Language Models II

Sandeep Soni

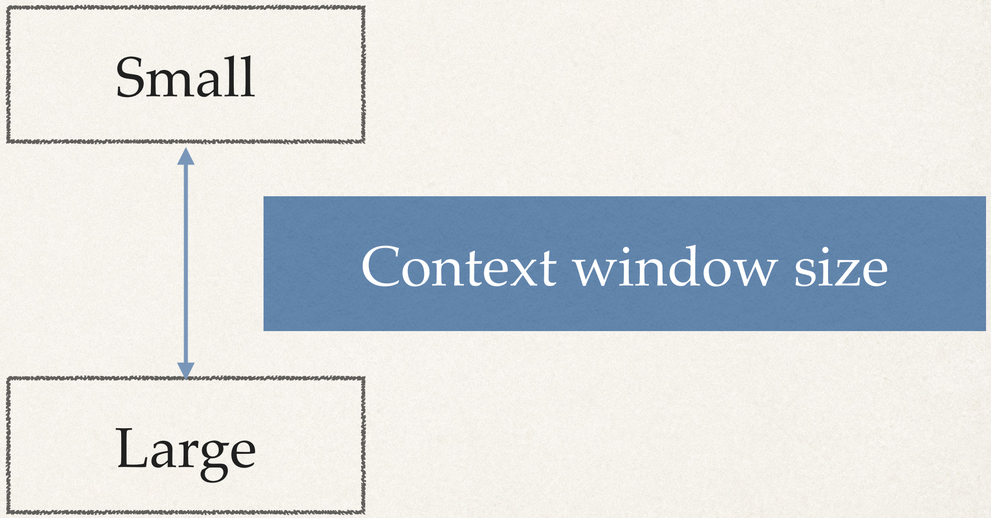
10/25/2023

Hw8

- What is the effect of context window size on word2vec embeddings?

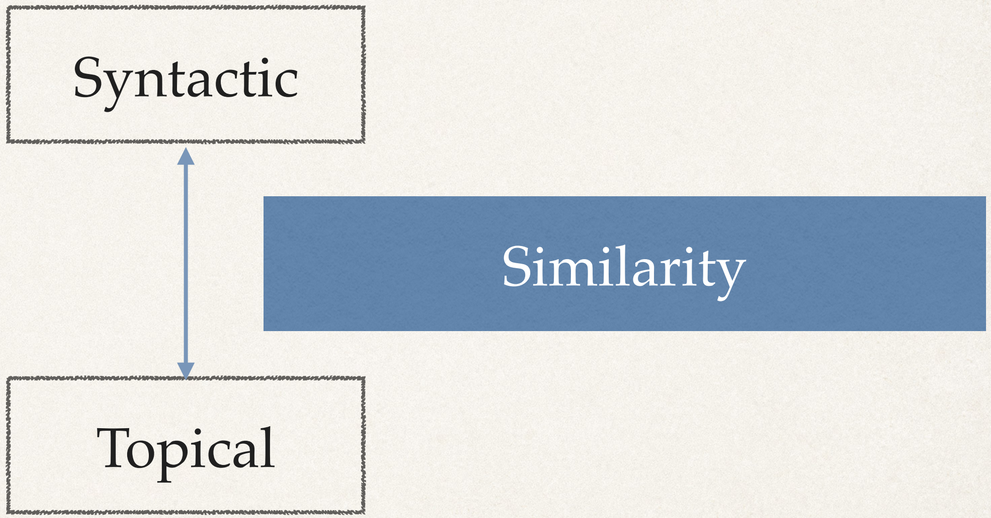
Effect of context size

- run is closer to running
- run is closer to walk
- run is closer to smile



Effect of context size

- run is closer to running
- run is closer to walk
- run is closer to smile



N-gram language models

$$P(x) = \prod_i P(x_i)$$

unigram

$$P(x) = \prod_i P(x_i \mid x_{i-1})$$

bigram

$$P(x) = \prod_i P(x_i \mid x_{i-2}, x_{i-1})$$

trigram

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

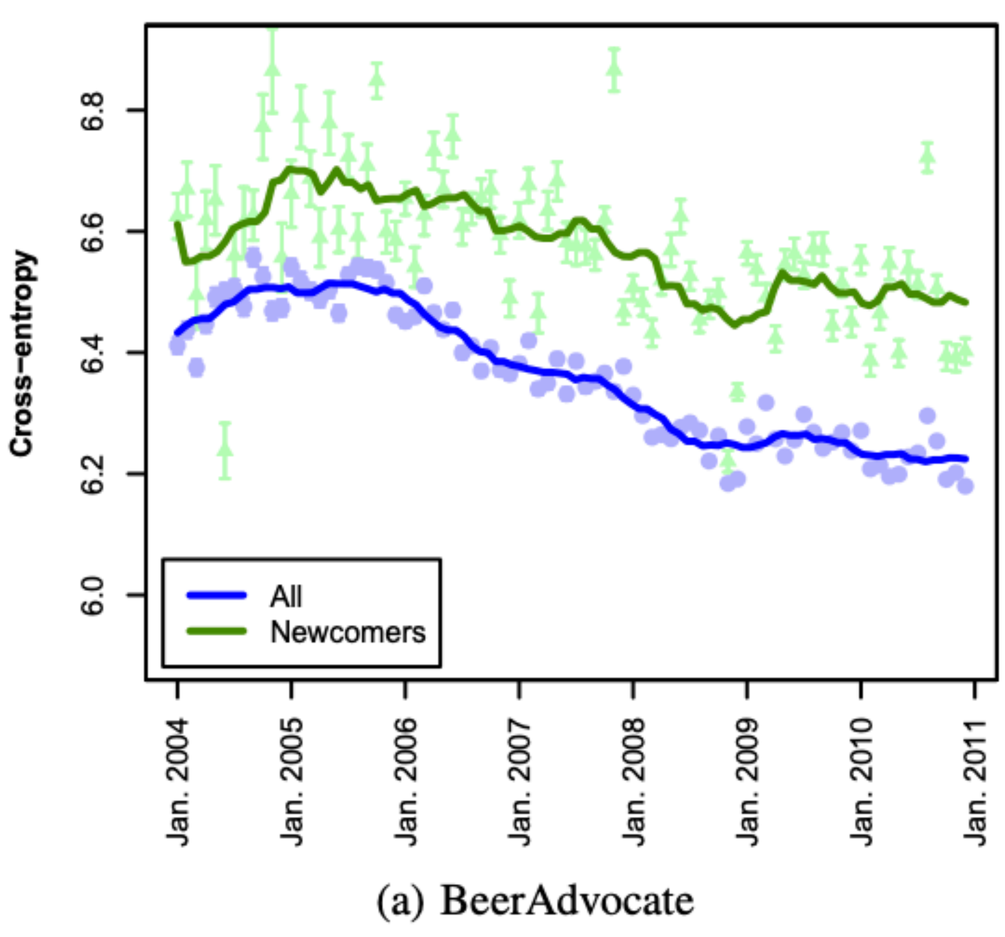
Jure Leskovec
Stanford University
jure@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

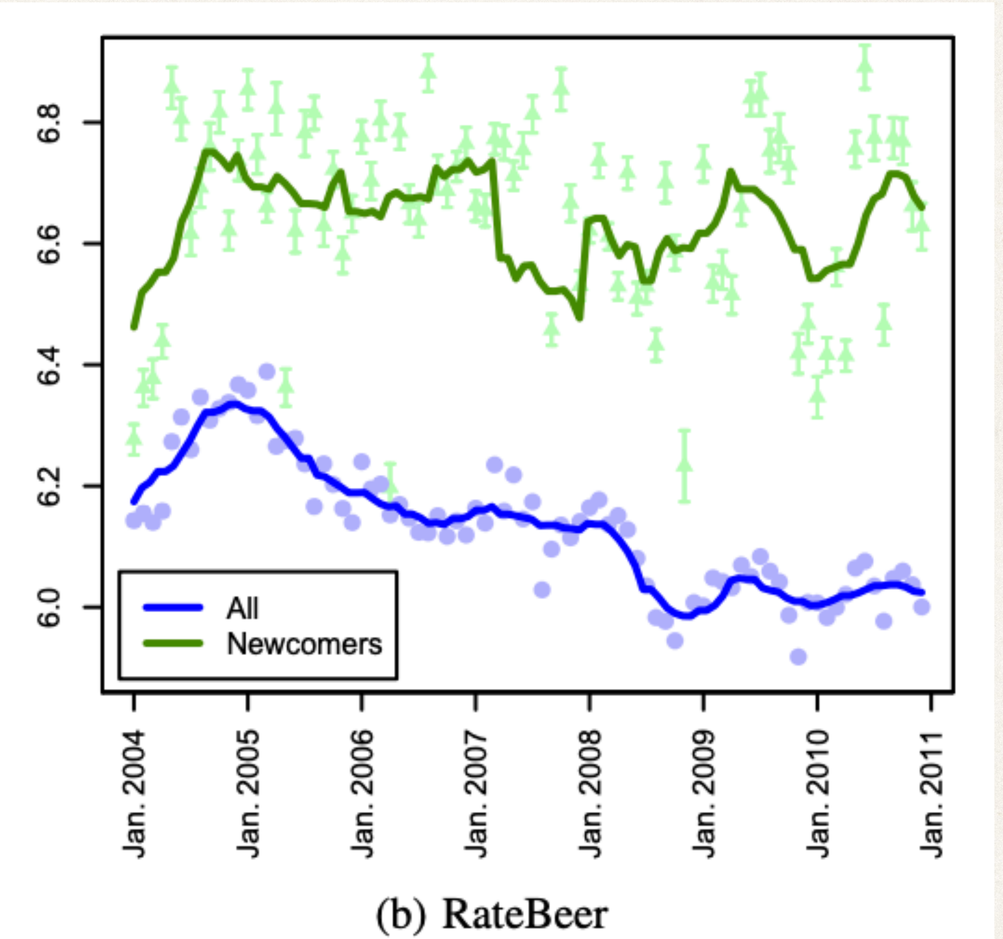
Who contributes to changing linguistic norms?

Snapshot language models

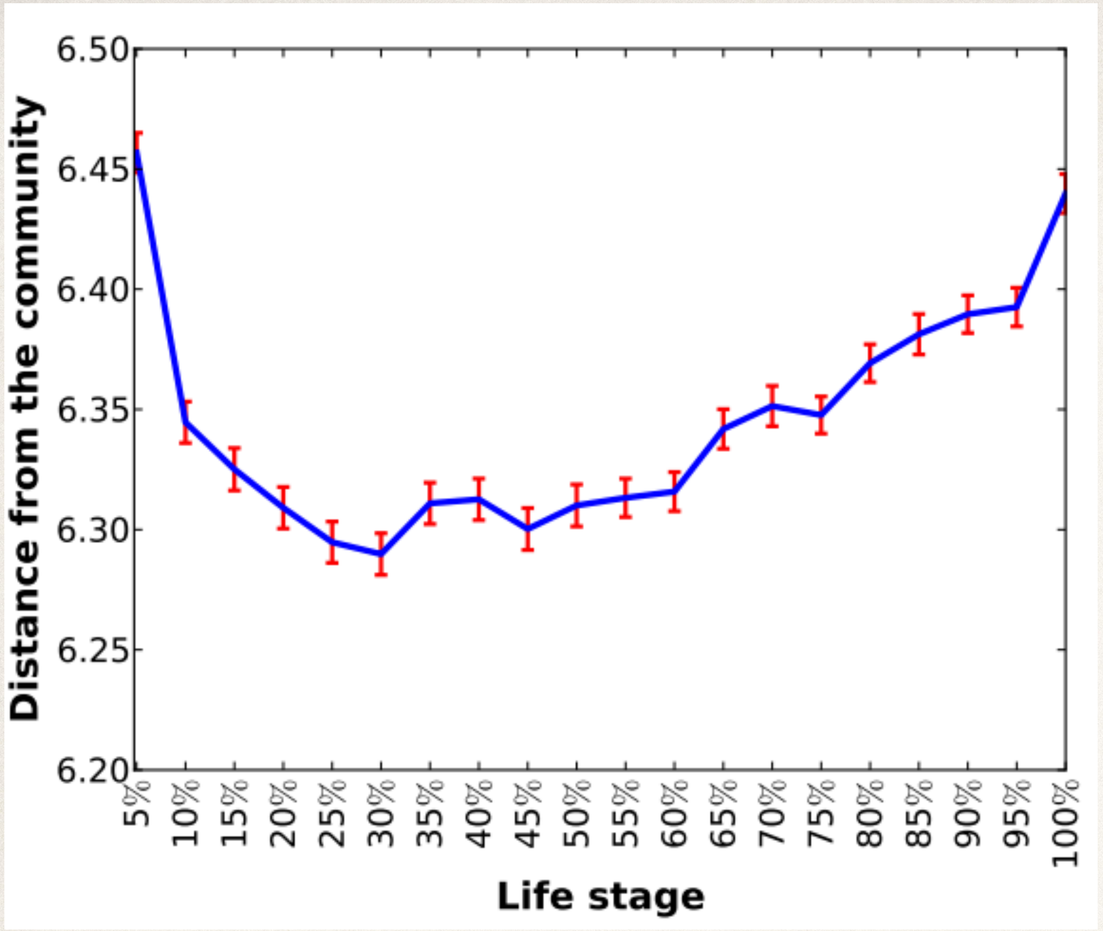
- Create a bigram language model SLM_m for every month m; this is called a snapshot language model
- Q. How surprising is a document with respect to any month?
- Calculate cross-entropy $H(p, SLM_m) = -\frac{1}{N}\sum_i \log P_{SLM_m}(b_i)$, where $b_1, b_2, \ldots, b_N$ are bigrams from the post

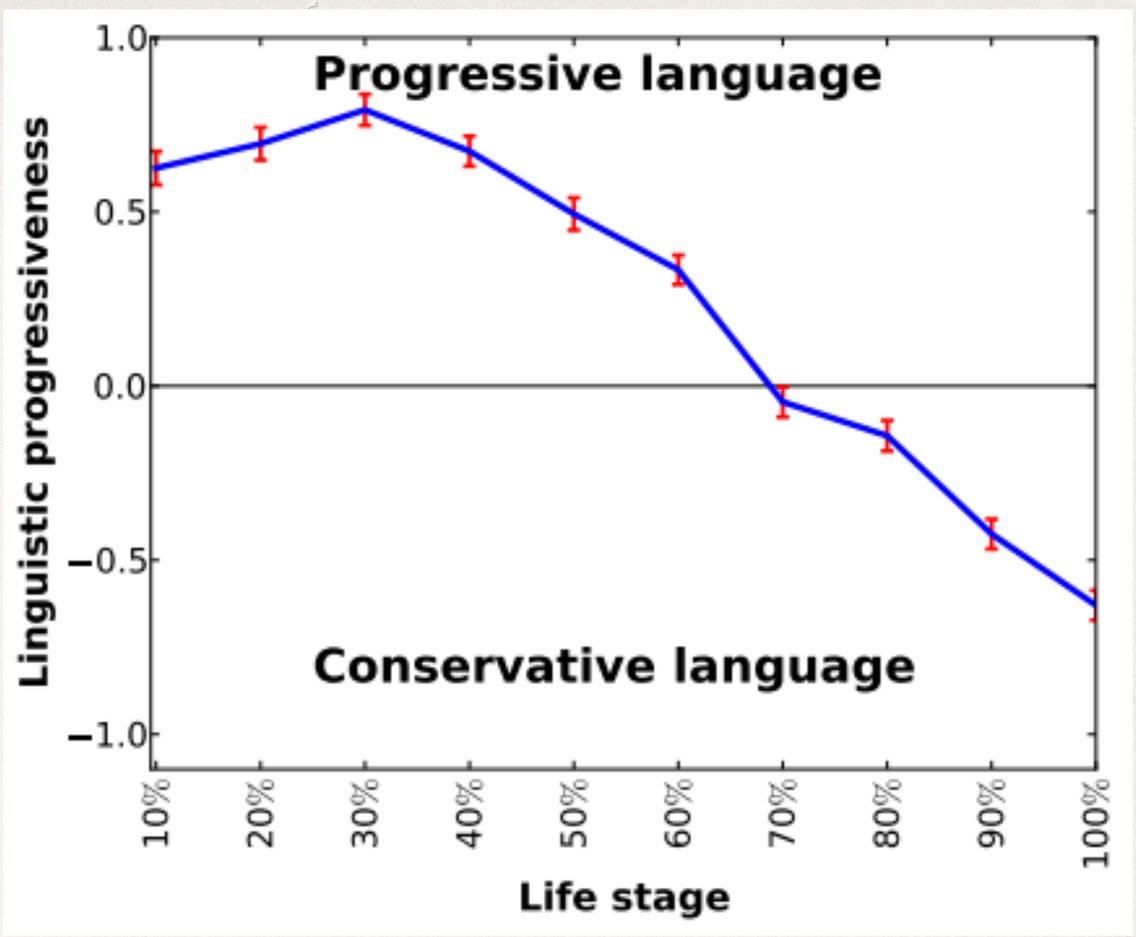


(a) BeerAdvocate



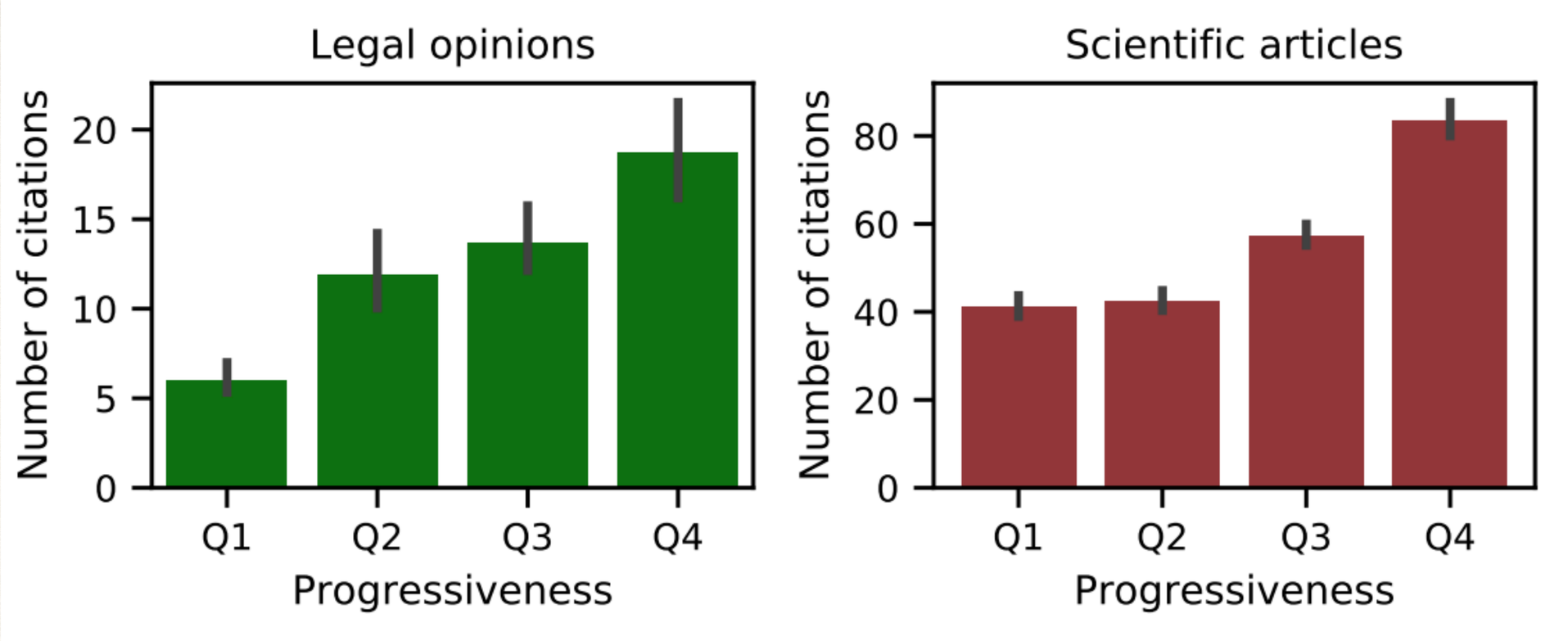
(b) RateBeer





- Users conform to the community's language initially but then stop adapting after some point

- Users are more innovative and trend-setting with their language use initially but then stabilize to rely increasingly more on past language



- A legal opinion or a scientific article is progressive if it uses words with meanings from the future than the past
- Learn two skipgram language models, one trained on future data and one trained on the past; then compare word senses with respect to both the models

Soni, Sandeep, Kristina Lerman, and Jacob Eisenstein. "Follow the leader: Documents on the leading edge of semantic change get more citations." *Journal of the Association for Information Science and Technology* 72.4 (2021): 478-492.

Can we do better than N-gram language models?

Language modeling

- Instead of modeling P(x), why not model P(w|c)?
- Rather than directly estimating the word probabilities from relative frequencies, this hints at language modeling as a learning task

Language modeling

Context	Word
May the Force be with	____

- Instead of modeling P(x), why not model P(w|c)?
- Rather than directly estimating the word probabilities from relative frequencies, this hints at language modeling as a learning task

Language modeling

Context | Word

May the Force be with | ____

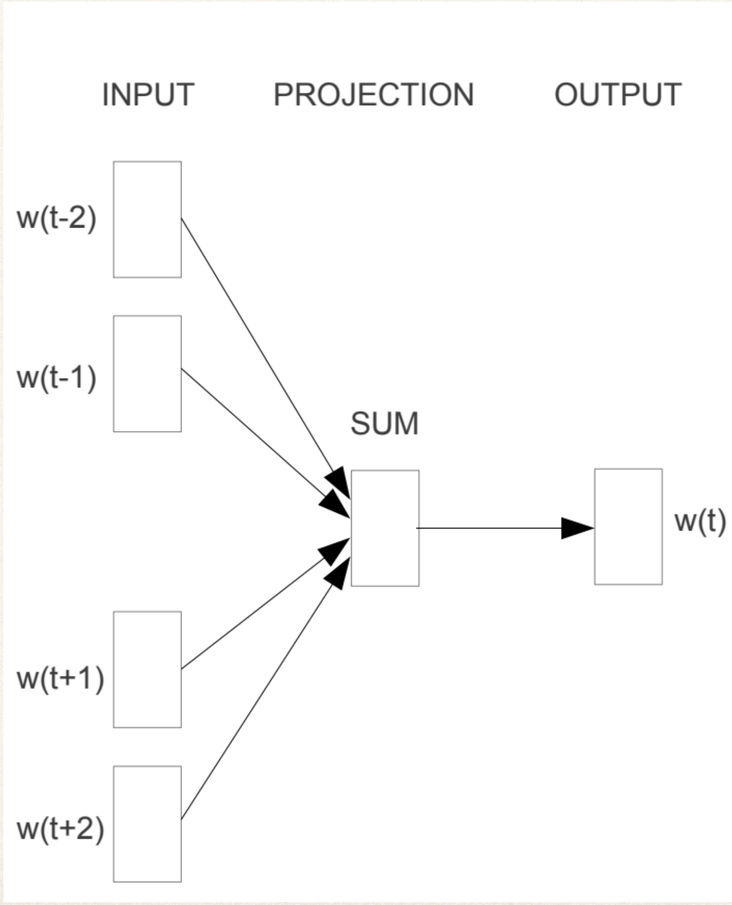
- Reparametrize the probability P(w|c) to depend on dense vectors
- $P(w \mid c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$, where β_w is a vector representation of w and v_c is the vector representation of the context

Word2Vec (CBoW)

- $$P(w \mid c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$$

- In CBoW model of word2vec, w is a word and c are words on the left and right of w

- v_c was calculated as a sum of output vectors



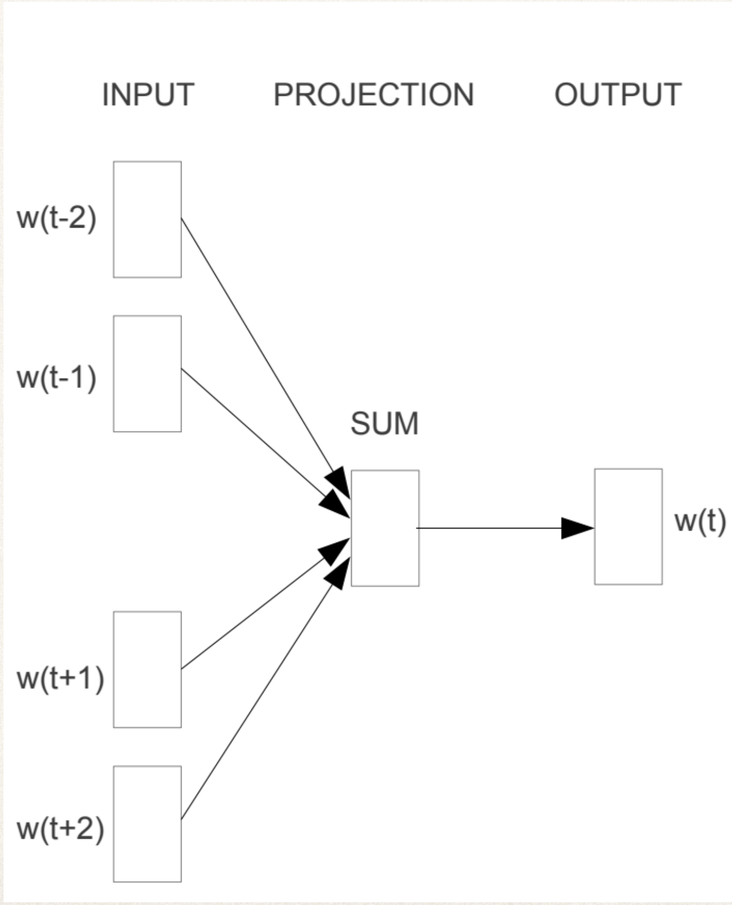
Mikolov et. al. 2013

Word2Vec (CBoW)

- $$P(w \mid c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$$

- In CBoW model of word2vec, w is a word and c are words on the left and right of w

- v_c was calculated as a sum of output vectors



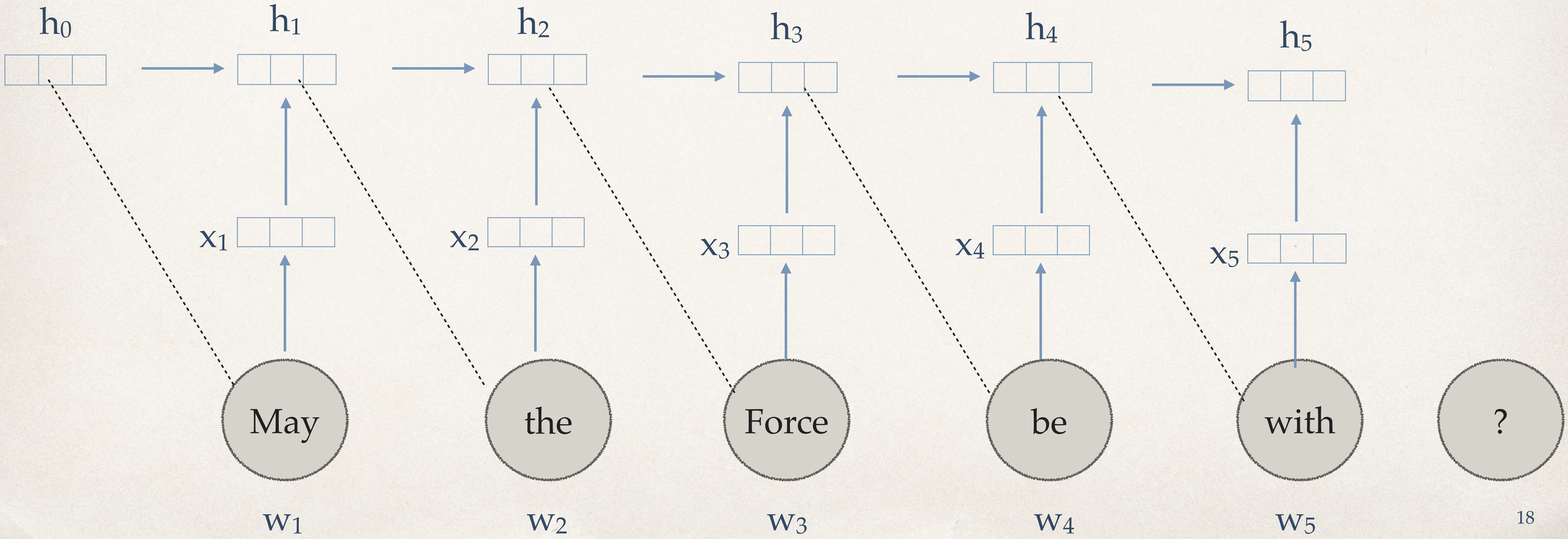
Mikolov et. al. 2013

What can be better ways of coming up with a vector representation of the context?

Context

Word

May the Force be with ____



Recurrent Neural Network LM

At every position m:

$$x_m = \text{Lookup}(\phi, w_m)$$

$$h_m = \text{RNN}(x_m, h_{m-1})$$

$$\boxed{h_m = g(\Theta h_{m-1} + x_m)} \quad \text{Elman unit}$$

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

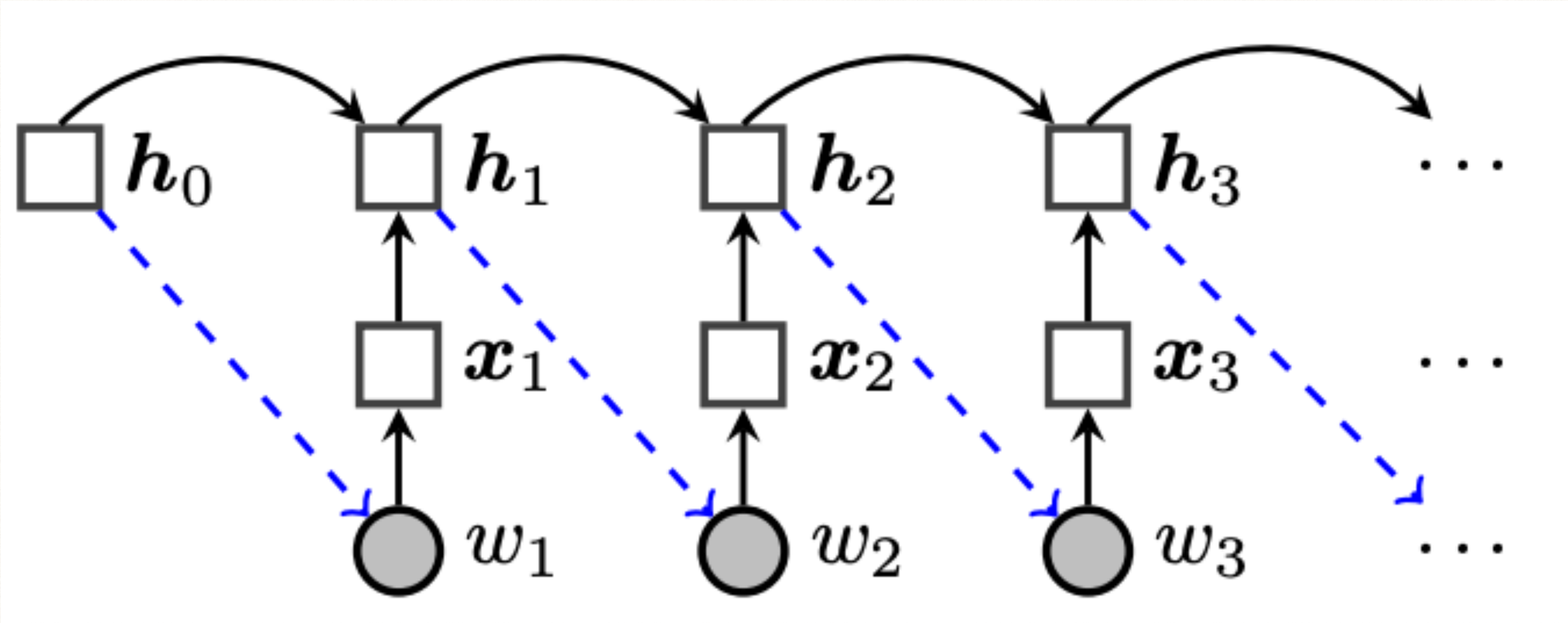
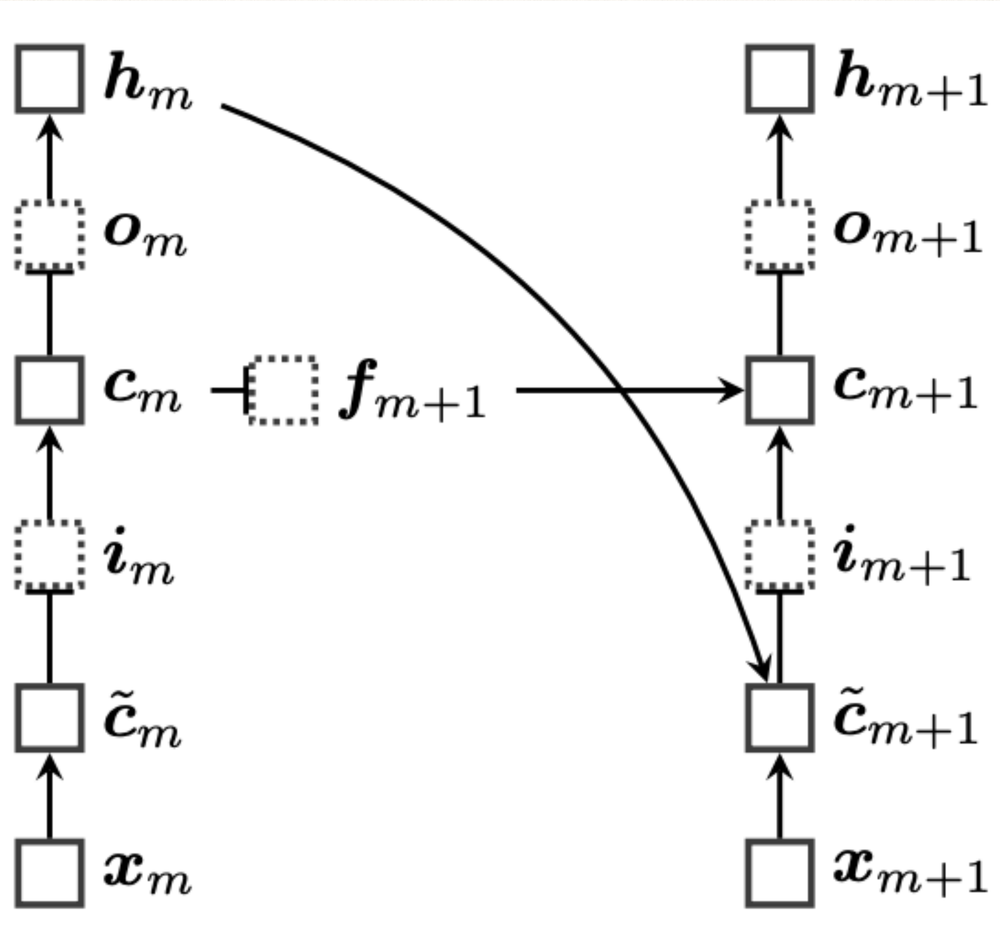


Figure taken from Eisenstein 2018

Long short-term Memories (LSTM)



- We can define the transformation from x to h in a more involved way by adding gating units
- This has the effect of preserving information propagation over long distances but also down weighting non-important contexts in the past

Can we do better?

- Which part of the context is more important or one we should attend?

Attention

- We can calculate a score to each part of a sequence by learning some parameters

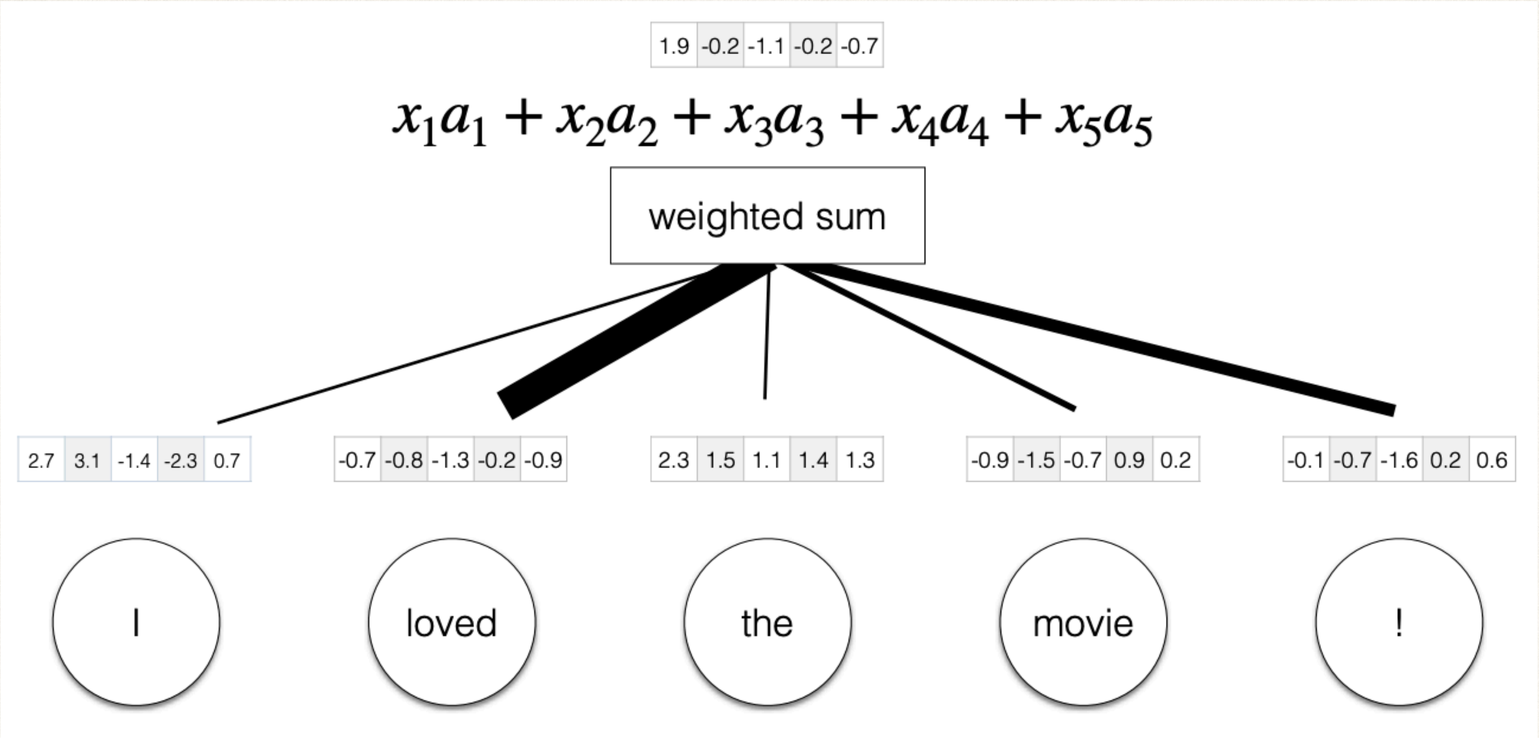


Figure taken from David Bamman's class slides

$$r_i = v^\top x_i$$

$$a = \text{softmax}(r)$$

Self Attention

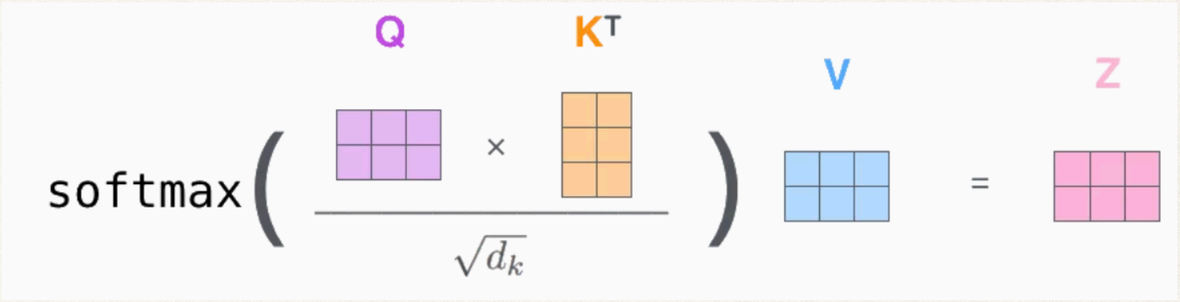
- Given a sequence find importance of every word over every other word in the sequence

Self Attention

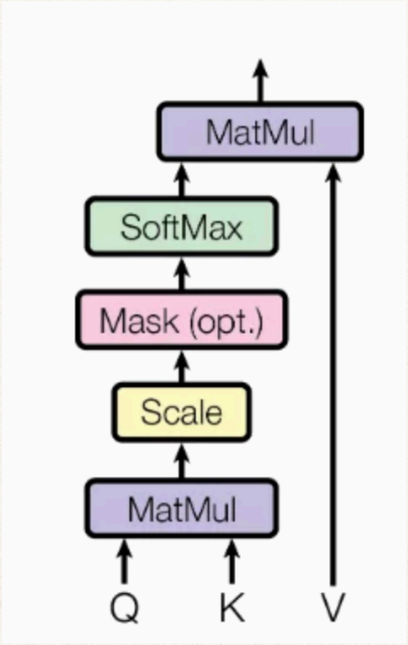
query $Q = XW^Q$

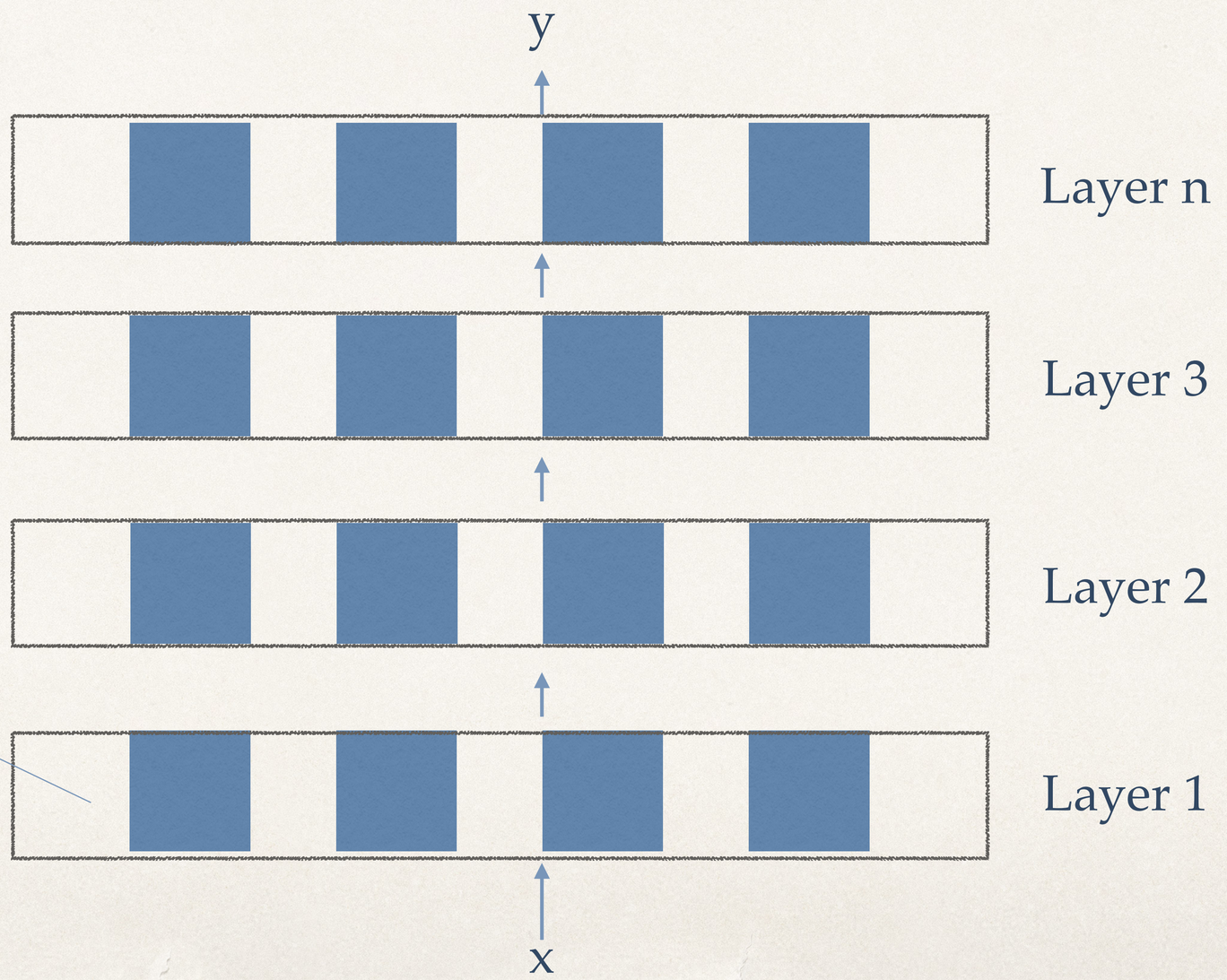
key $K = XW^K$

value $V = XW^V$



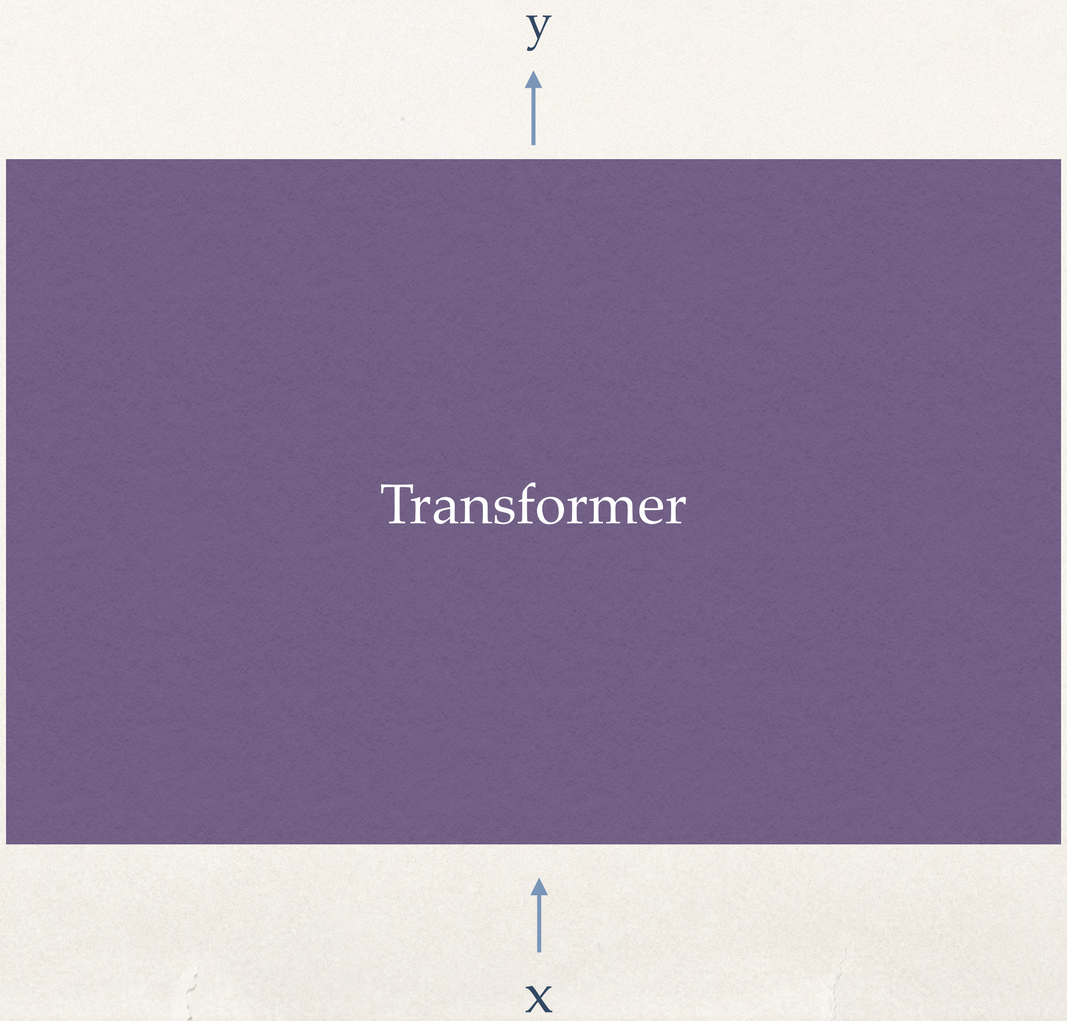
Single head attention





Self attention units
within an attention block

Transformers



LARGE LANGUAGE MODELS

- Most modern LLMs are based on the transformer architecture (e.g., BERT, GPT, etc)
- Many layers, high dimensional vector representations, and large corpora for pretraining are all hallmarks of contemporary LLMs