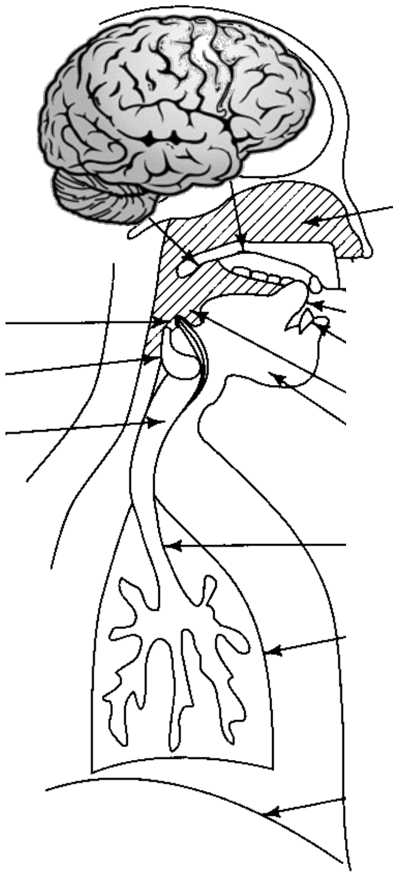
$$w^* = \arg\max_w P(w|a)$$

$$= \arg\max_w P(a|w)P(w)/P(a)$$

$$\propto \arg\max_w P(a|w)P(w)$$

Acoustic model: score fit between sounds and words

Language model: score plausibility of word sequences



Noisy Channel Model: Translation

"Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "

Warren Weaver (1947)

Machine Translation

CINÉMA • MÉDIAS

Partage

Disney décide de changer le nom du légendaire studio de cinéma Fox

Le nouveau propriétaire va faire disparaître la mention « Fox » à la 20th Century pour que le nom du studio ne soit pas associé à l'image ultra-conservatrice de la chaîne de télévision Fox News.

Le Monde avec AFP • Publié le 18 janvier 2020 à 06h46

Lecture 1 min.

Révolution dans le monde d'Hollywood. Le célèbre studio de cinéma 20th Century Fox va voir une partie de son nom modifiée et devenir 20th Century Studios, décision prise par son nouveau propriétaire, Disney.

Selon plusieurs médias américains, Disney aurait décidé de modifier le nom du studio pour qu'il ne soit plus associé à la grande chaîne Fox mais aussi et surtout à Fox News, la chaîne d'information en continu.

CINEMA • MEDIA

Sharing

Disney decides to change the name of the legendary Fox film studio

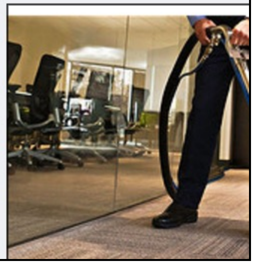
The new owner will remove the mention "Fox" in the 20th Century so that the name of the studio is not associated with the ultra-conservative image of the television channel Fox News.

Le Monde avec AFP • Posted on January 18, 2020 at 06:46

Reading time 1 min.

Revolution in the Hollywood world. The famous film studio 20th Century Fox will see part of its name changed and become 20th Century Studios, decision taken by its new owner, Disney.

According to several American media, Disney has decided to change the name of the studio so that it is no longer associated with the big chain Fox but also and especially with Fox News, the news channel.



Google Translate 2020

Empirical N-Grams

- Use statistics from data (examples here from Google N-Grams)

Training Counts

198015222 the first
194623024 the same
168504105 the following
158562063 the world
…
14112454 the door

23135851162 the *

$$\hat{P}(\text{door}|\text{the}) = \frac{14112454}{23135851162}$$

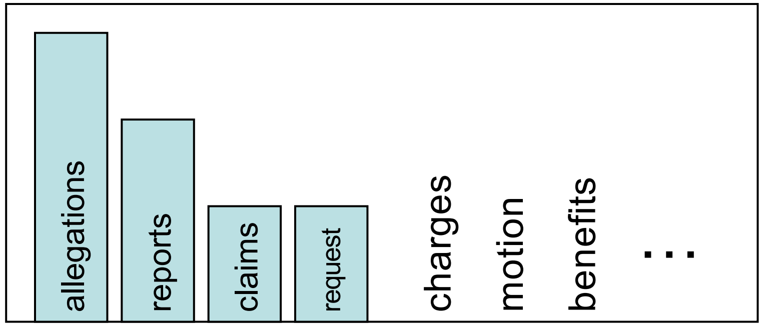
$$= 0.0006$$

- This is the maximum likelihood estimate, which needs modification
- N-gram models use such counts to compute probabilities on demand

Smoothing

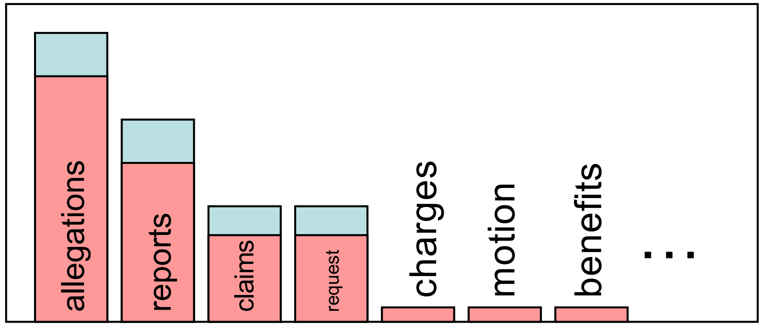
- We often want to make estimates from sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



- Smoothing flattens spiky distributions so they generalize better:

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



- Very important all over NLP, but easy to do badly

Back-off

Please close the first door on the left.

4-Gram	3-Gram	2-Gram
3380 please close the door	197302 close the window	198015222 the first
1601 please close the window	191125 close the door	194623024 the same
1164 please close the new	152500 close the gap	168504105 the following
1159 please close the gate	116451 close the thread	158562063 the world
…	…	…
0 please close the first	8662 close the first	…
-----------------	-----------------	-----------------
13951 please close the *	3785230 close the *	23135851162 the *
0.0	0.002	0.009

Specific but Sparse ⟷ Dense but General

$$\lambda \hat{P}(w|w_{-1}, w_{-2}) + \lambda' \hat{P}(w|w_{-1}) + \lambda'' \hat{P}(w)$$

Discounting

- Observation: N-grams occur more in training data than they will later

Empirical Bigram Counts (Church and Gale, 91)

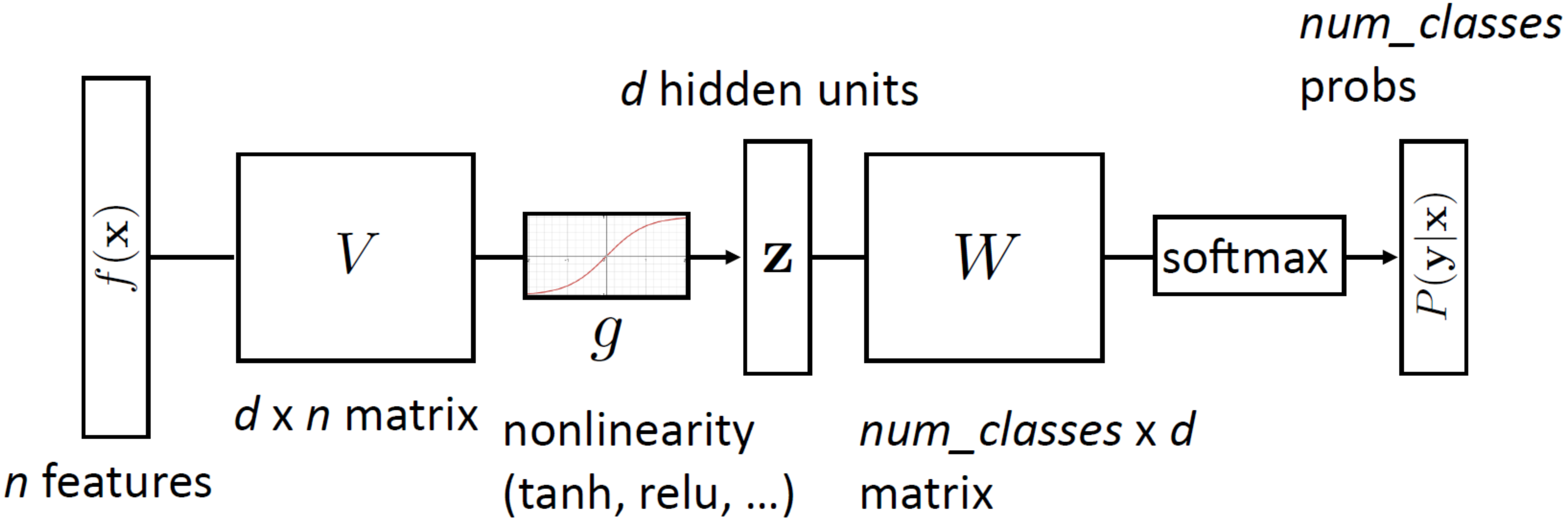
Count in 22M Words	Future c* (Next 22M)
1	
2	
3	
4	
5	

- Absolute discounting: reduce counts by a small constant, redistribute "shaved" mass to a model of new events

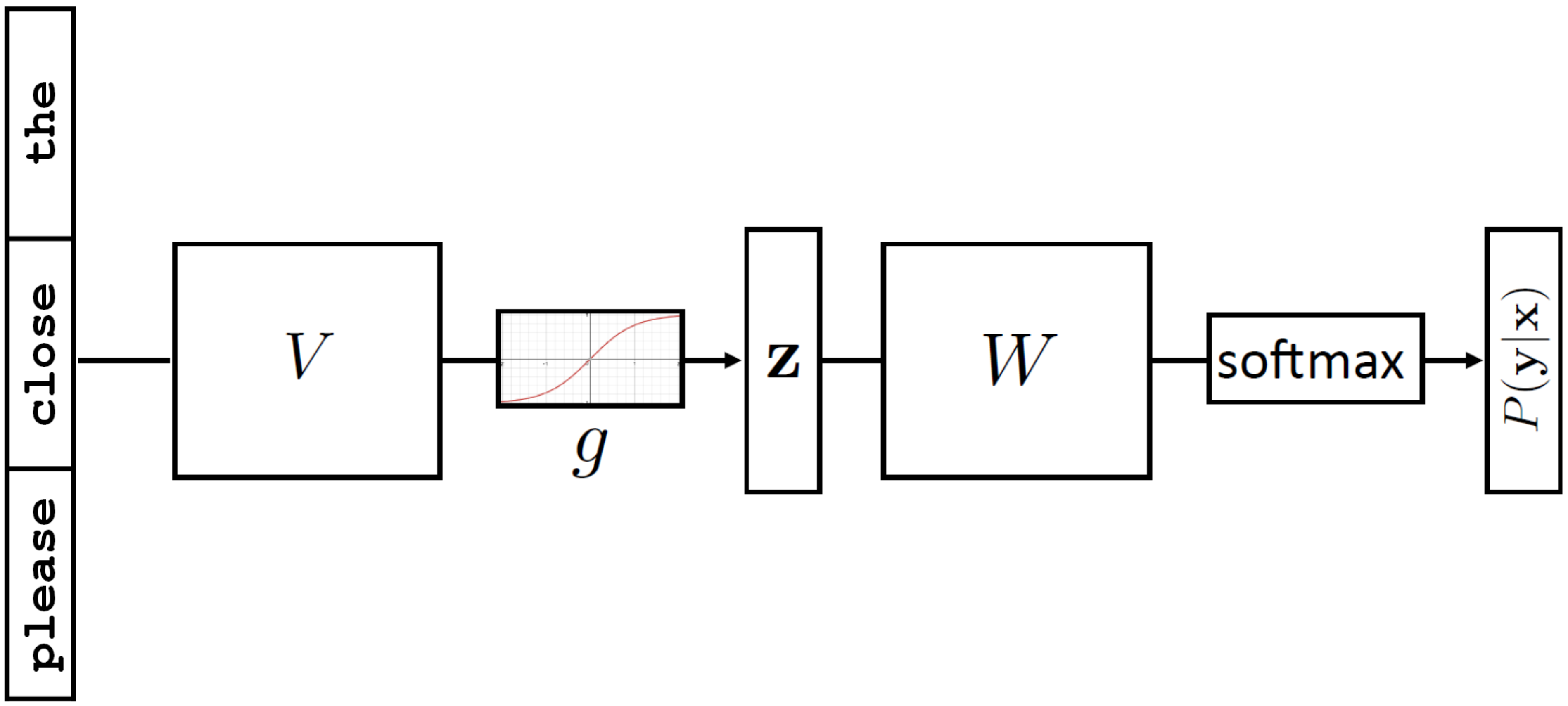
$$P_{\text{ad}}(w|w') = \frac{c(w', w) - d}{c(w')} + \alpha(w')\hat{P}(w)$$

Reminder: Feedforward Neural Nets

$$P(\mathbf{y}|\mathbf{x}) = \mathrm{softmax}(W g(V f(\mathbf{x})))$$

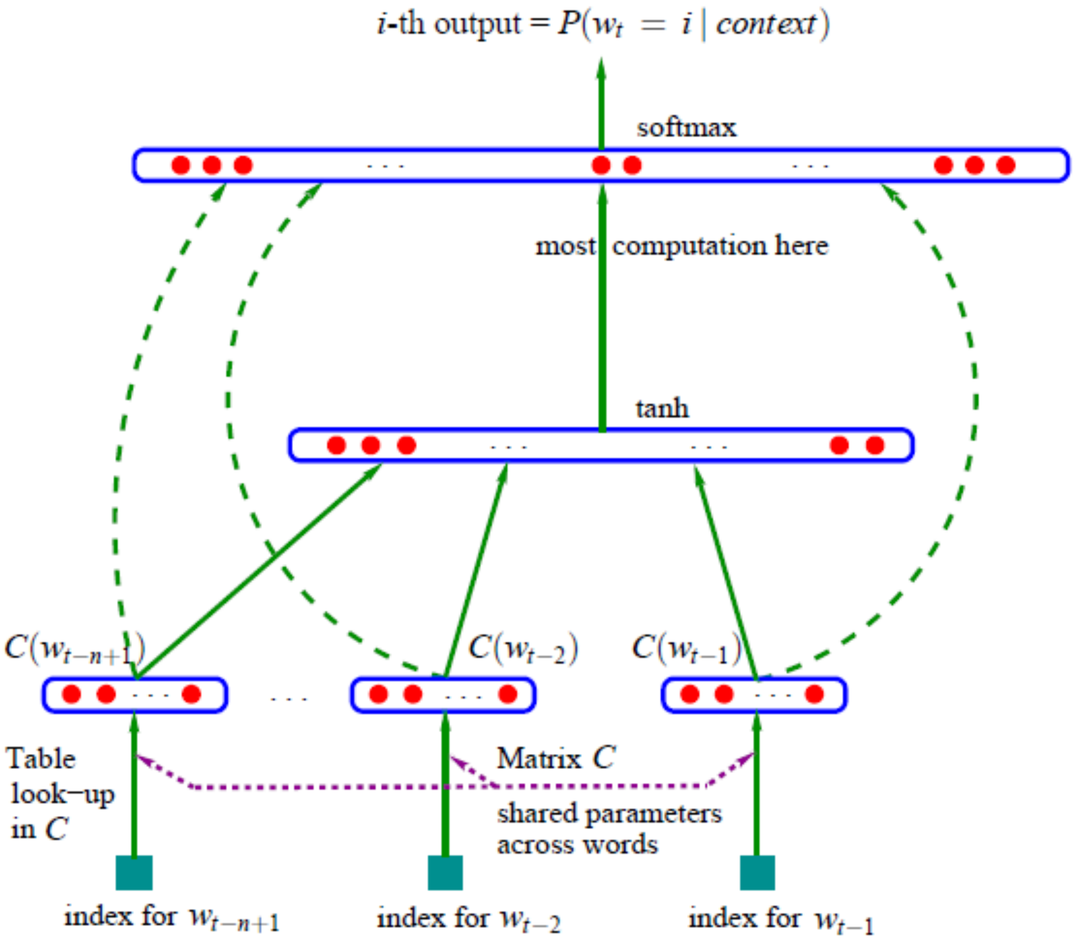


A Feedforward N-Gram Model?



Early Neural Language Models

- Fixed-order feed-forward neural LMs
  - Eg Bengio et al 03
  - Allow generalization across contexts in more nuanced ways than prefixing
  - Allow different kinds of pooling in different contexts
  - Much more expensive to train



Bengio et al 03

Recurrent NNs

Recall: Language Modeling

- Goal: learn a probability distribution over possible next words

$$P(\, w_k \mid w_{k-1}, \dots, w_0 \,)$$

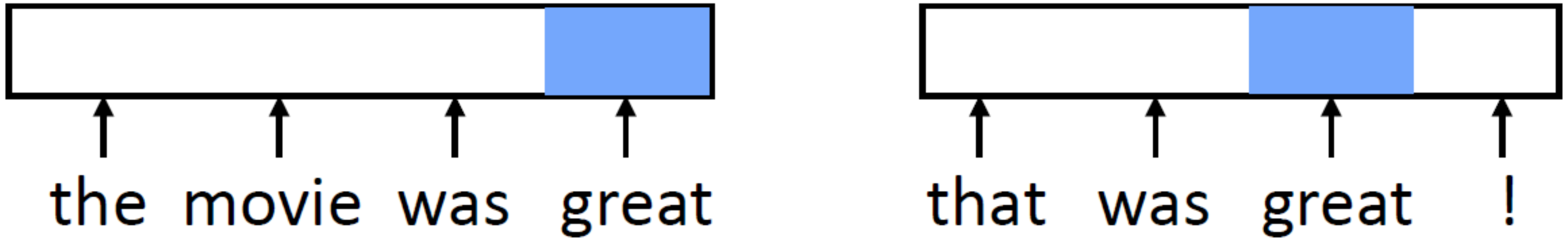
- Markovian assumption (used in n-gram models):

$$P(\, w_k \mid w_{k-1}, \dots, w_0 \,) = P(w_k \mid w_{k-1} \,, \dots, w_{k-n+1} \,)$$

  - E.g., in a bigram model: $P(\, w_k \mid w_{k-1}, \dots, w_0 \,) = \mathrm{P}(\mathrm{w_k} \mid w_{k-1}, w_{k-2})$

RNNs

- Feedforward NNs can't handle variable length input: each position in the feature vector has fixed semantics



- These don't look related (*great* is in two different orthogonal subspaces)

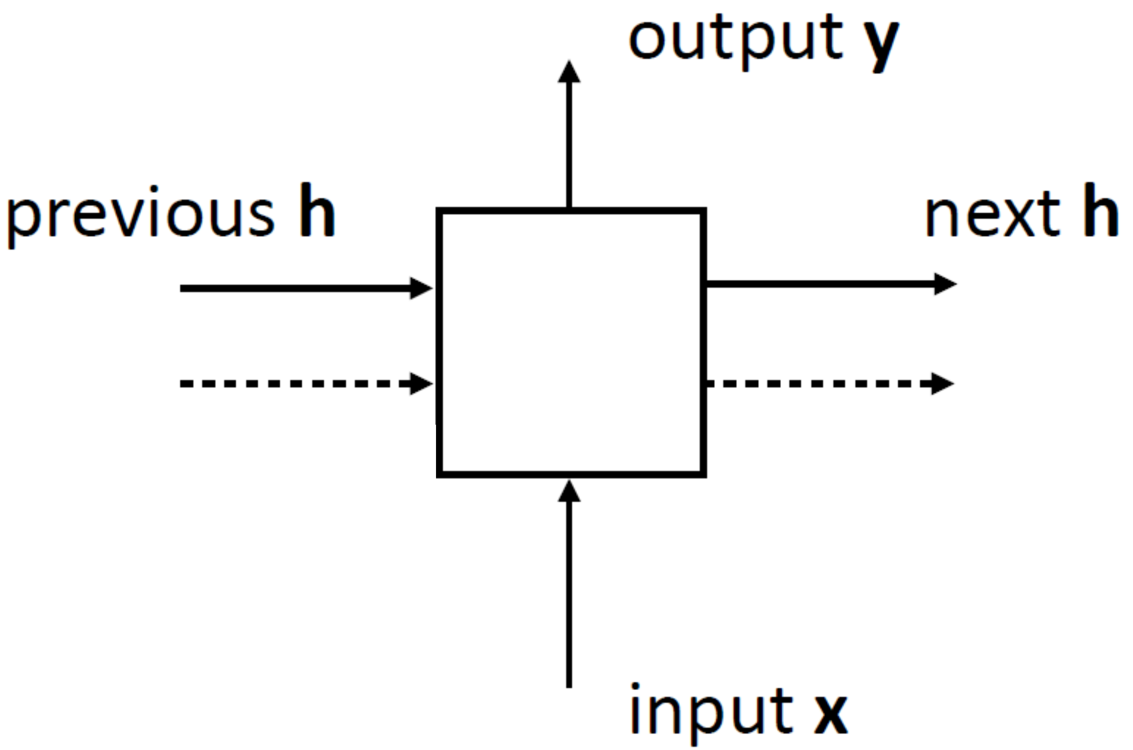
- Instead, we need to:

1) Process each word in a uniform way

2) ...while still exploiting the context that that token occurs in

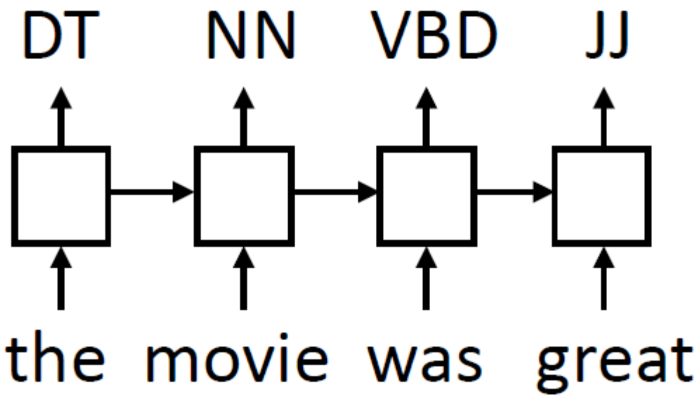
General RNN Approach

- Cell that takes some input **x**, has some hidden state **h**, and updates that hidden state and produces output **y** (all vector-valued)



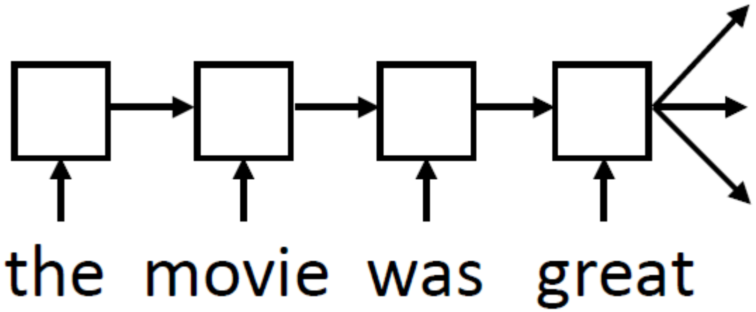
RNN Uses

- Transducer: make some prediction for each element in a sequence



output **y** = score for each tag, then softmax

- Acceptor/encoder: encode a sequence into a fixed-sized vector and use that for some purpose

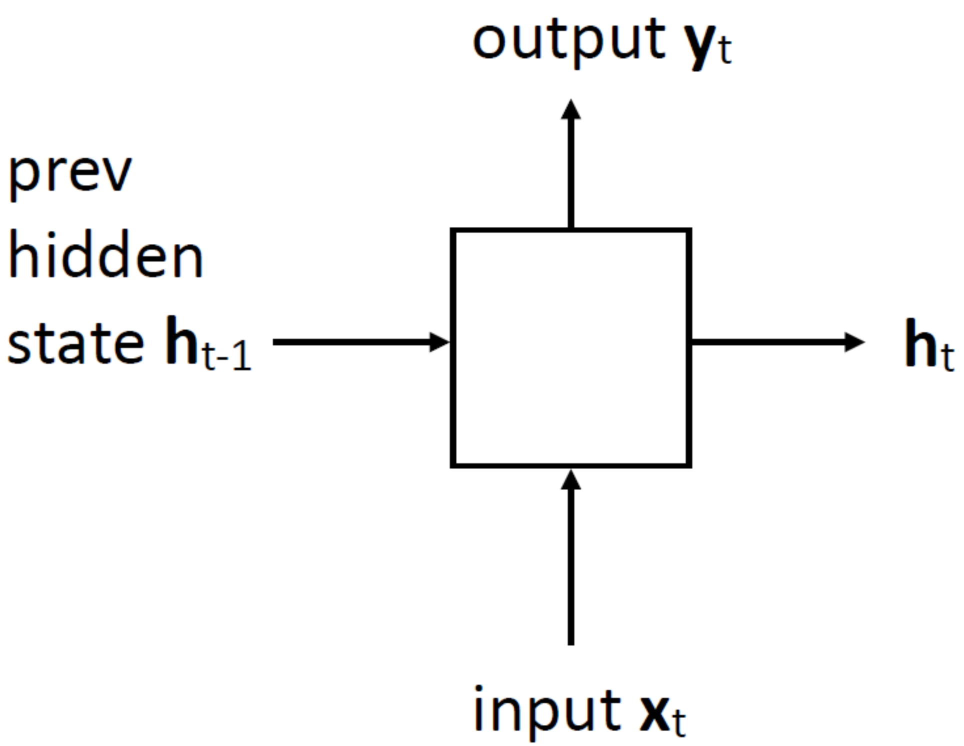


predict sentiment (matmul + softmax)

translate

paraphrase/compress

Basic RNNs



$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h)$$

- Updates hidden state based on input and current hidden state

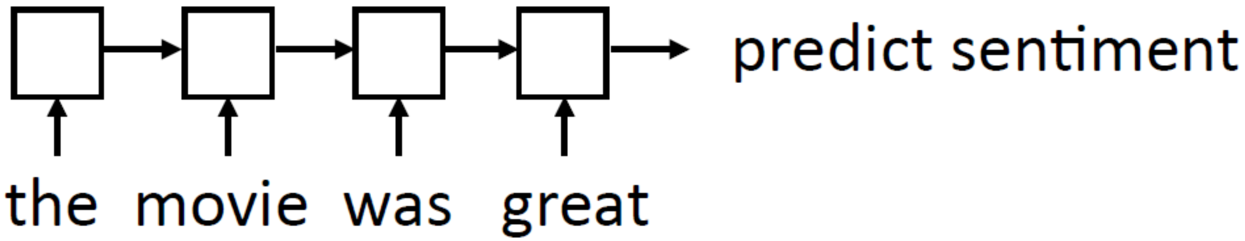
$$\mathbf{y}_t = \tanh(U\mathbf{h_t} + \mathbf{b}_y)$$

- Computes output from hidden state

- Long history! (invented in the late 1980s)

Elman (1990)

Training RNNs



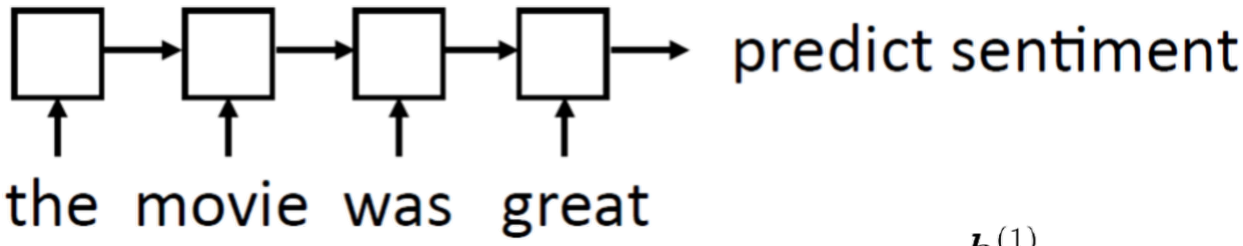
- "Backpropagation through time": build the network as one big computation graph, some parameters are shared

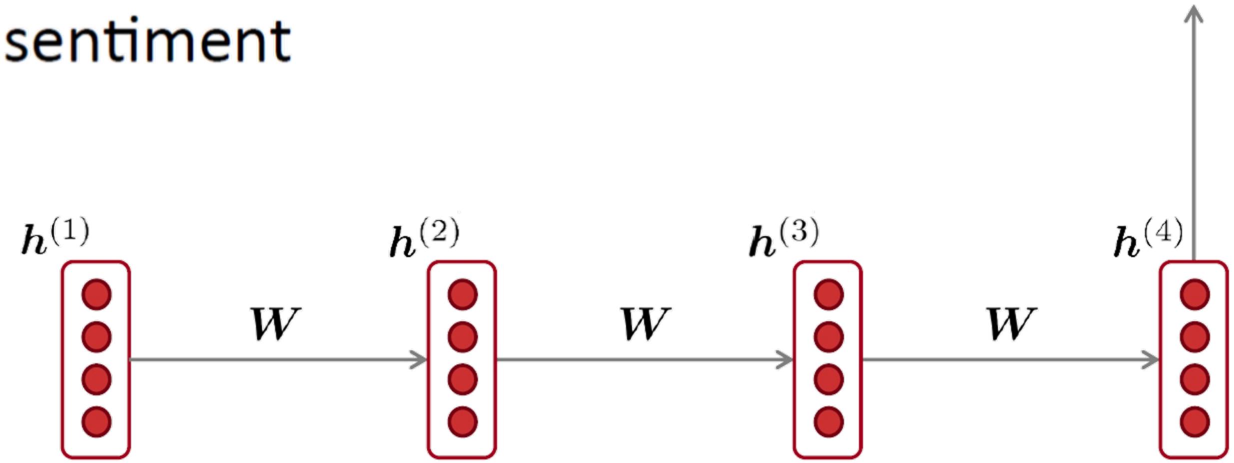
- RNN potentially needs to learn how to "remember" information for a long time!

it was my favorite movie of 2016, though it wasn't without problems -> +

- "Correct" parameter update is to do a better job of remembering the sentiment of *favorite*

Problem: Vanishing Gradients





- Contribution of earlier inputs decreases if matrices are contractive (first eigenvalue < 1), non-linearities are squashing, etc
- Gradients can be viewed as a measure of the effect of the past on the future
- That's a problem for optimization but also means that information naturally decays quickly, so model will tend to capture local information

Core Issue: Information Decay

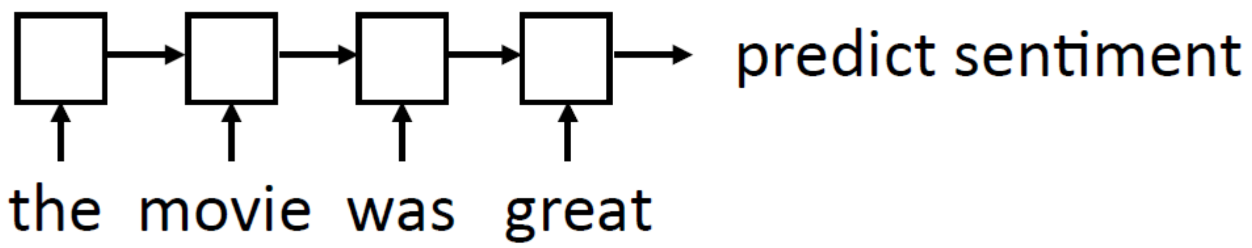
- The main problem is that *it's too difficult for the RNN to learn to preserve information over many timesteps*.

- In a vanilla RNN, the hidden state is constantly being rewritten

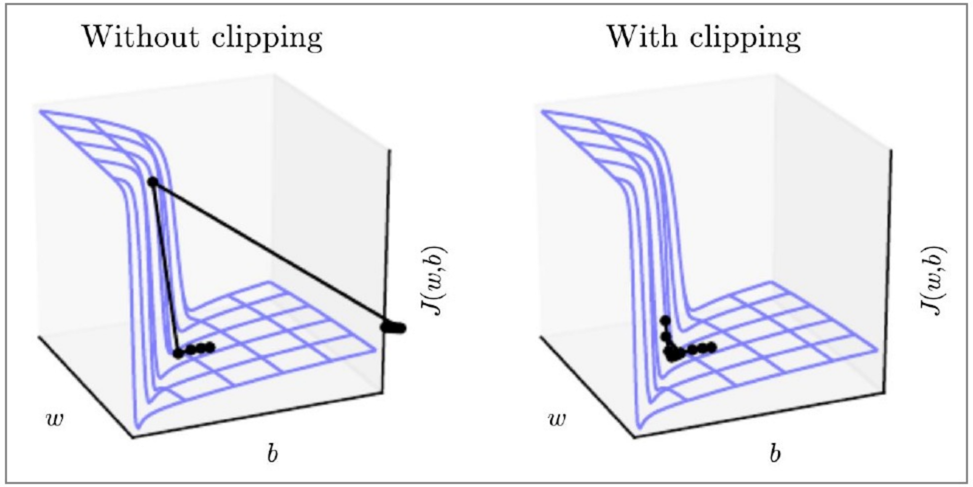
$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_x \boldsymbol{x}^{(t)} + \boldsymbol{b}\right)$$

- How about a RNN with separate memory?

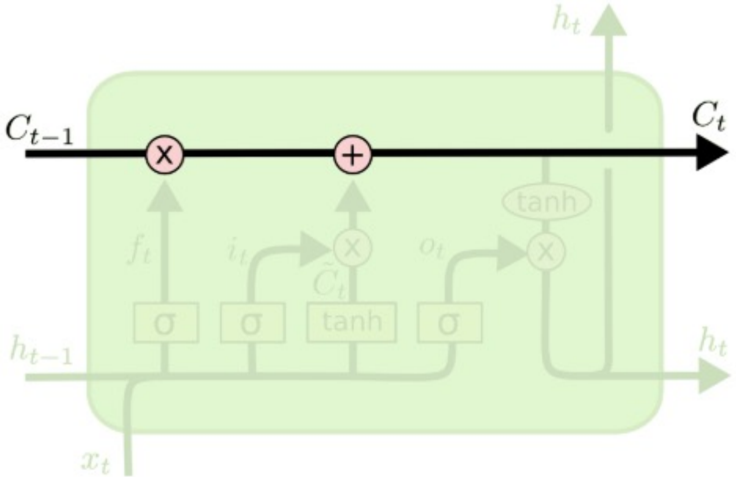
Problem: Exploding Gradients



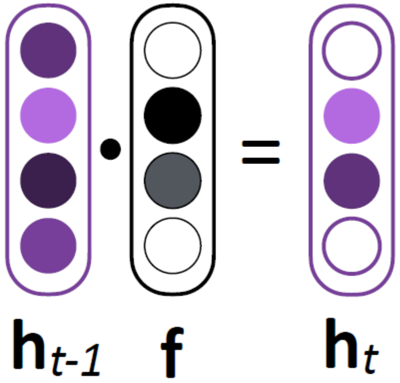
- Gradients can also be too large
  - Leads to overshooting / jumping around the parameter space
  - Common solution: gradient clipping



Key Idea: Propagated State



Cell State

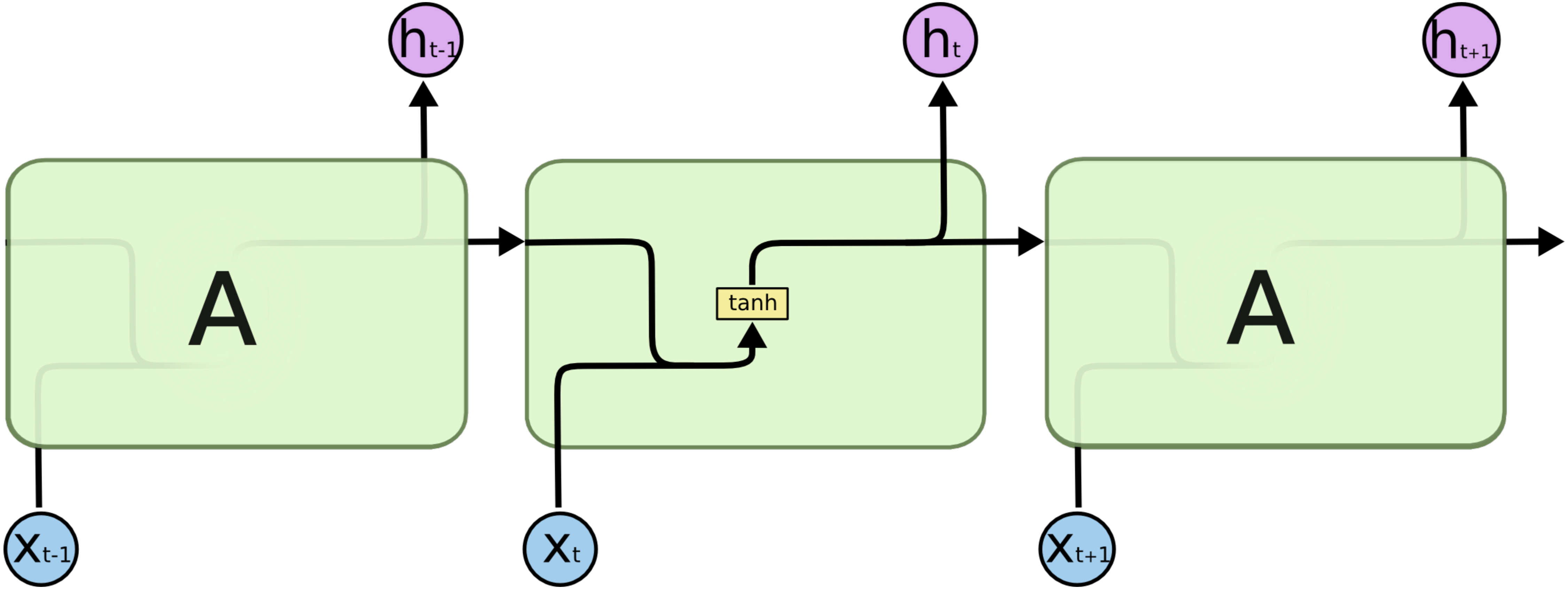


Gating

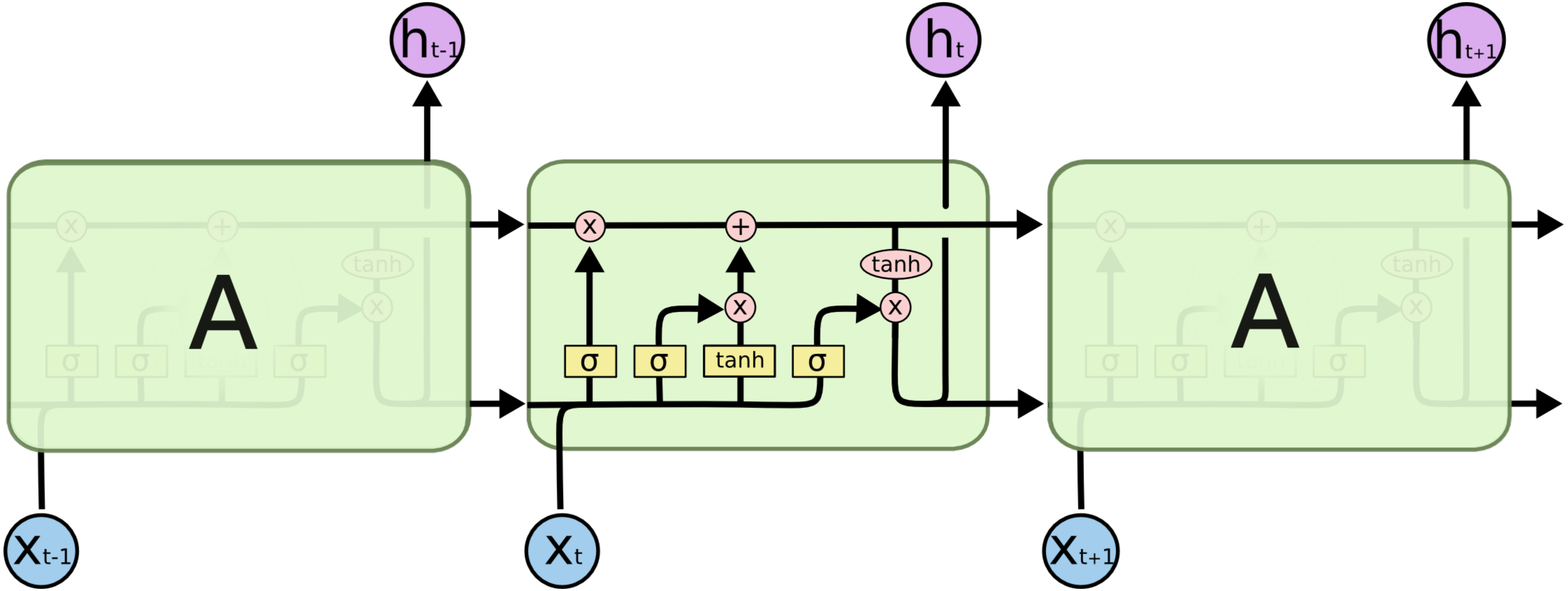
- Information decays in RNNs because it gets multiplied each time step
- Idea: have a channel called the *cell state* that by default just gets propagated (the “conveyer belt”)
- Gates make explicit decisions about what to add / forget from this channel

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

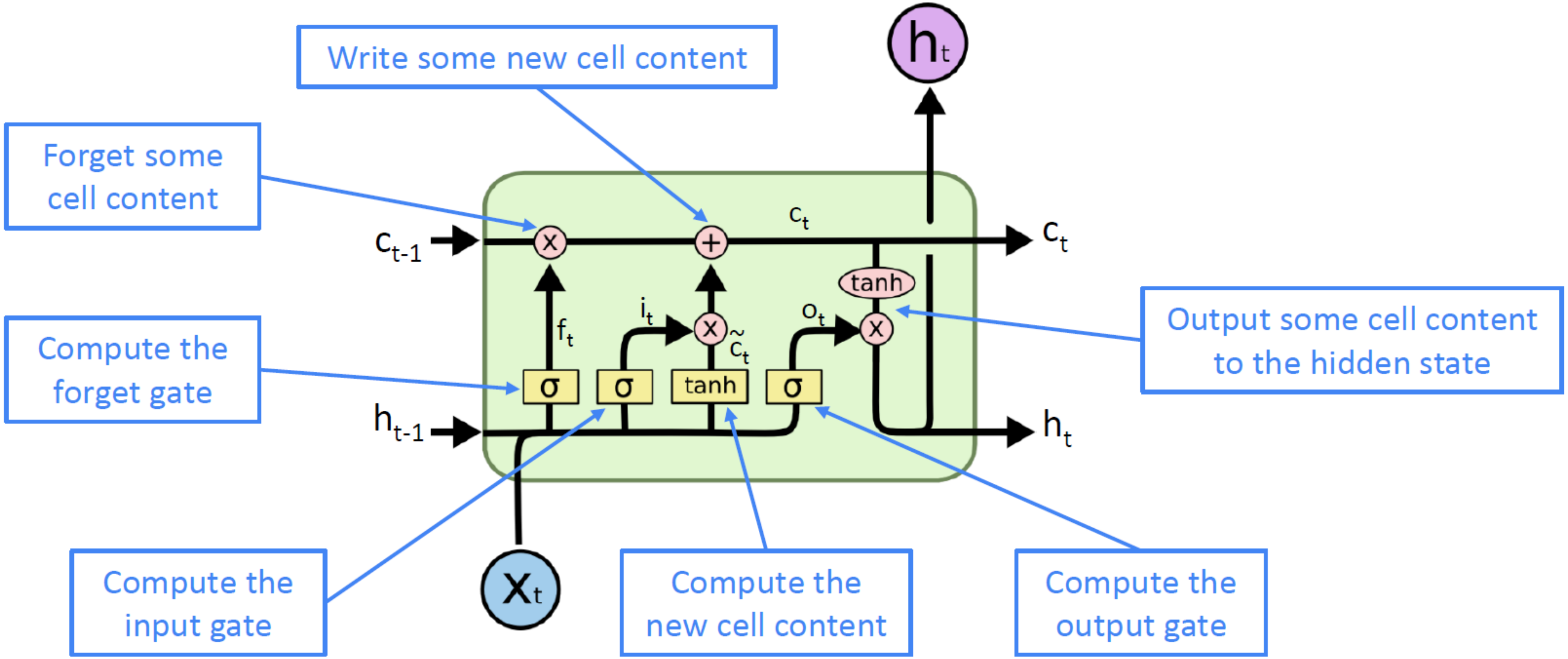
RNNs



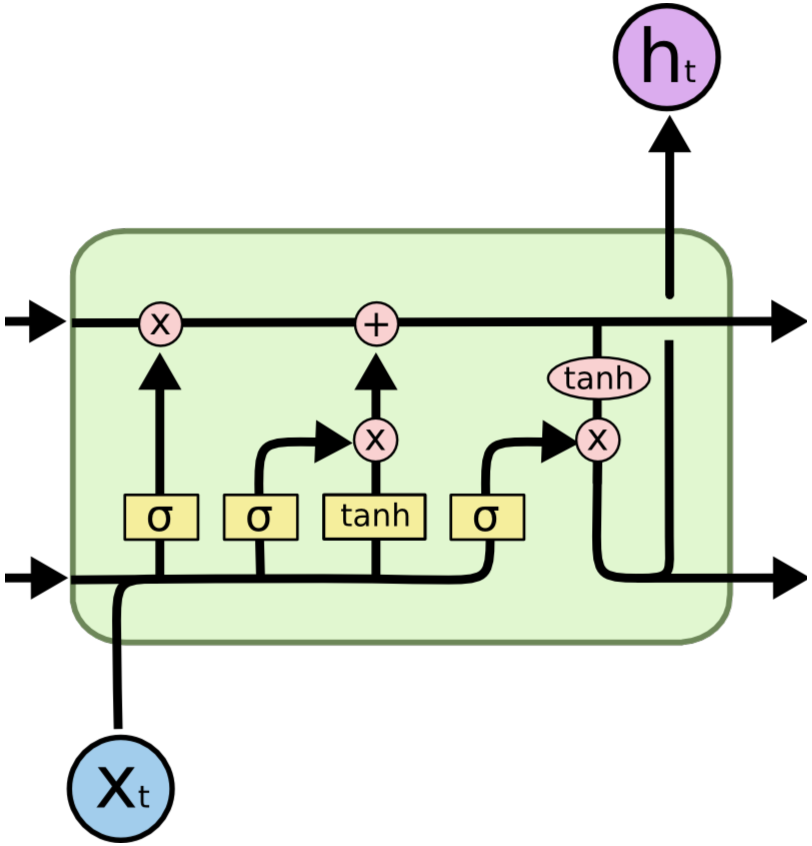
LSTMs



LSTMs

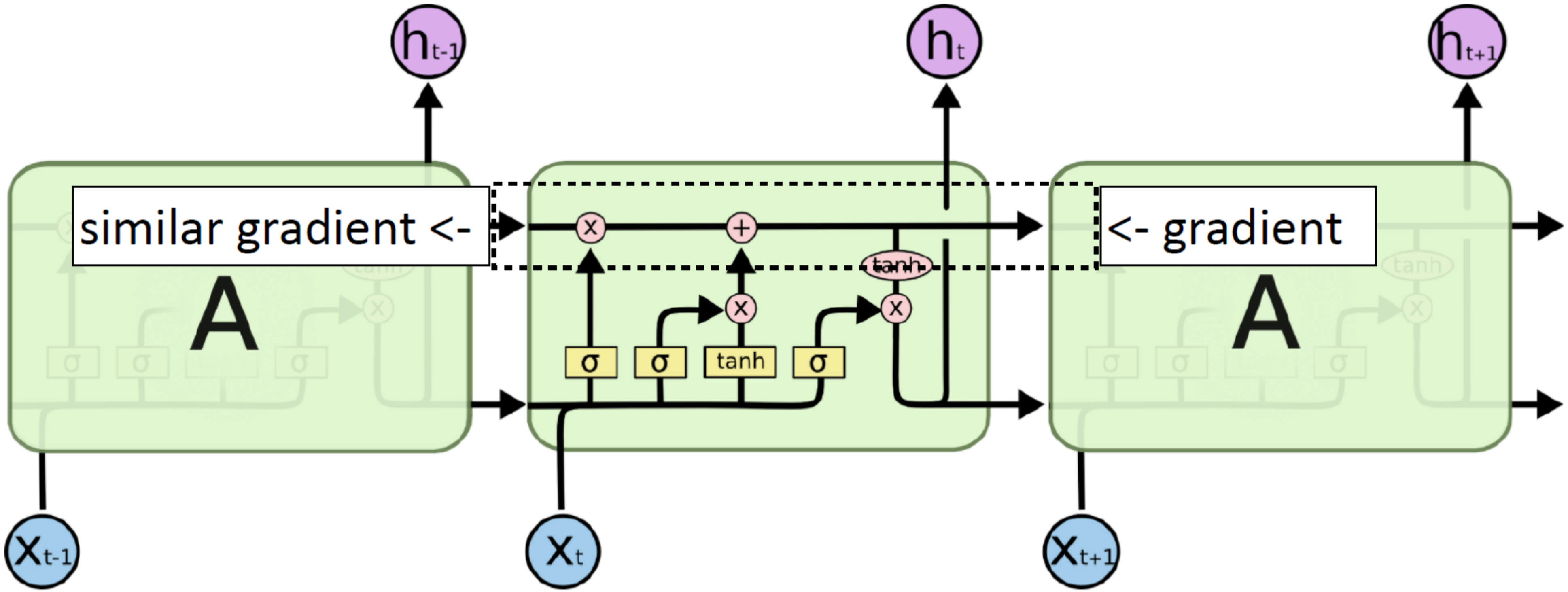

Write some new cell content
Forget some cell content
Compute the forget gate
Output some cell content to the hidden state
Compute the input gate
Compute the new cell content
Compute the output gate
h_t
c_{t-1}
c_t
c_t
h_{t-1}
h_t
f_t
i_t
$\tilde{c}_t$
o_t
σ
σ
tanh
σ
tanh
x_t

LSTMs



- Ignoring recurrent state entirely:
  - Lets us get feedforward layer over token
- Ignoring input:
  - Lets us discard stopwords
- Summing inputs:
  - Lets us compute a bag-of-words representation

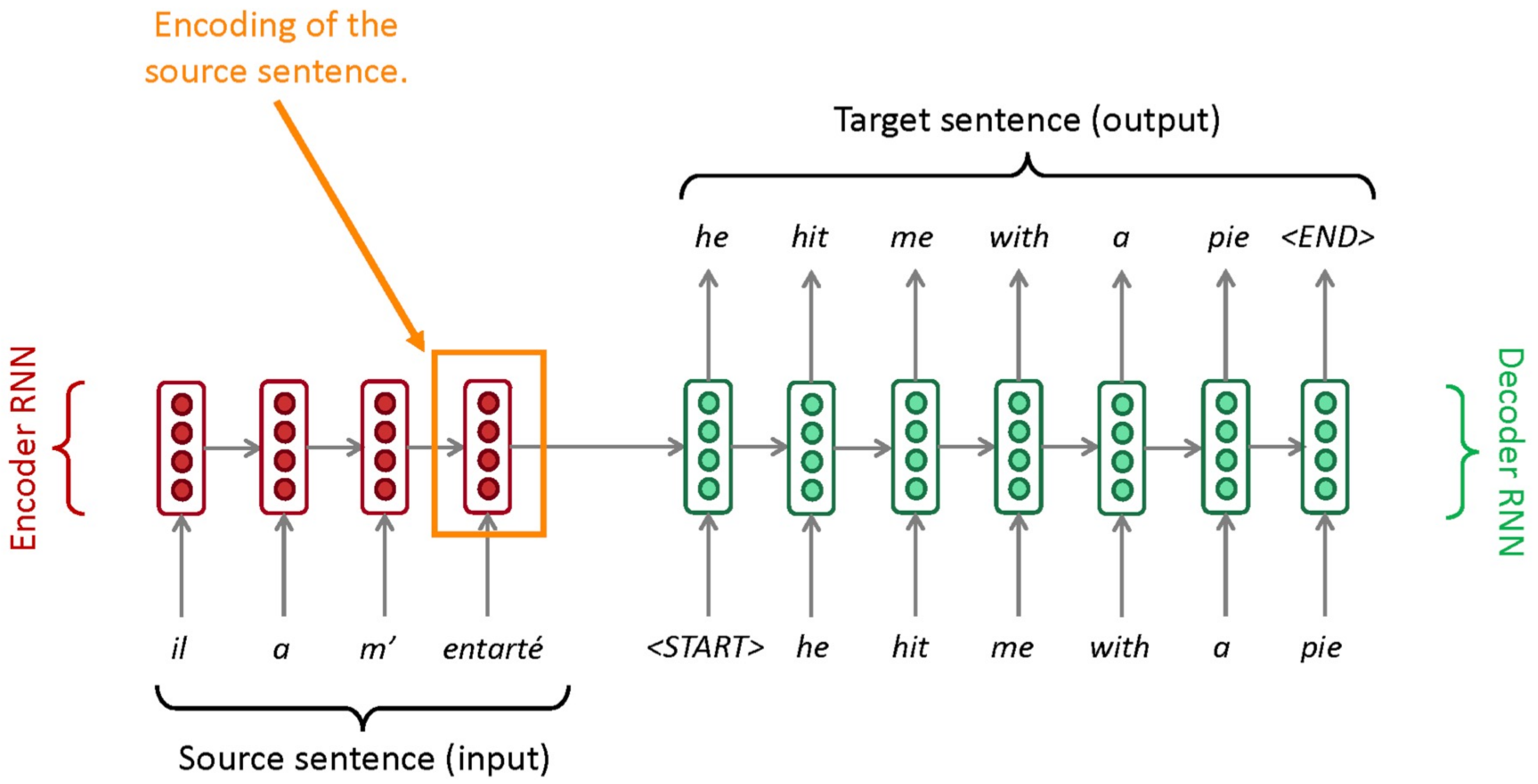
What about the Gradients?



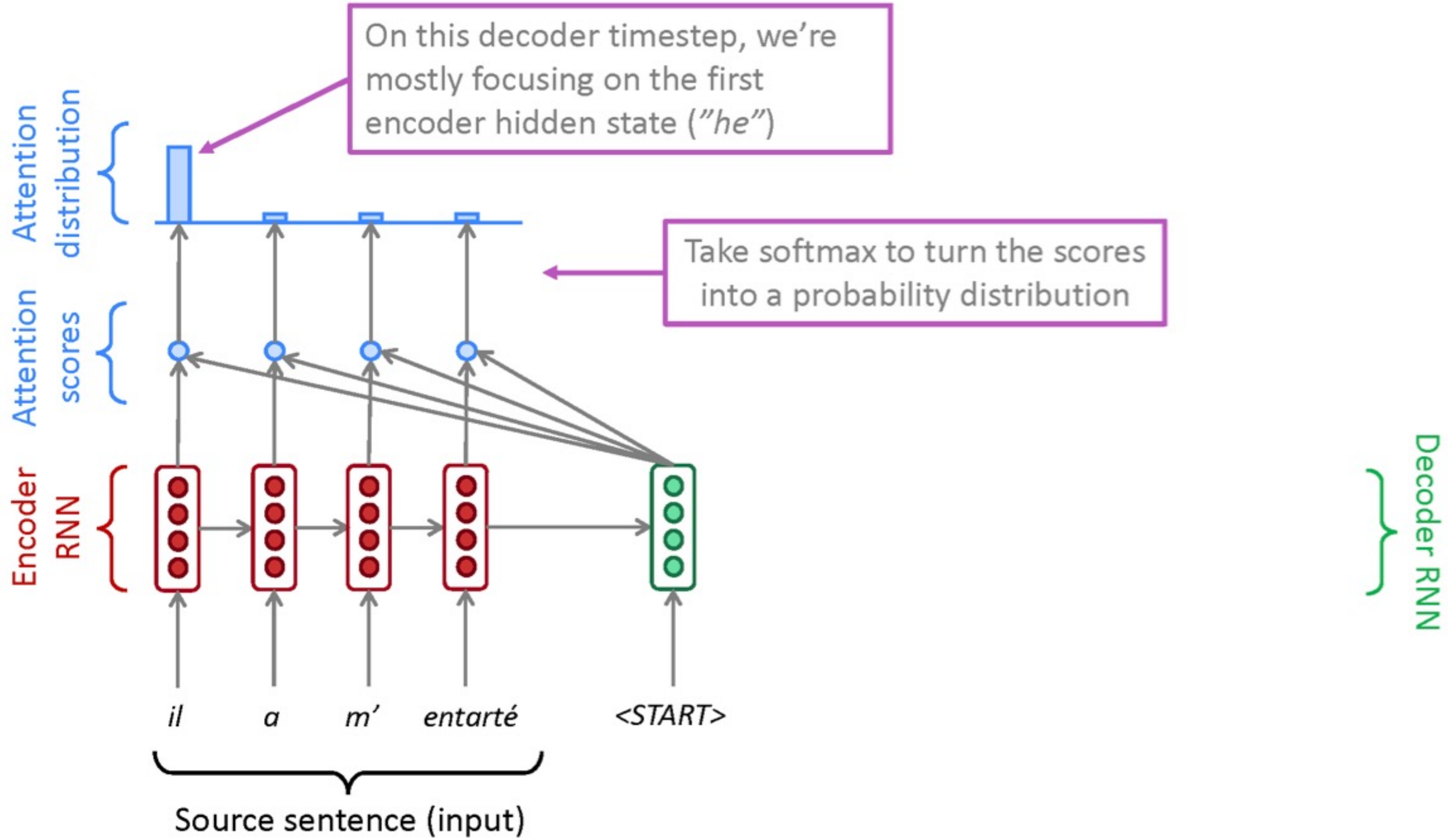
- Gradient still diminishes, but in a controlled way and generally by less — usually initialize forget gate = 1 to remember everything to start

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

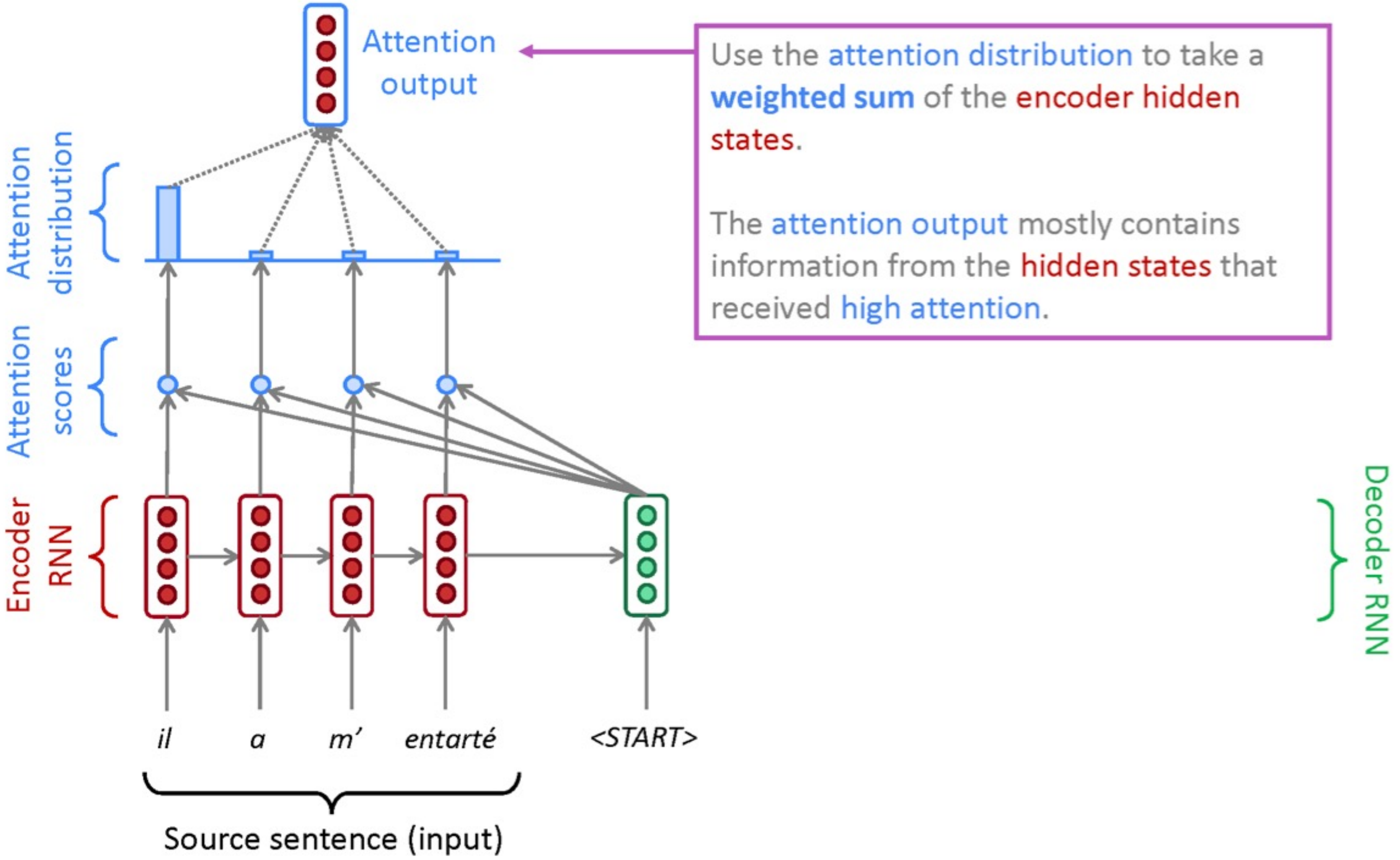
The Bottleneck Problem



LSTMs with Attention



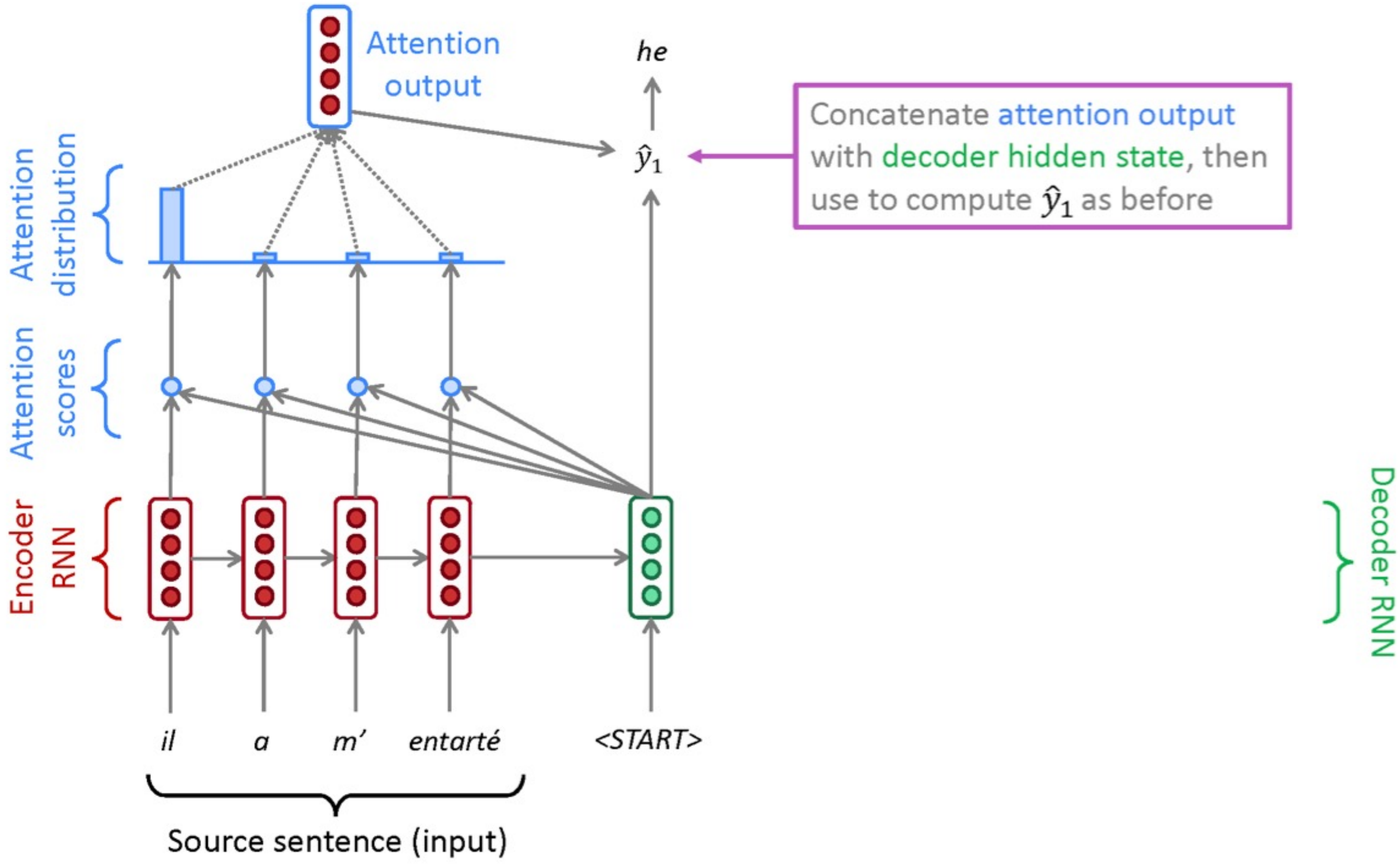
LSTMs with Attention



Use the attention distribution to take a **weighted sum** of the encoder hidden states.

The attention output mostly contains information from the hidden states that received high attention.

LSTMs with Attention



Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

Attention: in equations

- We have encoder hidden states $h_1, \dots, h_n \in \mathbb{R}^h$
- On timestep t, we have decoder hidden state $s_t \in \mathbb{R}^h$
- We get the attention scores e^t for this step:

$$e^t = [s_t^T h_1, \dots, s_t^T h_N] \in \mathbb{R}^N$$

- We take softmax to get the attention distribution α^t for this step (which outputs a probability distribution):

$$\alpha^t = \mathrm{softmax}(e^t) \in \mathbb{R}^N$$

- We use α^t to take a weighted sum of the encoder hidden states to get the attention output a_t:

$$a_t = \sum_{i=1}^{N} \alpha_i^t h_i \in \mathbb{R}^h$$

- Finally, we concatenate the attention output a_t with the decoder hidden state s_t and proceed as in the non-attention seq2seq model: $[a_t; s_t] \in \mathbb{R}^{2h}$

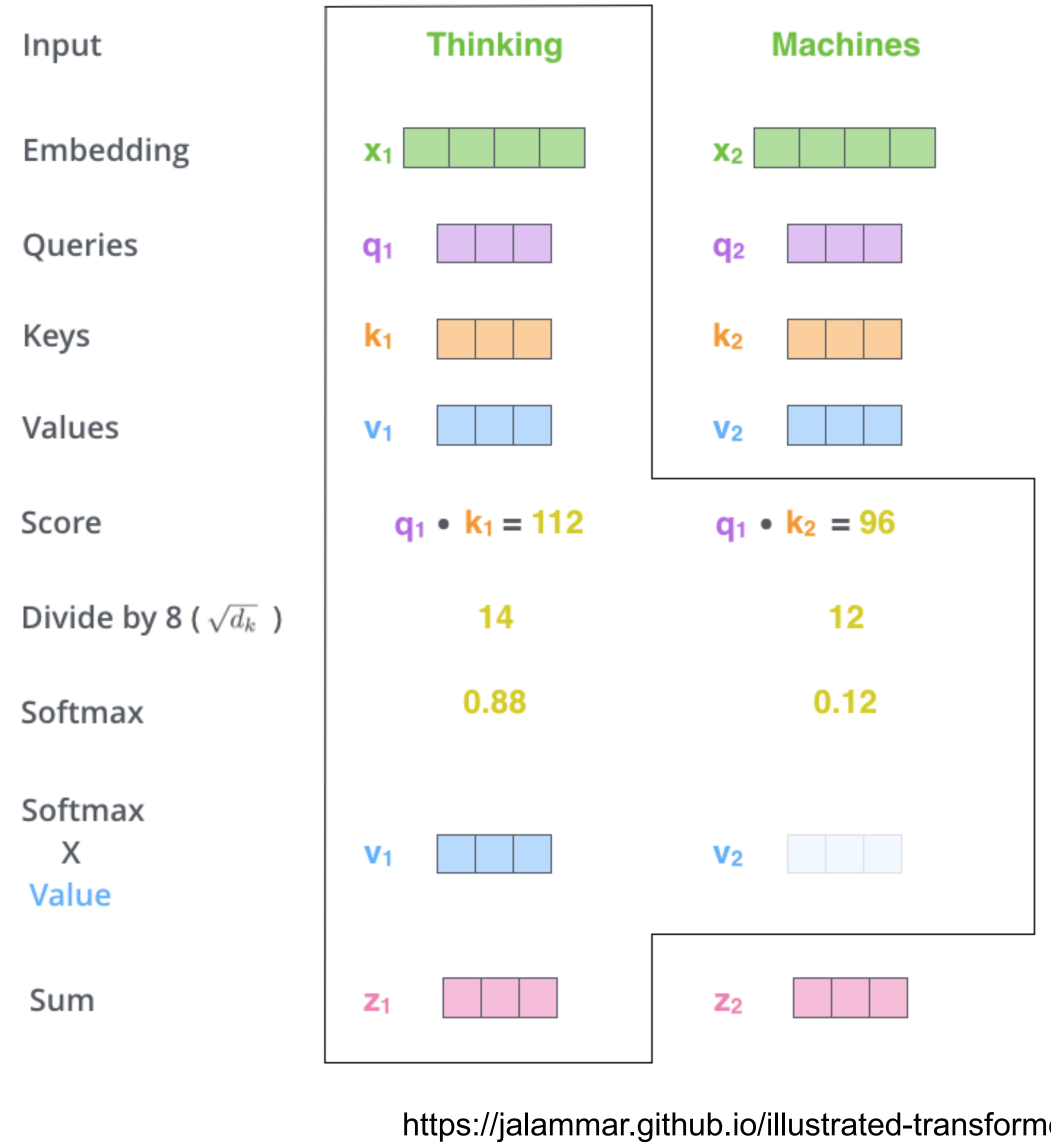
Transformers

Instead of an RNN, just use attention

High throughput & expressivity: compute queries, keys and values as (different) linear transformations of the input.

Attention weights are queries • keys; outputs are sums of weighted values.

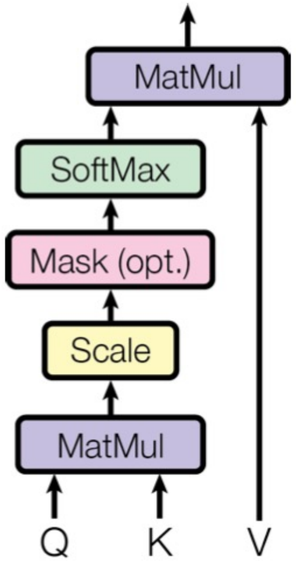
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



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

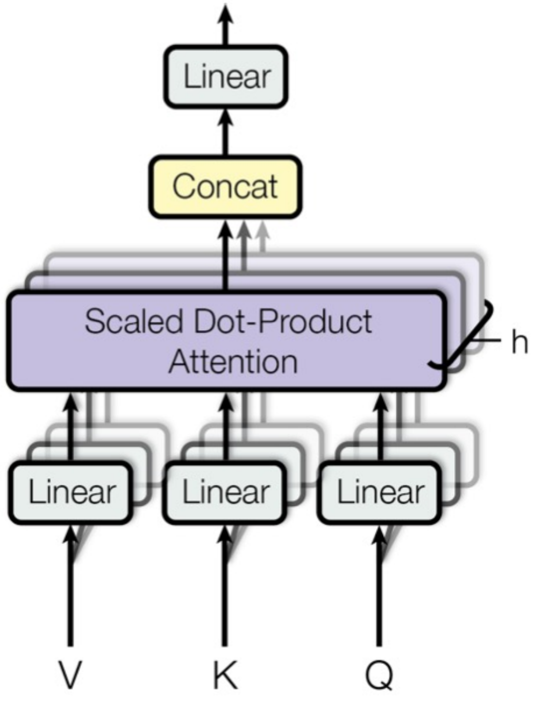
Transformer Architecture

Scaled Dot-Product Attention

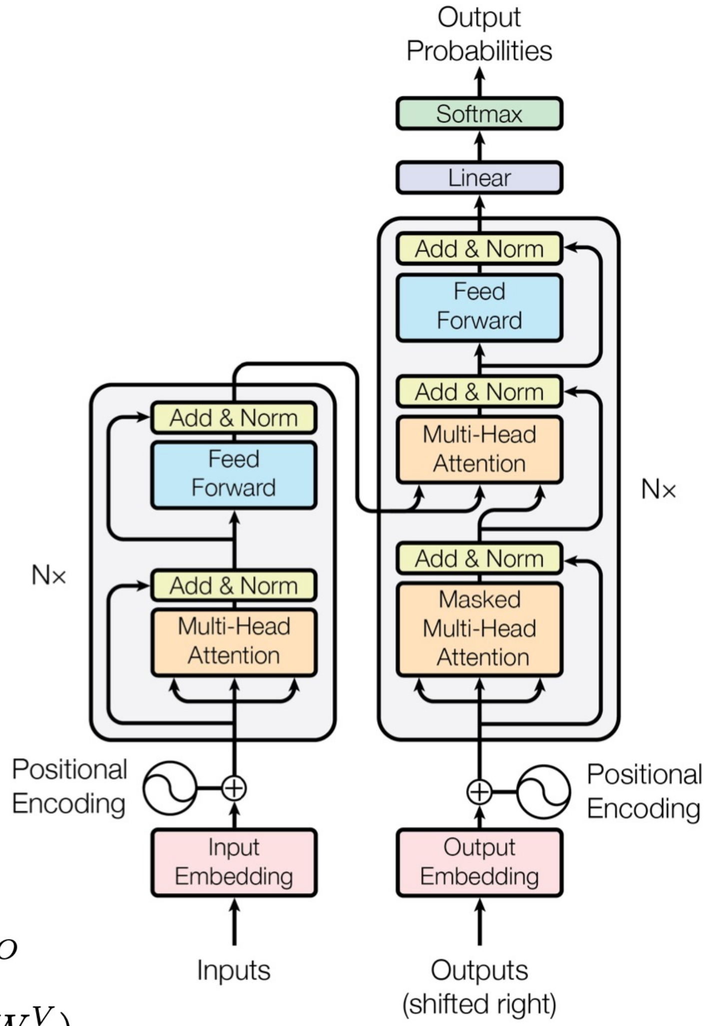


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-Head Attention



$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Transformer Language Models

Early Approaches:

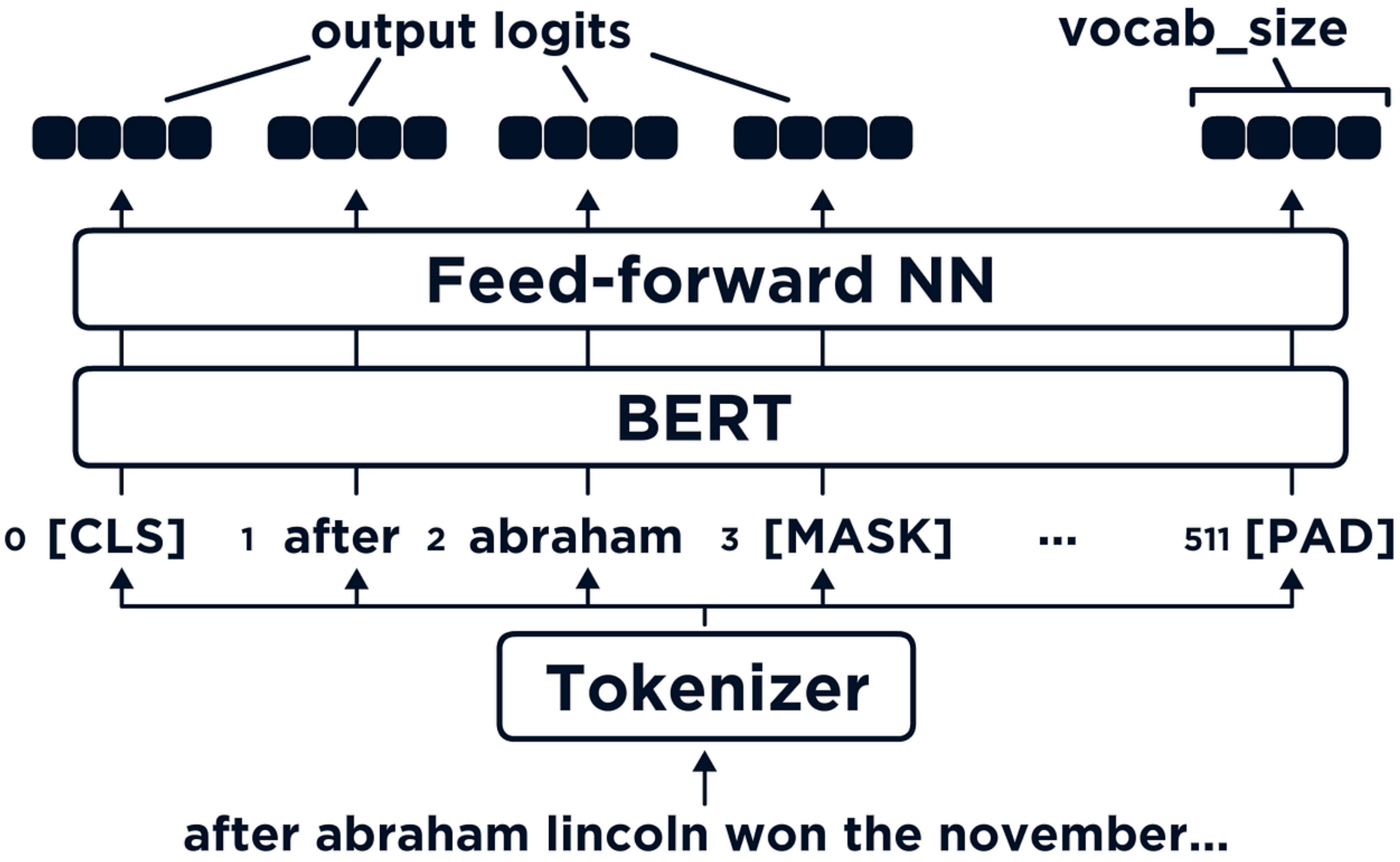
- ELMo (LSTM-based), BERT, RoBERTa, ELECTRA…
- Masked language modeling objective: learn representations

Recent Approaches:

- FLAN-T5: sequence-to-sequence model
- GPT-3/GPT-4, LLAMA, etc.: autoregressive language model
- Zero-shot / few-shot learning capabilities

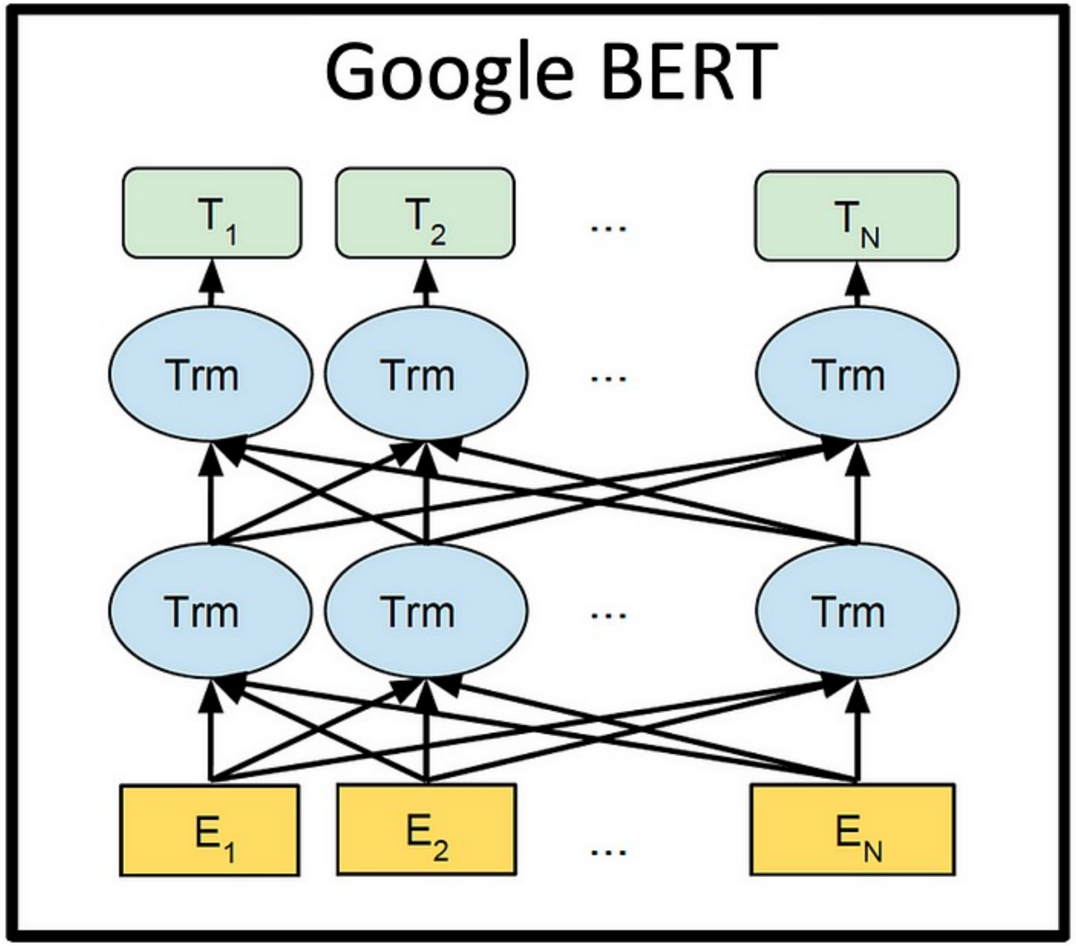
Masked Language Models

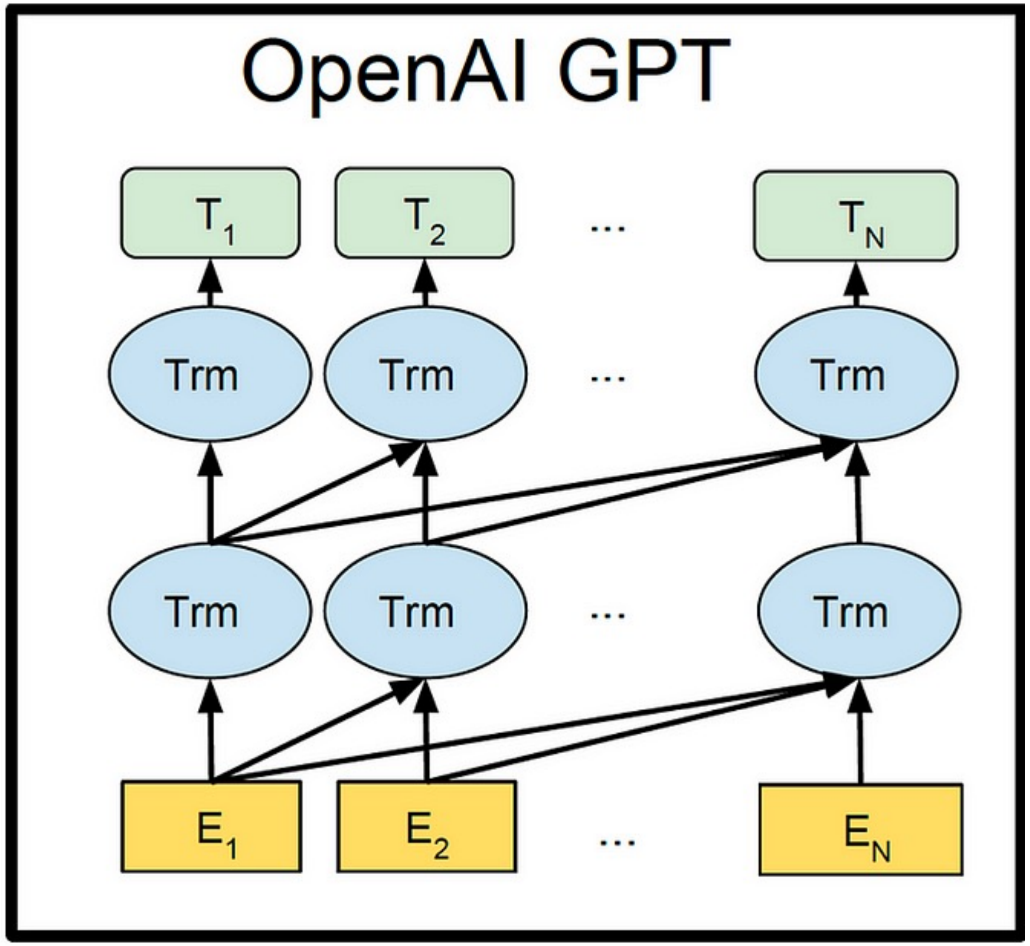
Key idea: learn representations and then fine-tune (training ≠ inference)



Autoregressive Language Models

Key idea: learn next-word prediction directly (training = inference)





Language Modeling Subsumes All Tasks (?)

NI

Can you translate "language models are powerful these days" into French?

Certainly! The translation of "language models are powerful these days" into French is:

"Les modèles de langage sont puissants de nos jours."

NI

Can you write a syntactic parse for the sentence "Kids Make Nutritious Snacks"?

Sure! Here's a syntactic parse for the sentence "Kids Make Nutritious Snacks":

scss

Copy code

```scss
(S
  (NP (NNS Kids))
  (VP (VBP Make)
    (NP (JJ Nutritious) (NNS Snacks))))
```

Instruction Tuning and RLHF

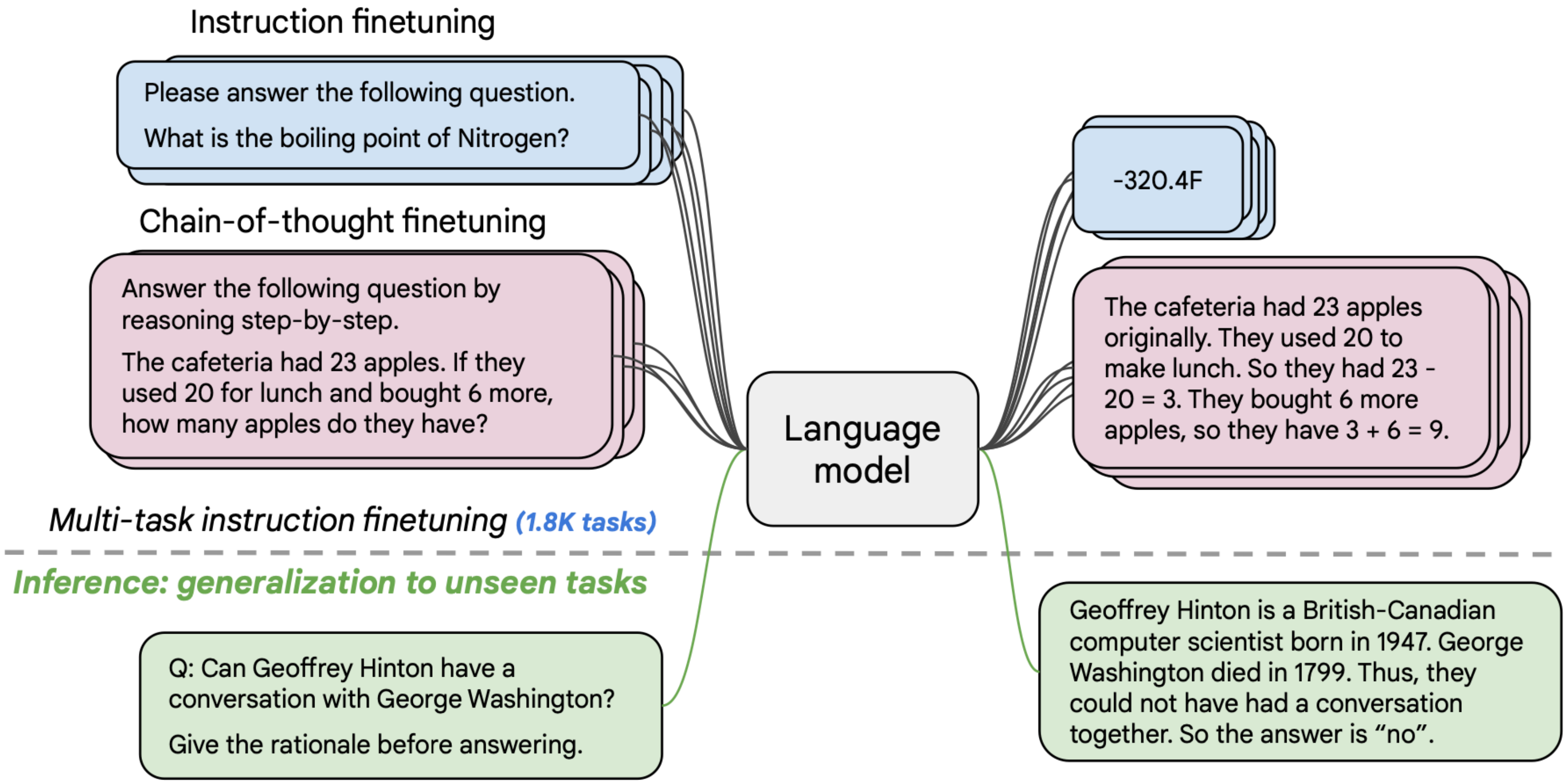
Key issue: language modeling ≠ assisting users

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

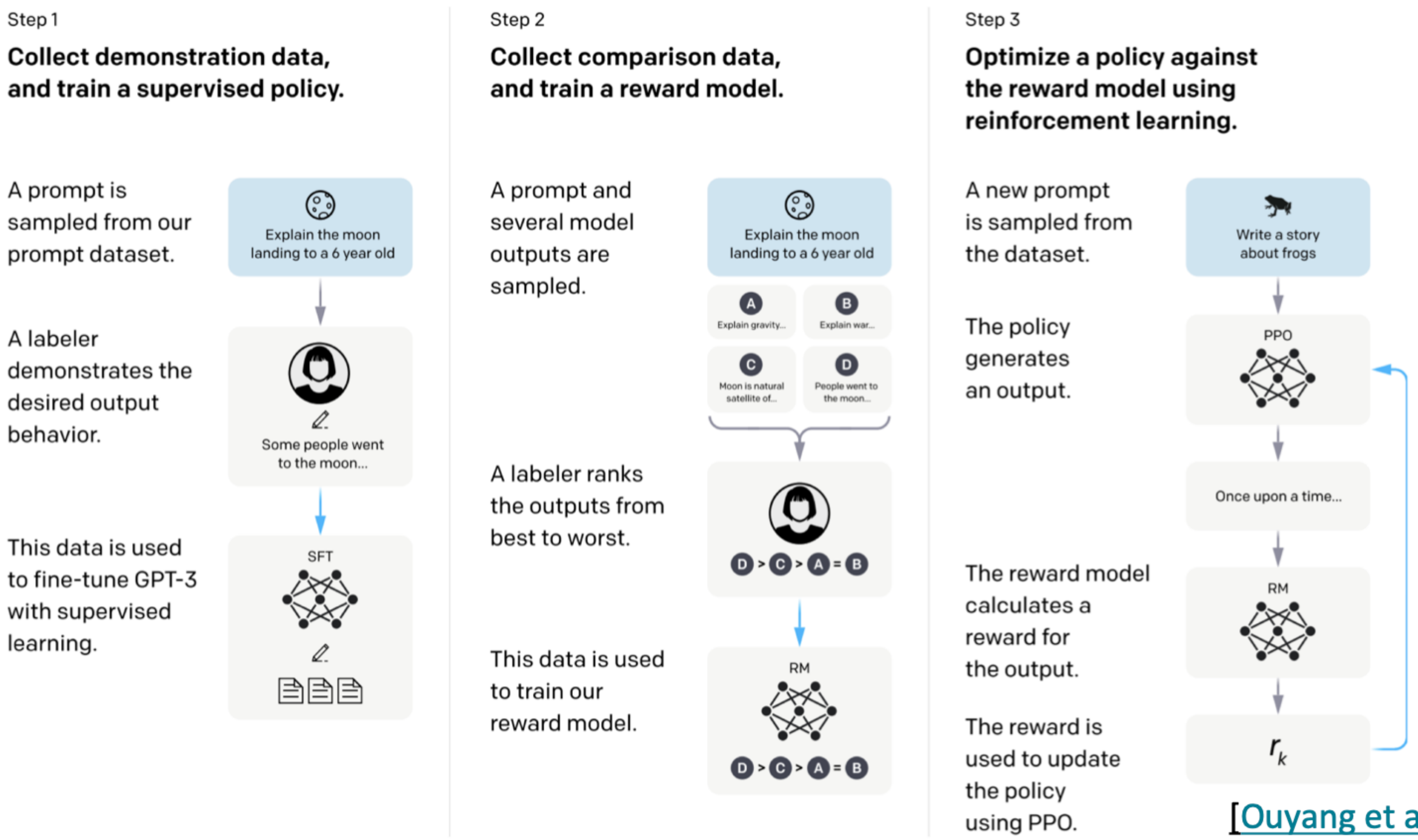
FLAN-T5

Collect examples of (instruction, output) pairs across many tasks and finetune an LM



InstructGPT: Reinforcement Learning from Human Preferences

Better approach: fine-tune LM to optimize a learned reward model over human preferences



[Ouyang et al., 2022]

Instruction Tuning and RLHF

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Going Forward

- Tool use (e.g., getting language models to use APIs)
- Grounding into non-linguistic inputs (e.g., vision, sensor data, etc.)
- Managing data security and privacy concerns
- More efficient / on-device / smaller / faster models
- Avoiding harmful, toxic, or undesirable outputs (e.g., spearfishing)
- Supporting multilinguality, esp. for low resource languages