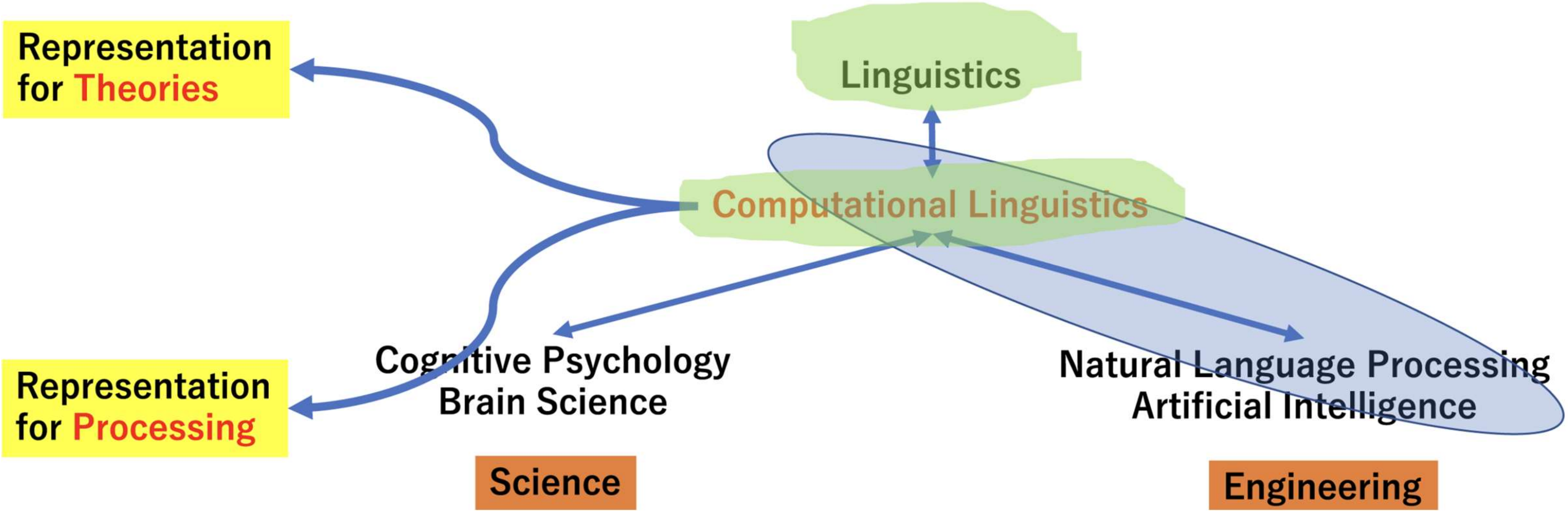
Natural Language Processing basics

NLP approaches

- Empirical - Statistical
  - Machine Learning – Corpora
    - Unsupervized
    - + Annotation - Supervized
    - Vector space models; Word embeddings
    - Neural Networks
  - Shallow parsing
- Rationalistic - Grammar-based
  - Parsing
  - Knowledge-based
  - Ontologies
  - Knowledge graphs



Junichi Tsujii; Natural Language Processing and Computational Linguistics (2021) https://doi.org/10.1162/coli_a_00420

Linguistics – the science that studies natural language

- Phonetics and phonology
- Morphology - lexicons
- Syntax - grammars
- Semantics – knowledge bases, ontologies, semantic spaces, embeddings
- Pragmatics and discourse

Grammars (Syntax)

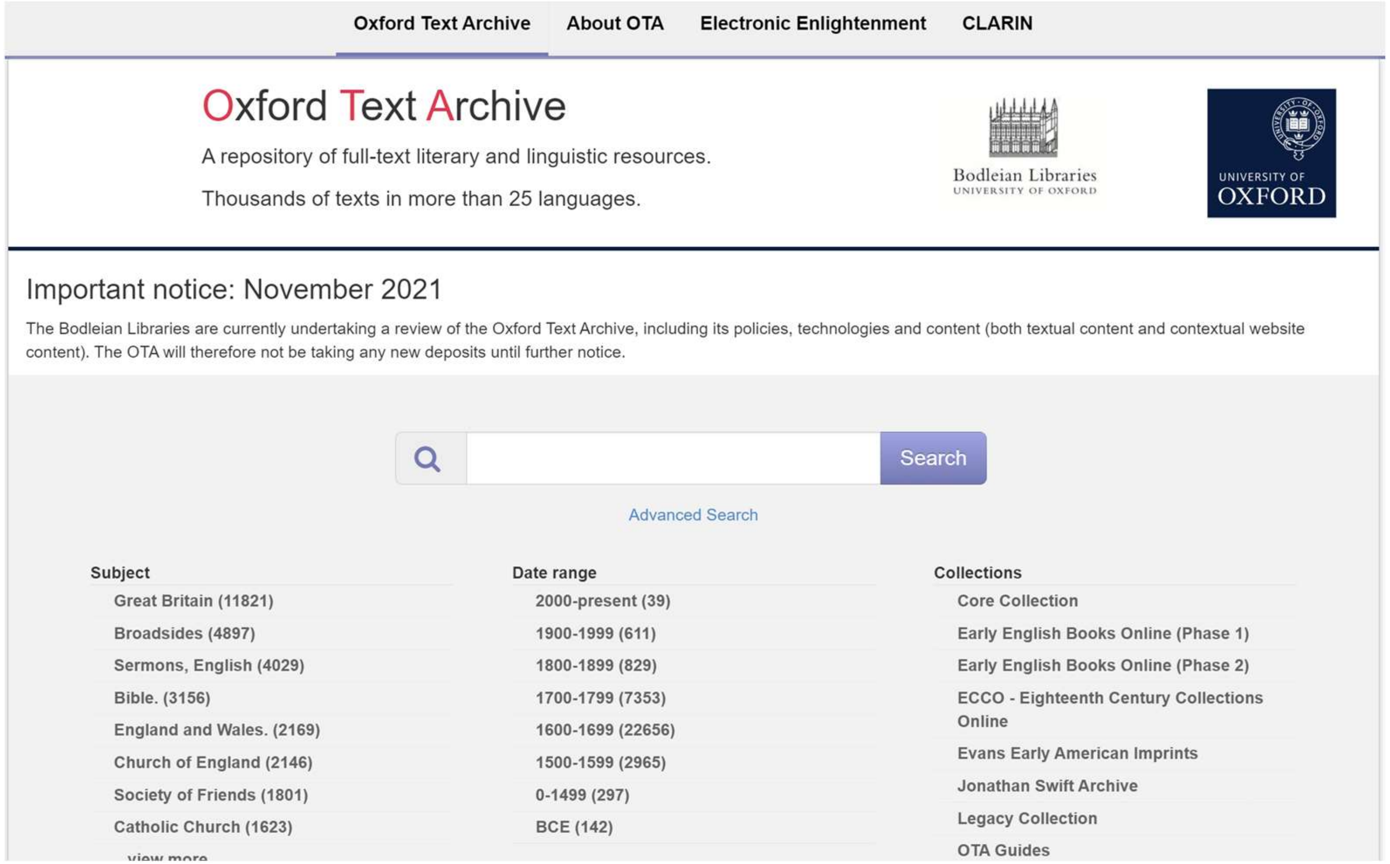
- Regular, Context Free, Context Dependent, General (Chomsky’s hierarchy)
- Dependency
- GPSG
- HPSG
- LFG
- (L)TAG

Corpus linguistics

- Empirical approach (based on datasets, not on rationalism)
- Based on corpora
- It may or not use computational techniques
- Introduced by John Sinclair (without NLP)

Corpus-Corpora

Collection(s) of naturally-occurring language text, chosen to characterize a state or variety of a language.
(John Sinclair, 1991)



https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm

Zipf's law – law of corpora

- A corpus of general text should satisfy a number of constraints, for example, Zipf's law (https://www.youtube.com/watch?v=fCn8zs912OE), which is specific to any language
- Other constraints should be satisfied

- It is an instance of a power law (Barabasy), which reflect natural properties of social networks and phenomena (e.g. the number of friends in a social networks)

Types of corpora

- Raw vs annotated
- Speech vs text
- General vs. specific
- Parallel corpora

Examples of general language corpora

- British National Corpus (BNC)
  http://www.natcorp.ox.ac.uk/
- Corpus of Contemporary American English (COCA)
  https://corpus.byu.edu/coca/
- Open American National Corpus (OANC)
  http://www.anc.org/
- CoRoLa - Corpus de referință pentru limba română contemporană
  http://89.38.230.23/corola_sound_search/

For various NLP learning tasks are many corpora (textual datasets)

See the Linguistic Data Consortium - https://www.ldc.upenn.edu/

Text structuring

- Tokenization (at the level of words)
- Bracketing (syntactical structures)
- Text segmentation
- Coreference resolution
- Discourse
- Rhetoric schema identification

Text annotation (in corpora)

- Syntactic
  - Part of speech – ex. noun, verb, …
  - “Bracketing” – syntactical structures
  - Treebanks – parsing trees
- Semantic – senses for words
- Pragmatic
  - Anaphoric annotation
  - Speech act annotation
- Discourse
- Rhetoric

Annotation languages

- SGML
- XML
- TEI
  (Text Encoding Initiative - https://tei-c.org/)
- Others

Machine learning with corpora

- Hidden Markov Models
- Naïve Bayes
- Support Vector Machines
- ...
- Neural Networks

Deep Neural Networks for NLP

Natural Language Processing with Deep Neural Networks

- Convolutional Neural Networks
- Recurrent Neural Networks (RNN)
  - Long Short-Term Memory (LSTM)
  - Bi-directional LSTM
  - Gated Recurrent Units (GRU)
  - Enconder-Decoder
  - Enconder-Decoder with Attention
- Transformers (Bert, GPT-2, GPT-3, GPT3.5, GPT4, **ChatGPT**…)

Natural Language Generation with Deep Neural Networks

- Deep Neural Networks
  - For example, **ChatGPT**
- Training with a corpus of literature
- Generating new texts in the “style” of the learned corpus

Pre-processing for DNN

- Tokenization
- Embedding

Language specific features considered in NLP with DNN

- Syntactic structures, meaning (semantics and pragmatics), and discourse are statistically learned from raw corpora

- Time sequencing

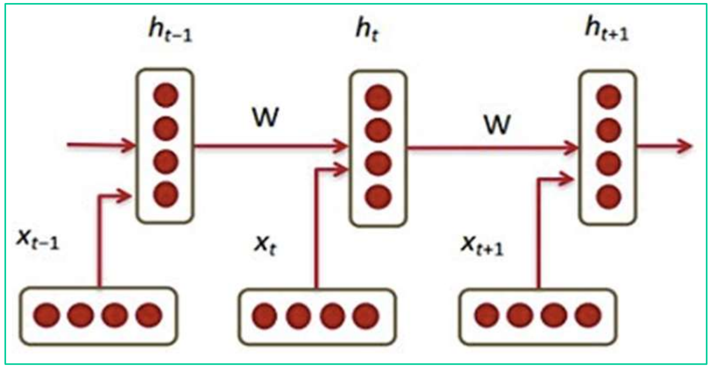
- Long distance dependencies

Recurrent Neural Networks (RNNs)

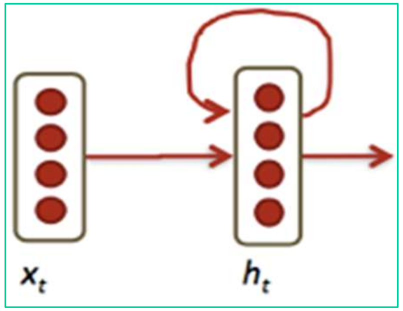
Main RNN idea for text:

Condition on **all previous words**

Use same set of weights at all time steps $h_t = \sigma(W^{(hh)}h_{t-1} + W^{(hx)}x_t)$

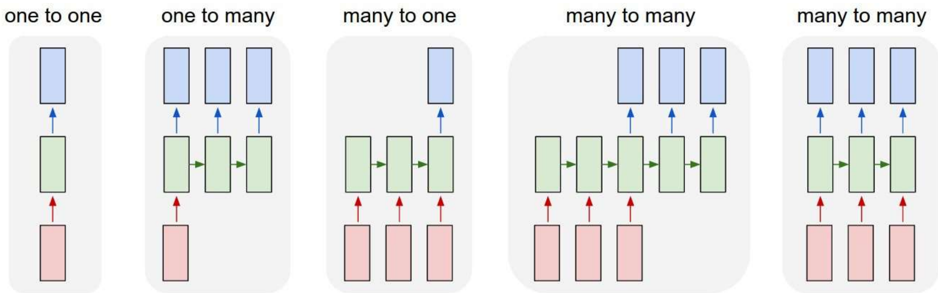


$\triangleq$



Feed Backward Network

https://pbs.twimg.com/media/C2j-8j5UsAACgEK.jpg



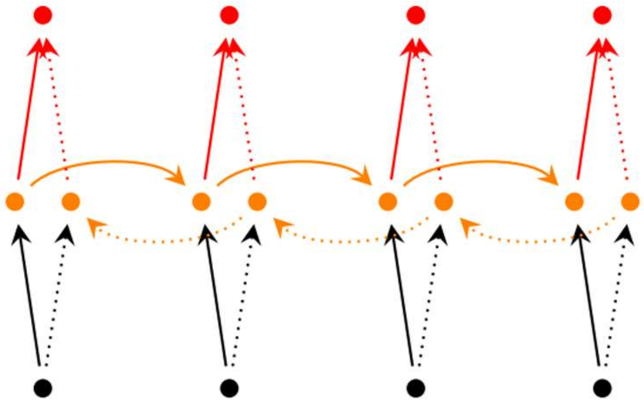
😁 Stack them up, Lego fun!

😞 **Vanishing gradient problem**

https://discuss.pytorch.org/uploads/default/original/1X/6415da0424dd66f2f5b134709b92baa59e604c55.jpg

Bidirectional RNNs

Main idea: incorporate both left and right context
output may not only depend on the **previous** elements in the sequence, but also **future** elements.



http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

$$\overrightarrow{h}_t = \sigma(\overrightarrow{W}^{(hh)}\overrightarrow{h}_{t-1} + \overrightarrow{W}^{(hx)}x_t)$$

$$\overleftarrow{h}_t = \sigma(\overleftarrow{W}^{(hh)}\overleftarrow{h}_{t+1} + \overleftarrow{W}^{(hx)}x_t)$$

$$y_t = f\left(\left[\overrightarrow{h}_t; \overleftarrow{h}_t\right]\right)$$

past and future around a single token

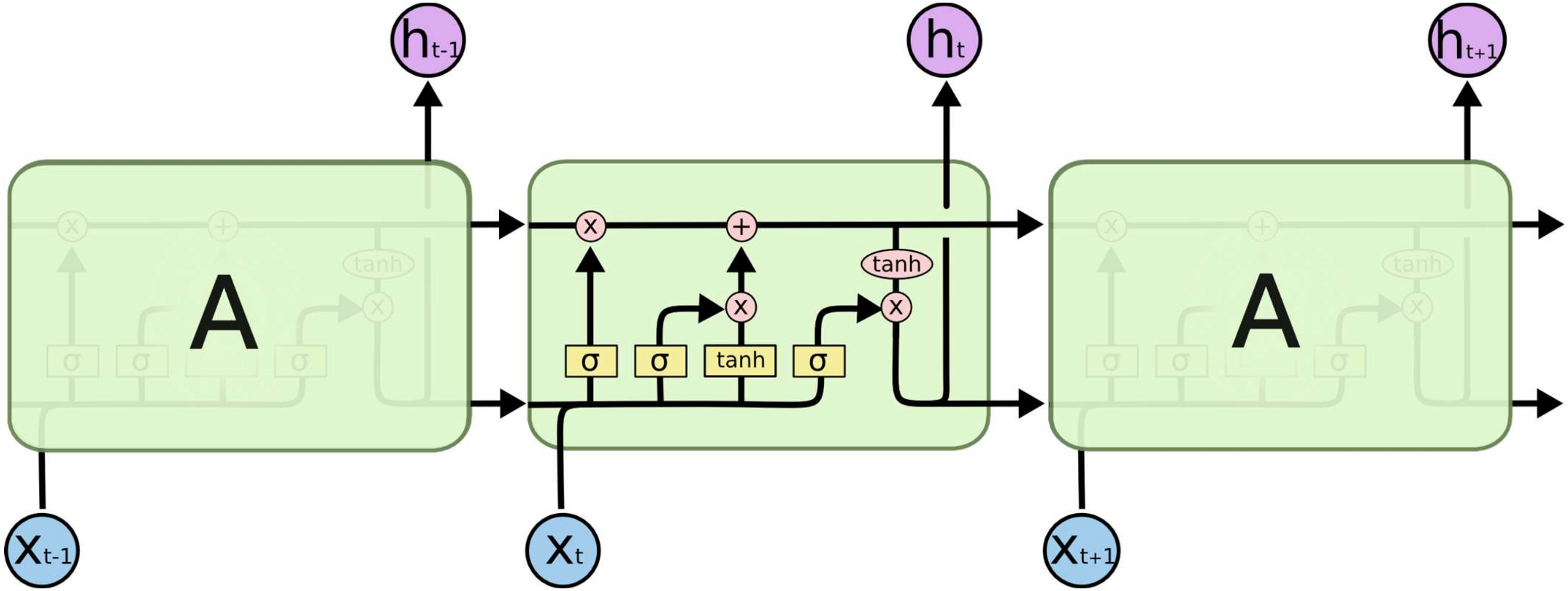
two RNNs stacked on top of each other

output is computed based on the hidden state of both RNNs $\left[\overrightarrow{h}_t; \overleftarrow{h}_t\right]$

Ismini Lourentzou

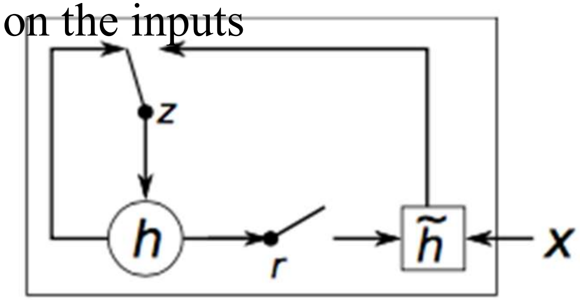
Long-Short Term Memory (LSTM)

- a special kind of RNN, capable of learning long-term dependencies
- some information is forgoten



Gated Recurrent Units (GRUs)

Simpler case of LSTM
Main idea:
keep around memory to capture **long dependencies**
Allow error messages to flow at **different strengths** depending on the inputs



http://www.wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/

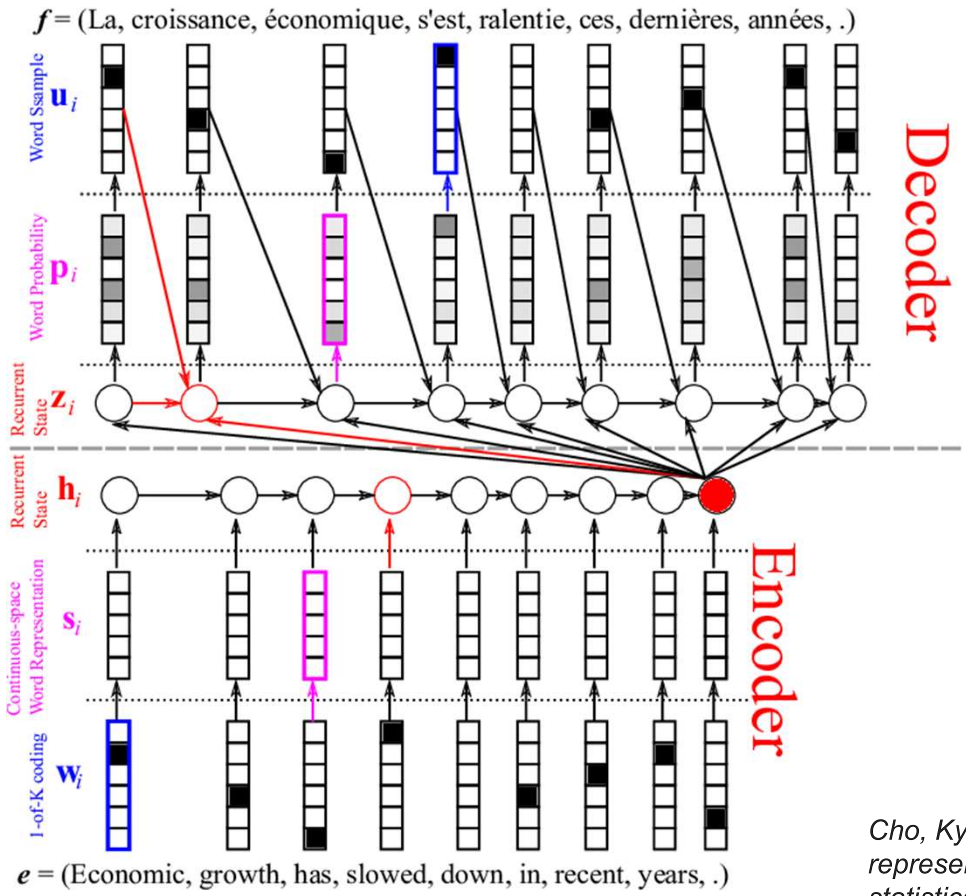
Standard RNN computes hidden layer at next time step directly
$h_t = \sigma(W^{(hh)} h_{t-1} + W^{(hx)} x_t)$

Compute an update gate based on current input word vector and hidden state

$$z_t = \sigma(U^{(z)} h_{t-1} + W^{(z)} x_t)$$

Controls how much of past state should matter now
If z close to 1, then we can copy information in that unit through many steps!

Sequence2Sequence or Encoder-Decoder model



Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP 2014

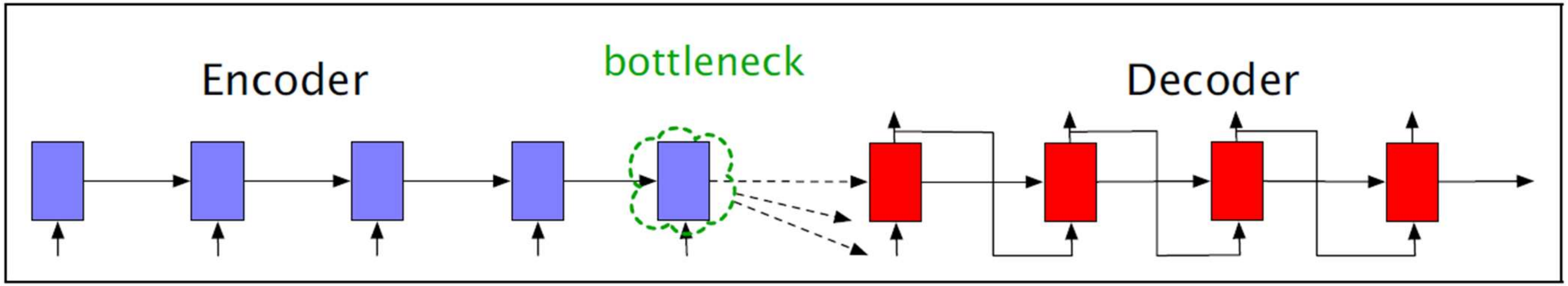
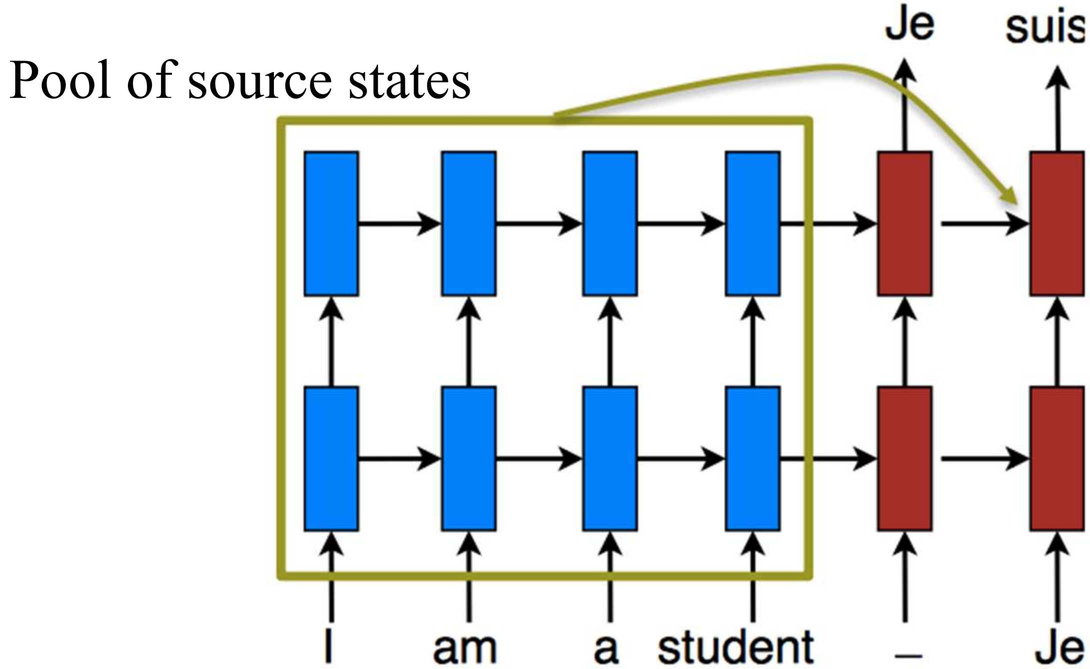


Figure 9.21 Requiring the context *c* to be only the encoder's final hidden state forces all the information from the entire source sentence to pass through this representational bottleneck.

(Jurafsky & Martin, 2024)

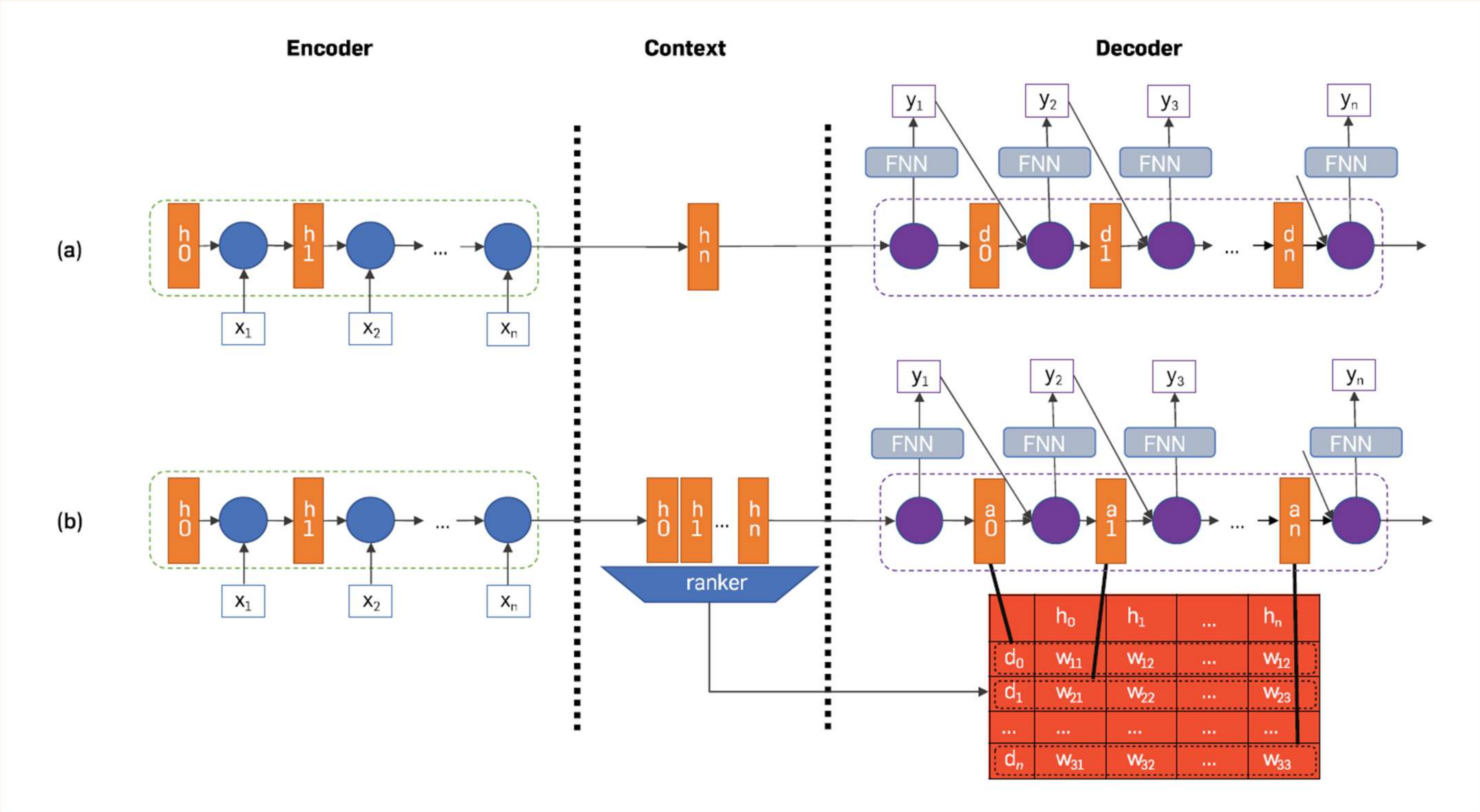
Attention Mechanism



Bahdanau D. et al. "Neural machine translation by jointly learning to align and translate." ICLR (2015)

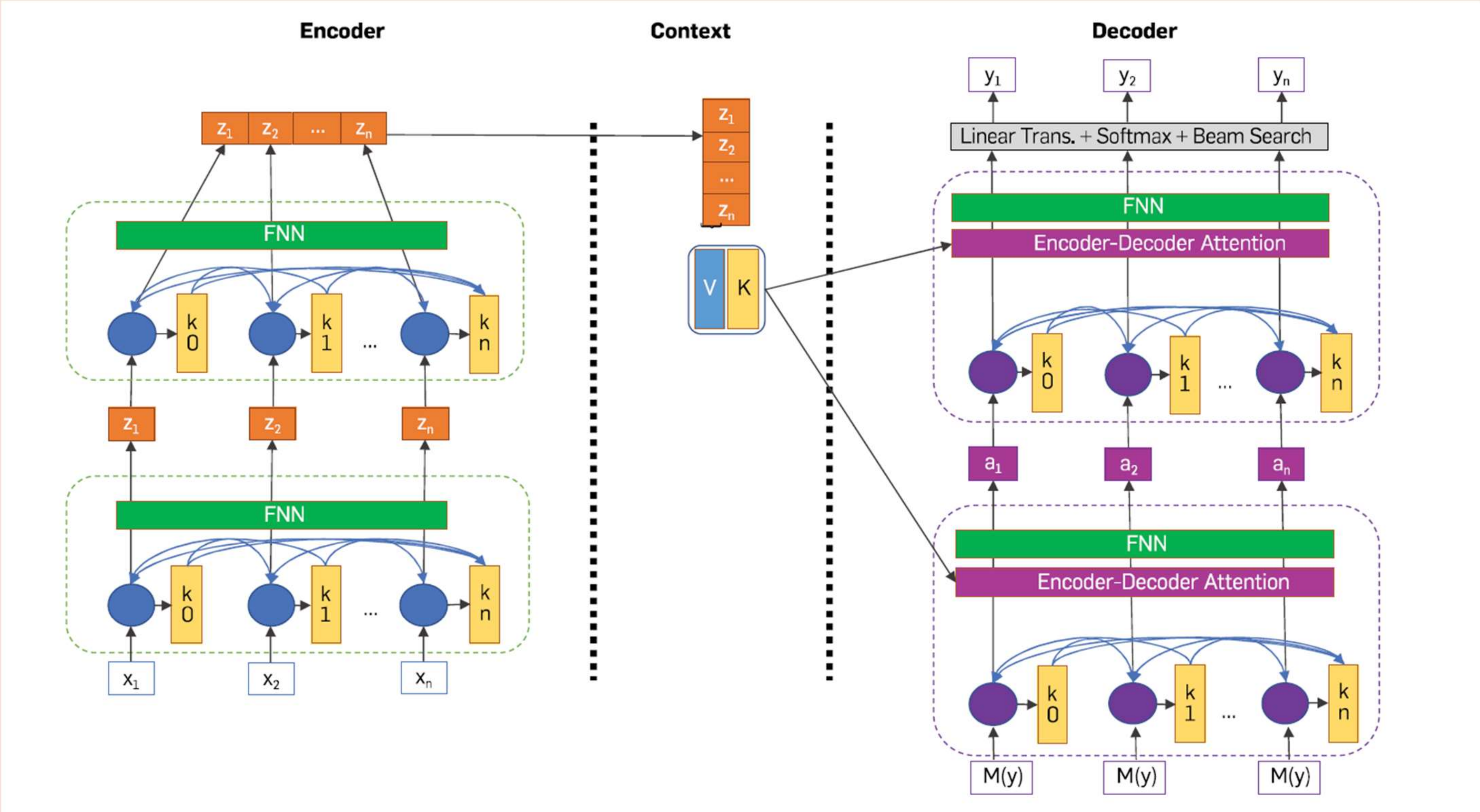
Main idea: retrieve as needed

Figure 1. Difference between encoder-decoder methods (a) without and (b) with attention. Notice that the circles represent the same set of weights changing at different timesteps.

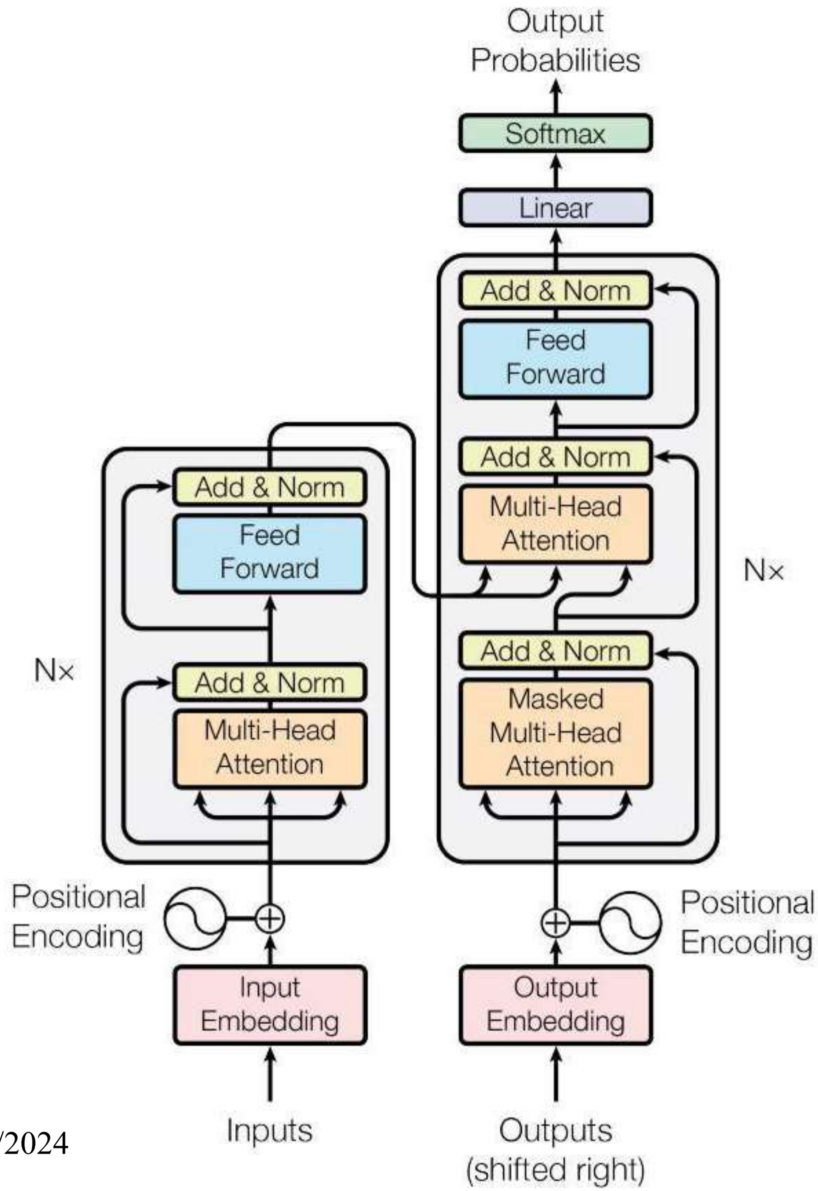


(Souza dos Reis et al, 2021, https://cacm.acm.org/research/transformers-aftermath/)

Figure 2. An example of Transformers composed of two encoders and two decoders. Notice that the decoders receive the context—projected in two vectors *v* and *k*—from the topmost encoder.

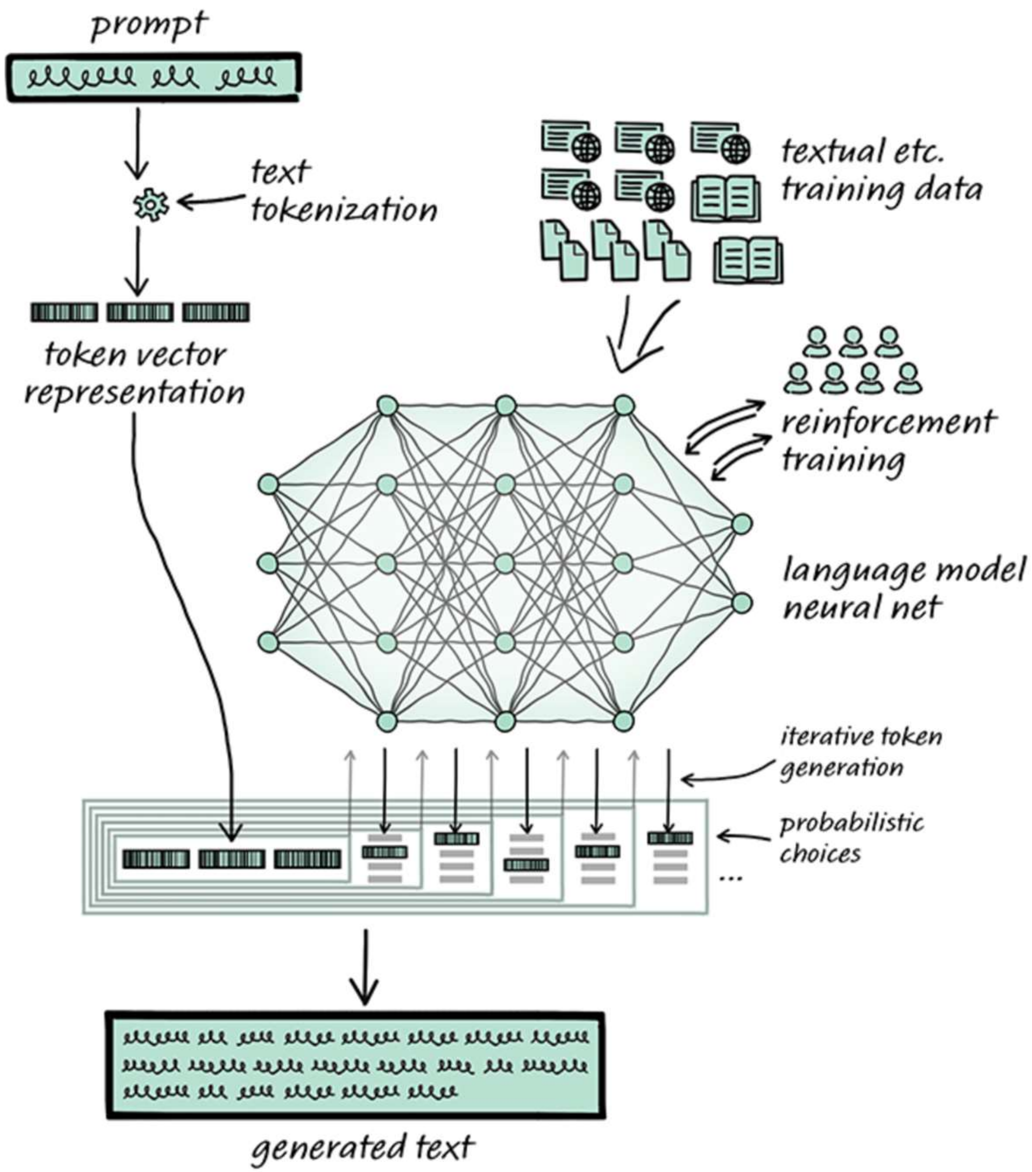


(Souza dos Reis et al, 2021, https://cacm.acm.org/research/transformers-aftermath/)



Attention is all you needs!

Transformer



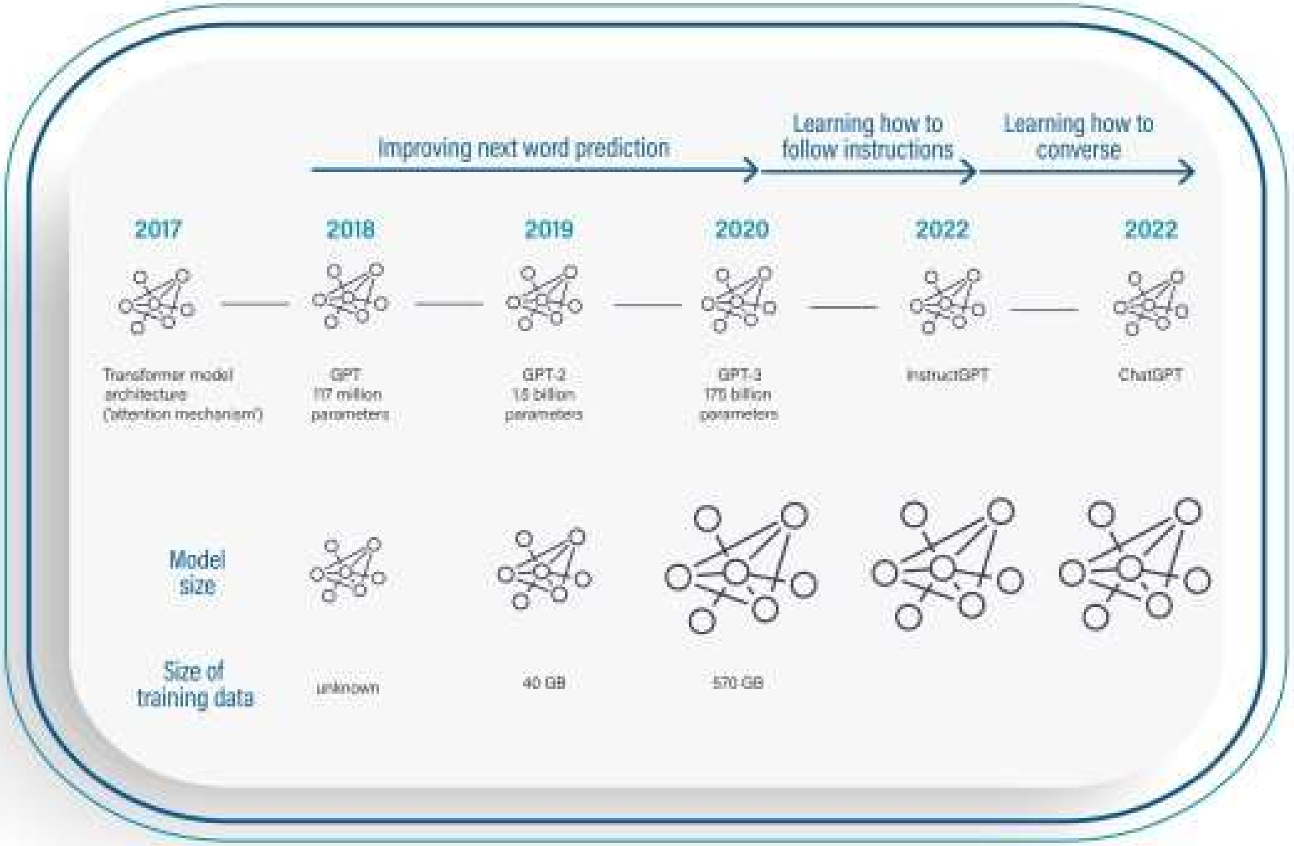
ChatGPT

(Chat Generative Pretraining Transformer)

Is a Large Language Model (LLM)

https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/

Evolution of Large Language Models (ChatGPT)



https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/

ChatGPT has a number of neurons comparable to a human brain

- 100 billion neurons
- over 100 layers
- 100 trillion synapses

https://medium.com/@fenjiro/chatgpt-gpt-4-how-it-works-10b33fb3f12b3/7/2024

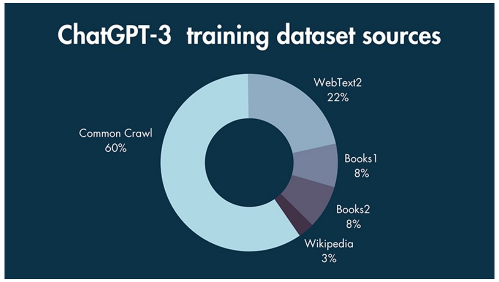
- Human Brain - 100 billion neurons and 10× more glial cells.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2776484/

Training of ChatGPT

- Some ChatGPT commentators have estimated that if ChatGPT was to be trained on a single NVIDIA Tesla V100 'Graphics Processing Unit' (GPU) that it would take around 355 years to complete ChatGPT's training on its training dataset.
- OpenAI reportedly used 1,023 A100 GPUs to train ChatGPT, so **it is possible that the training process was completed in as little as 34 days**.
- The costs of training ChatGPT is estimated to be just under $5 million dollars.

https://lambdalabs.com/blog/demystifying-gpt-3



Training of ChatGPT

- **60%** of ChatGPT-3's dataset was based on a filtered version of what is known as 'common crawl' data, which consists of web page data, metadata extracts and text extracts from over 8 years of web crawling.
- **22%** of ChatGPT-3's dataset came from 'WebText2', which consists of Reddit posts that have three or more upvotes.
- **16%** of ChatGPT-3's dataset come from two Internet-based book collections. These books included fiction, non-fiction and also a wide range of academic articles.
- **3%** of ChatGPT-3's dataset comes from the English-language version of Wikipedia.
- **93%** of ChatGPT-3's data set was in English

https://arxiv.org/pdf/2005.14165.pdf