NLU lecture 5: Word representations and morphology

Adam Lopez
alopez@inf.ed.ac.uk

- Essential epistemology
- Word representations and word2vec
- Word representations and compositional morphology

Reading: Mikolov et al. 2013, Luong et al. 2013

Essential epistemology

	Exact sciences	Empirical sciences	Engineering
Deals with	Axioms & theorems	Facts & theories	Artifacts
Truth is	Forever	Temporary	It works
Examples	Mathematics C.S. theory F.L. theory	Physics Biology **Linguistics**	Many, including applied C.S. e.g. **NLP**

Essential epistemology

Exact sciences	Empirical sciences	Engineering

Essential epistemology

Exact sciences	Empirical sciences	Engineering
	morphological properties of words (facts)	

Essential epistemology

Exact sciences	Empirical sciences	Engineering
	morphological properties of words (facts)	
	Optimality theory	

Essential epistemology

Exact sciences	Empirical sciences	Engineering
Optimality theory is finite-state	morphological properties of words (facts)	
	Optimality theory	

Essential epistemology

Exact sciences	Empirical sciences	Engineering
Optimality theory is finite-state	morphological properties of words (facts) Optimality theory	We can represent morphological properties of words with finite-state automata

The Bandwagon

CLAUDE E. SHANNON

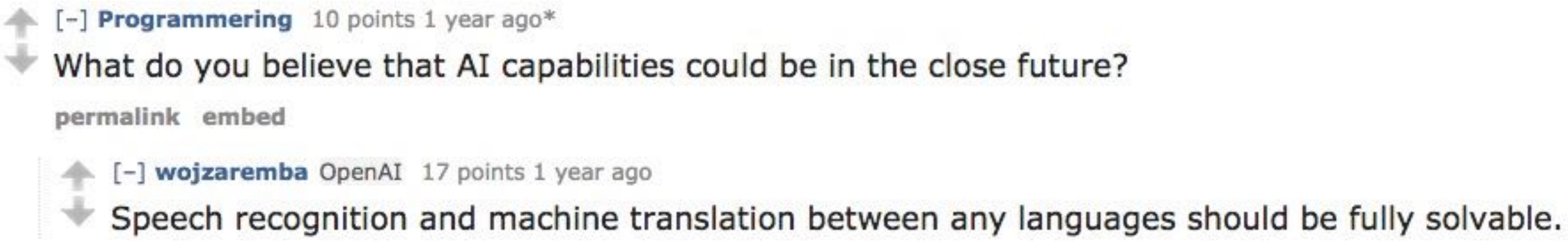
INFORMATION theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, it has perhaps been ballooned to an importance beyond its actual accomplishments. Our fellow scientists in many different fields, attracted by the fanfare and by the new avenues opened to scientific analysis, are using these ideas in their own problems. Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. In short, information theory is currently partaking of a somewhat heady draught of general popularity.

Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. While we feel that information theory is indeed a valuable tool in providing fundamental insights into the nature of communication problems and will continue to grow in importance, it is certainly no panacea for the communication engineer or, *a fortiori*, for anyone else. Seldom do more than a few of nature's secrets give way at one time. It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information*, *entropy*, *redundancy*, do not solve all our problems.

What can be done to inject a note of moderation in this situation? In the first place, workers in other fields should realize that the basic results of the subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications. I personally believe that many of the concepts of information theory will prove useful in these other fields—and, indeed, some results are already quite promising—but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification. If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations.

Secondly, we must keep our own house in first class order. The subject of information theory has certainly been sold, if not oversold. We should now turn our attention to the business of research and development at the highest scientific plane we can maintain. Research rather than exposition is the keynote, and our critical thresholds should be raised. Authors should submit only their best efforts, and these only after careful criticism by themselves and their colleagues. A few first rate research papers are preferable to a large number that are poorly conceived or half-finished. The latter are no credit to their writers and a waste of time to their readers. Only by maintaining a thoroughly scientific attitude can we achieve real progress in communication theory and consolidate our present position.

Remember the bandwagon



[–] **Programmering** 10 points 1 year ago*

What do you believe that AI capabilities could be in the close future?

permalink embed

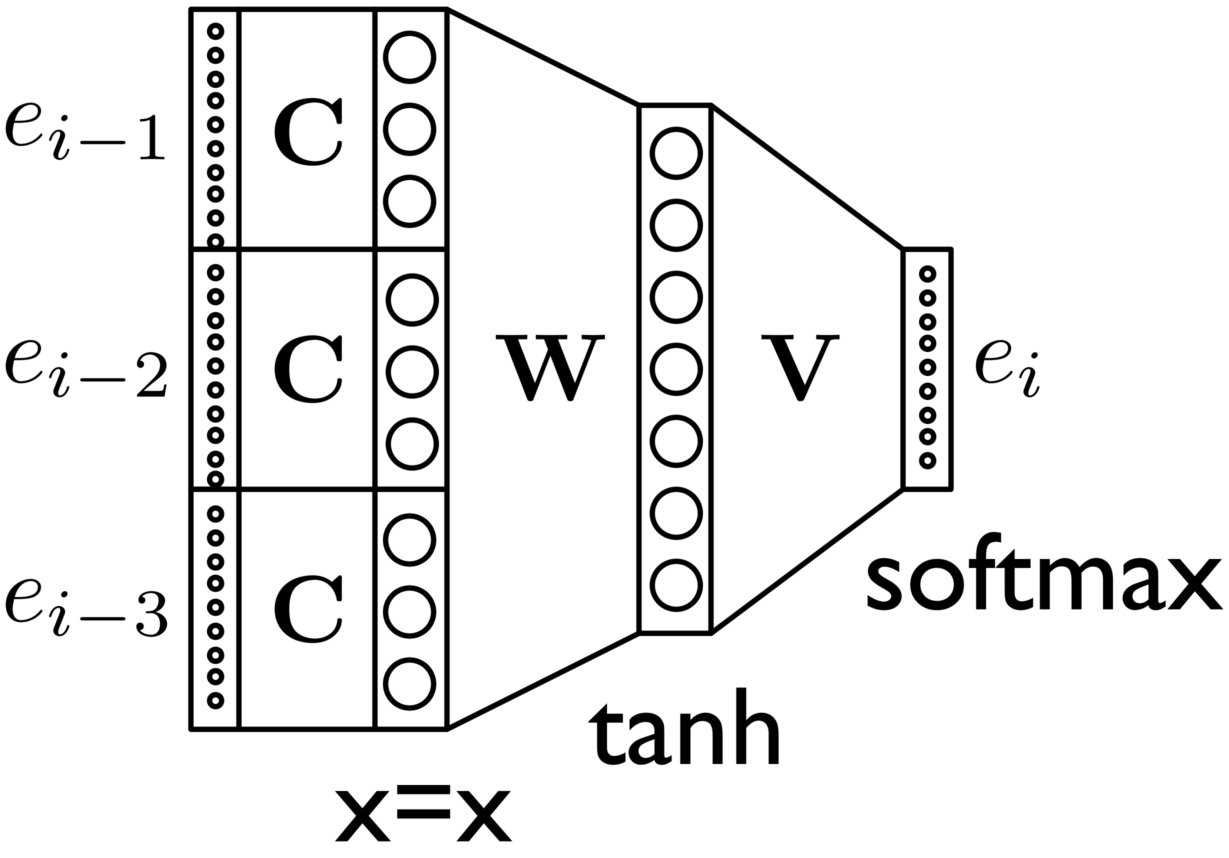
> [–] **wojzaremba** OpenAI 17 points 1 year ago
>
> Speech recognition and machine translation between any languages should be fully solvable.

Word representations

Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

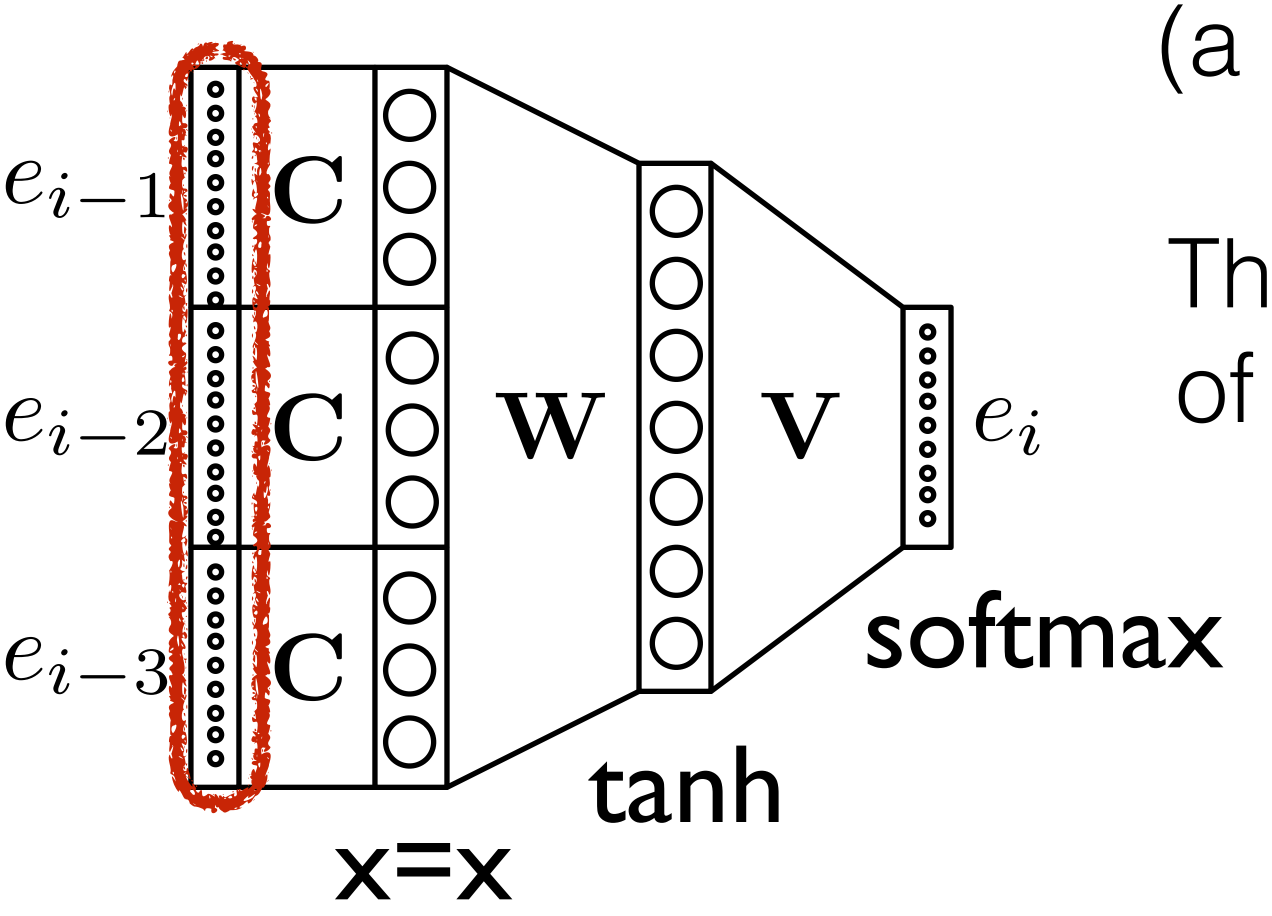
$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$



Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$



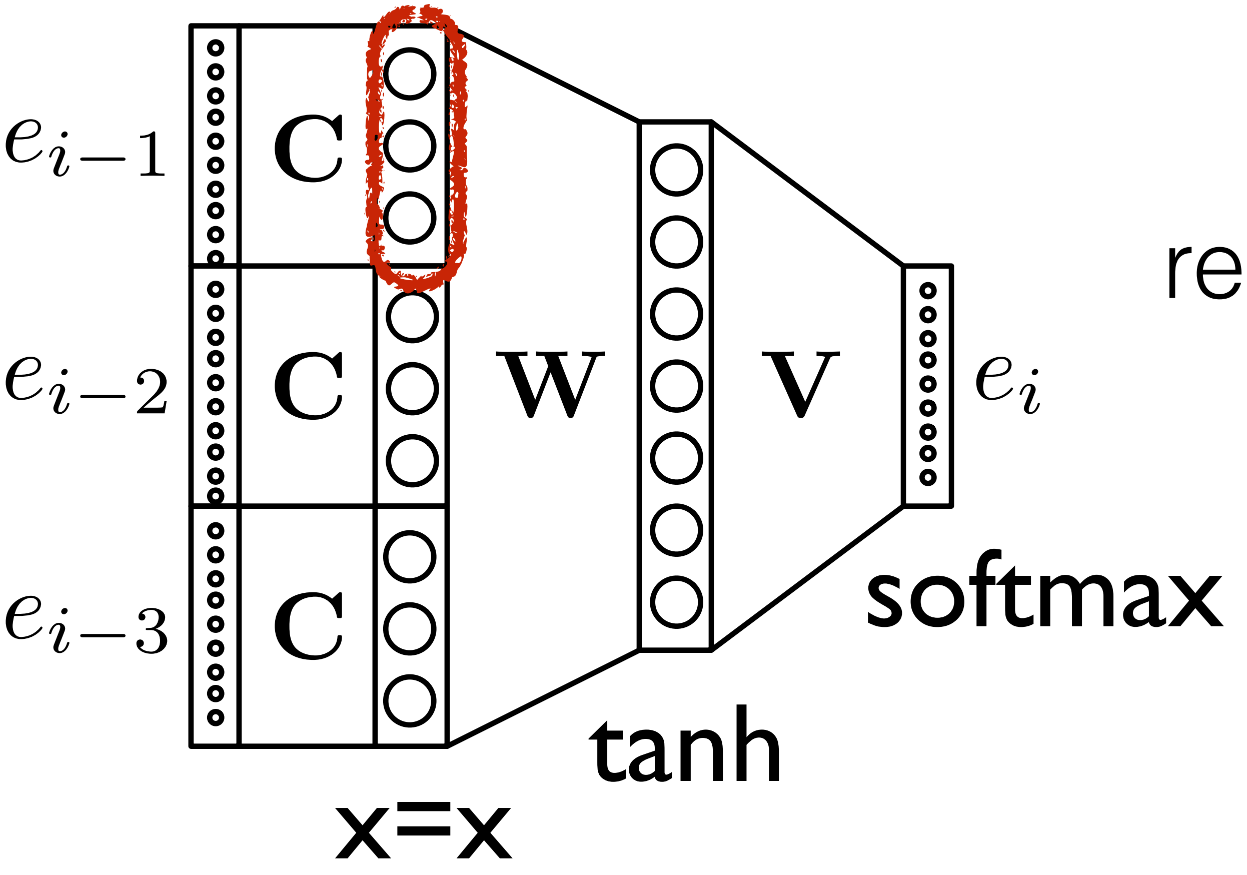
Every word is a vector (a one-hot vector)

The concatenation of these vectors is an *n*-gram

Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$

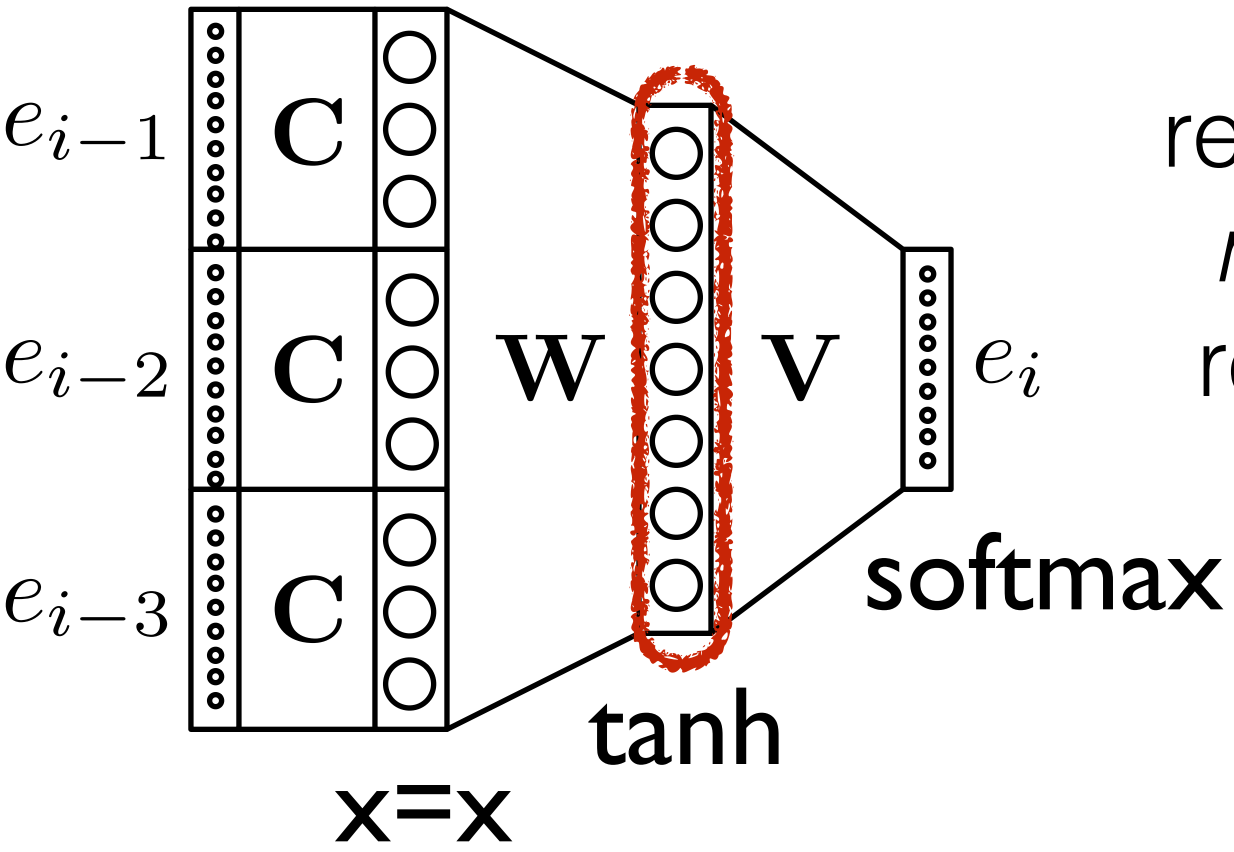


Word embeddings are vectors: continuous representations of each word.

Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$

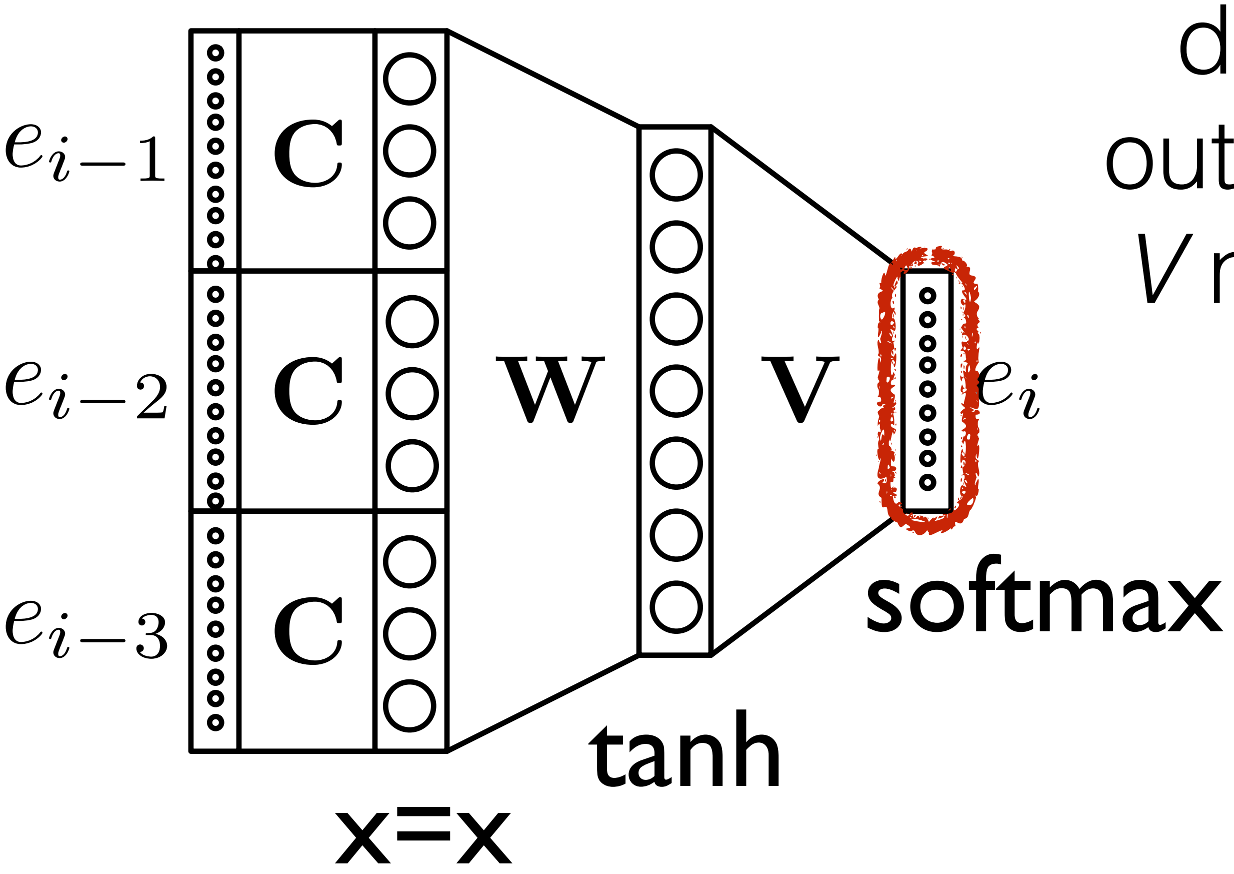


n-grams are vectors: continuous representations of *n*-grams (or, via recursion, larger structures)

Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$

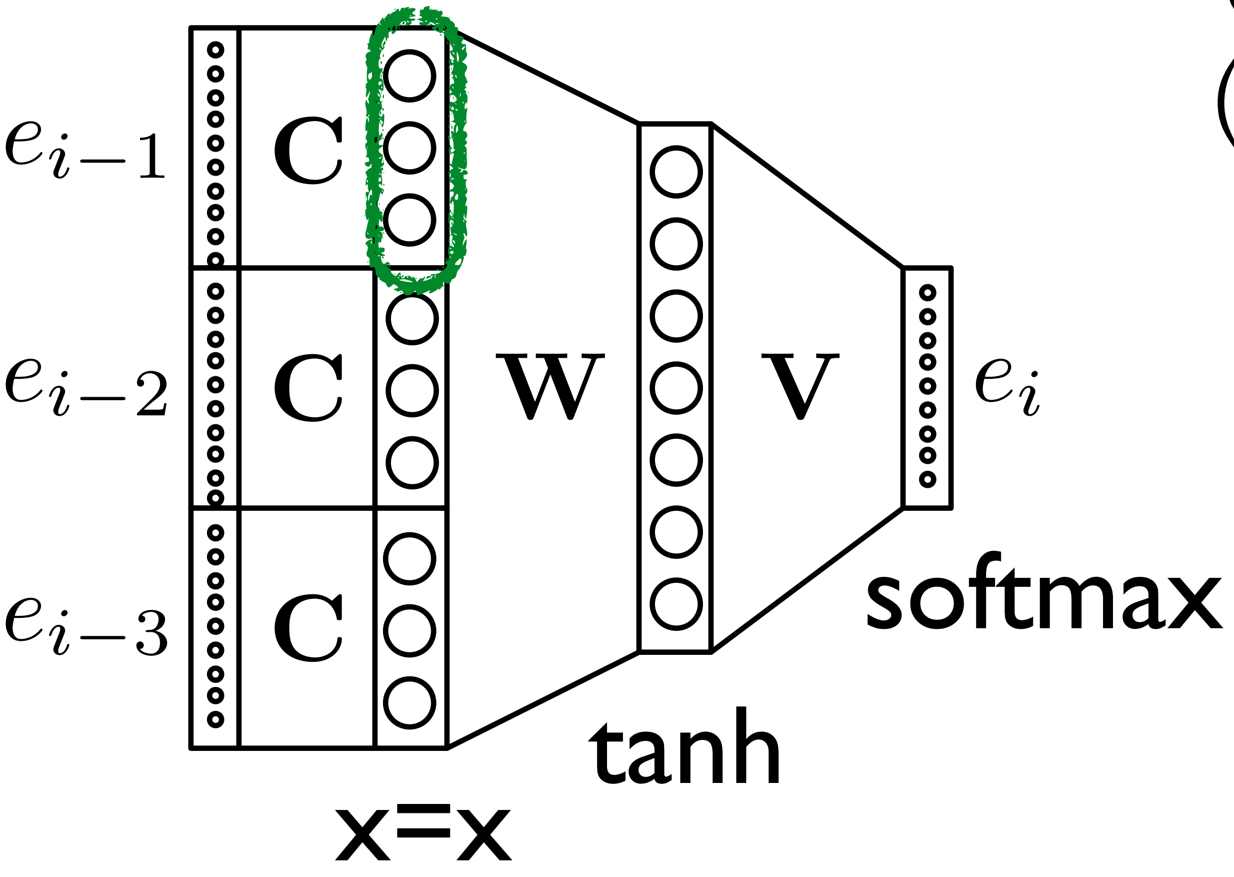


a discrete probability distribution over *V* outcomes is a vector: *V* non-negative reals summing to 1.

Feedforward model

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} p(e_i \mid e_{i-n+1}, \ldots, e_{i-1})$$

$$p(e_i \mid e_{i-n+1}, \ldots, e_{i-1}) =$$



No matter what we do in NLP, we'll (almost) always have words… Can we reuse these vectors?

Design a POS tagger using an RRNLM

Design a POS tagger using an RRNLM

What are some difficulties with this?

What limitation do you have in learning a POS tagger that you don’t have when learning a LM?

Design a POS tagger using an RRNLM

What are some difficulties with this?

What limitation do you have in learning a POS tagger that you don't have when learning a LM?

One big problem:
LIMITED DATA

“You shall know a word by the company it keeps”

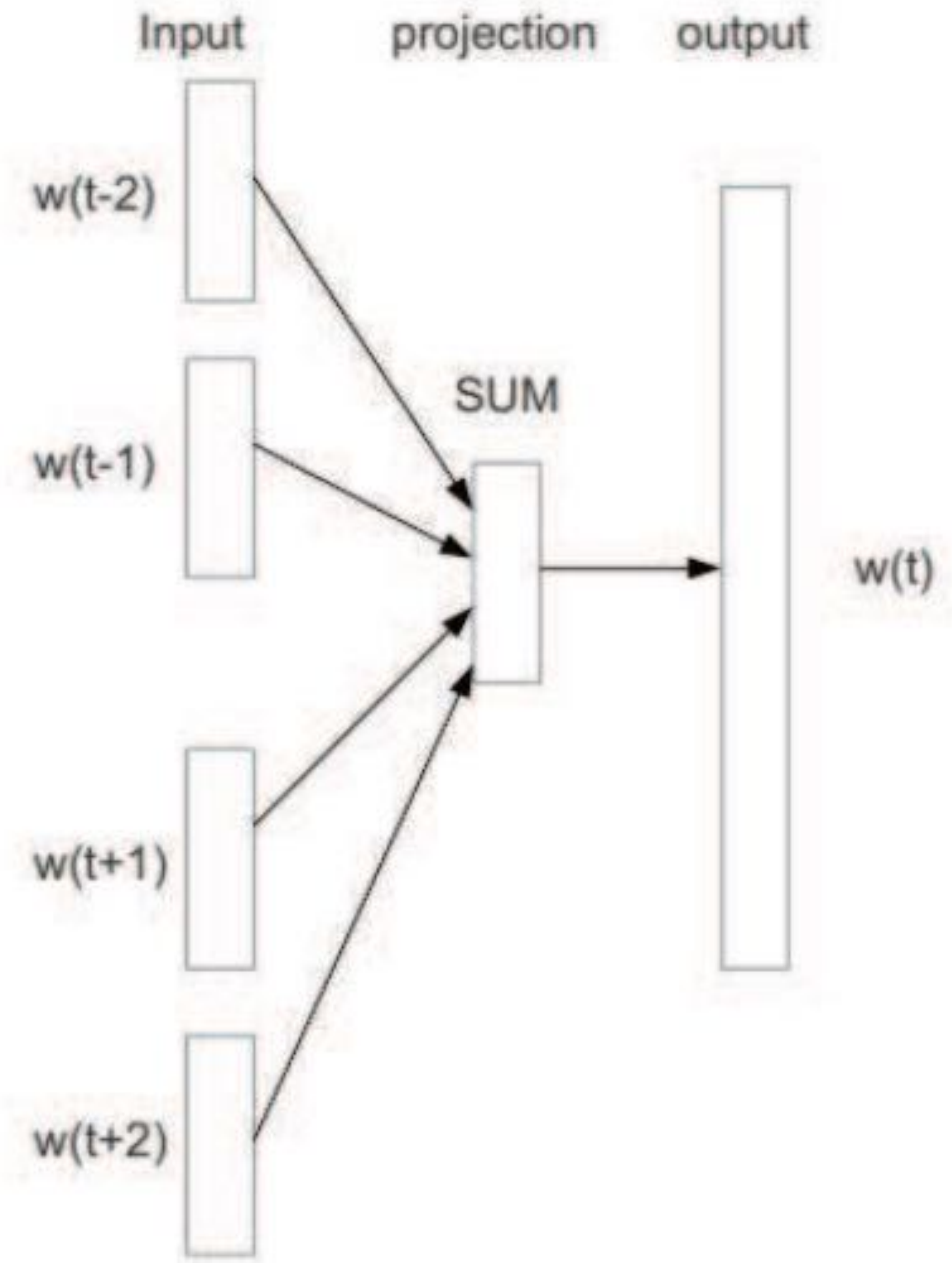
–John Rupert Firth (1957)

Learning word representations using language modeling

- Idea: we'll learn word representations using a language model, then reuse them in our POS tagger (or any other thing we predict from words).

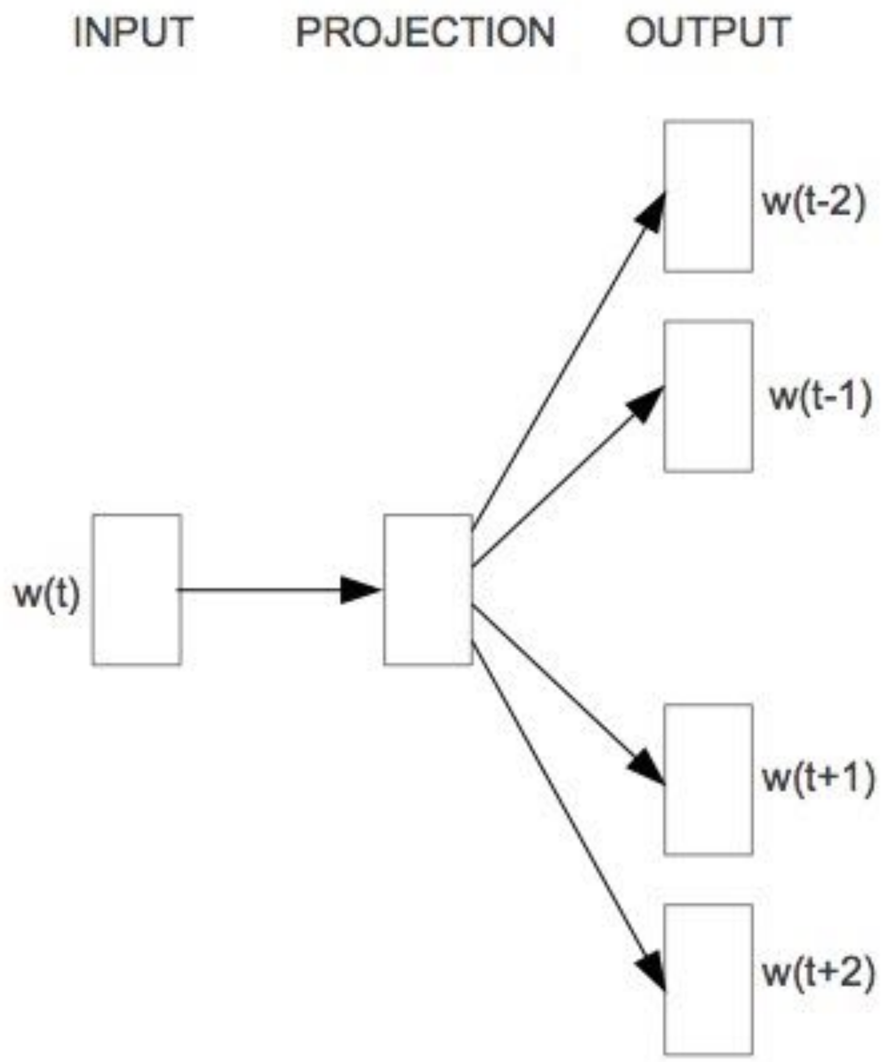
- Problem: Bengio language model is slow. Imagine computing a softmax over 10,000 words!

Continuous bag-of-words (CBOW)



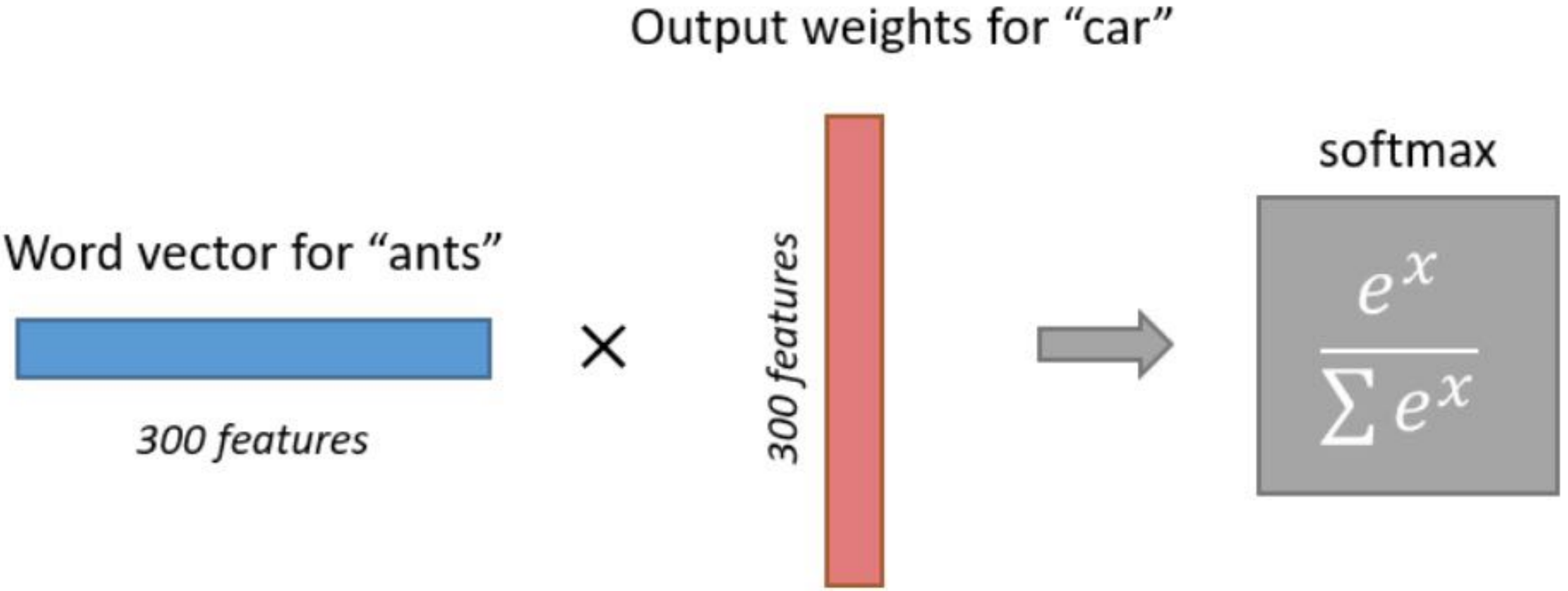
- Mikolov et al. (2013, ICLR)
- CBOW adds inputs from words within short window to predict the current word
- The weights for different positions are shared
- Computationally much more efficient than normal NNLM
- The hidden layer is just linear

Skip-gram



- We can reformulate the CBOW model by predicting surrounding words using the current word
- Find word representations useful for predicting surrounding words in a sentence or document.

Skip-gram



Output weights for “car”

softmax

Word vector for “ants”

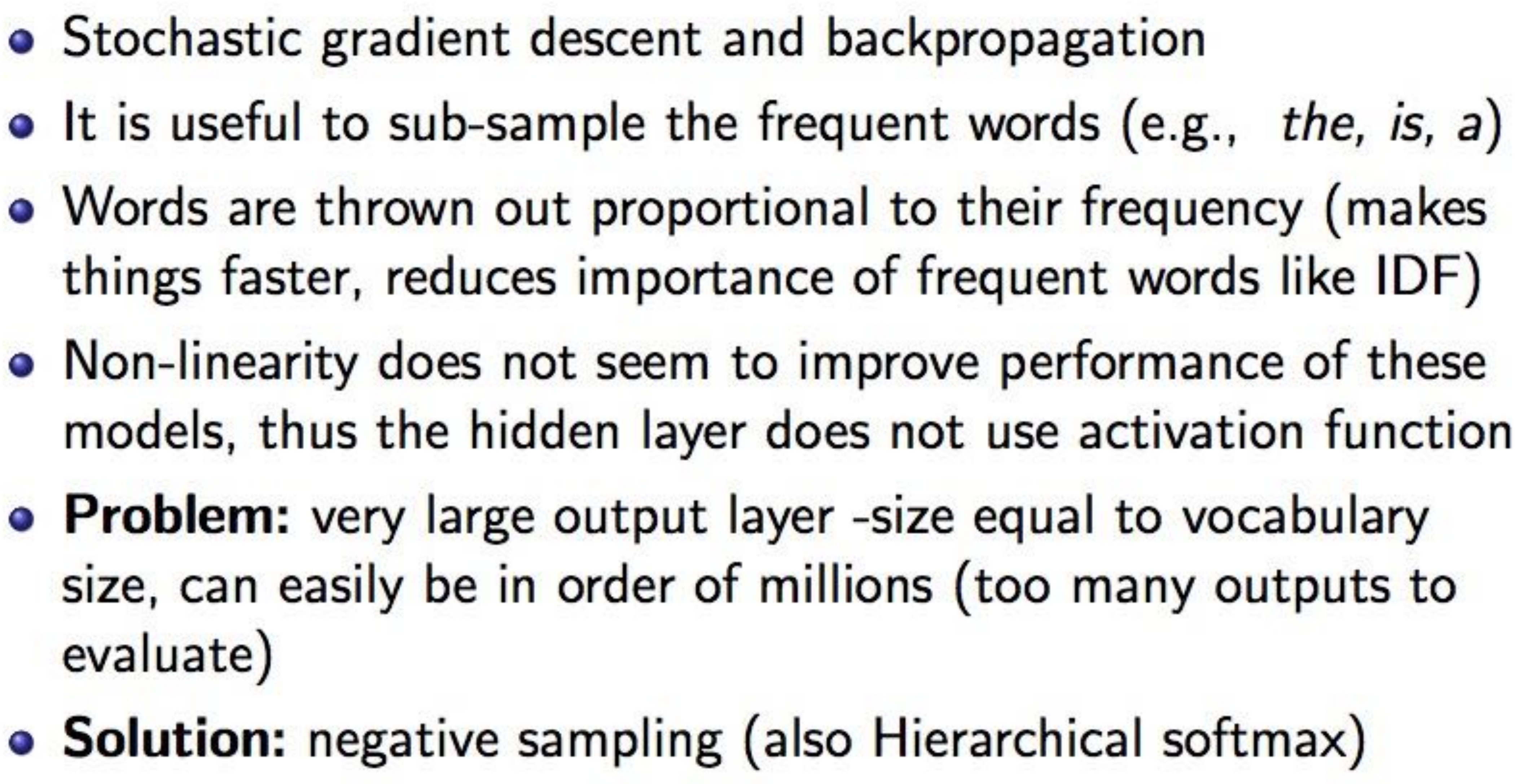
×

300 features

300 features

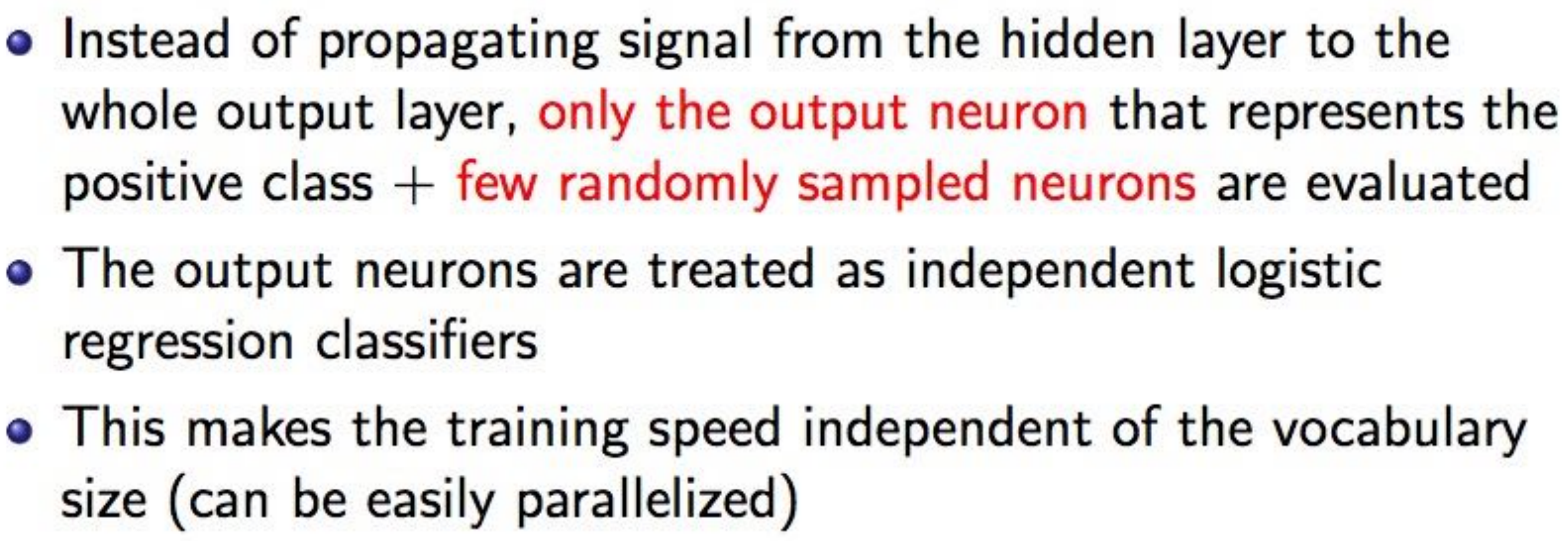
$$\frac{e^x}{\sum e^x}$$

Learning skip-gram



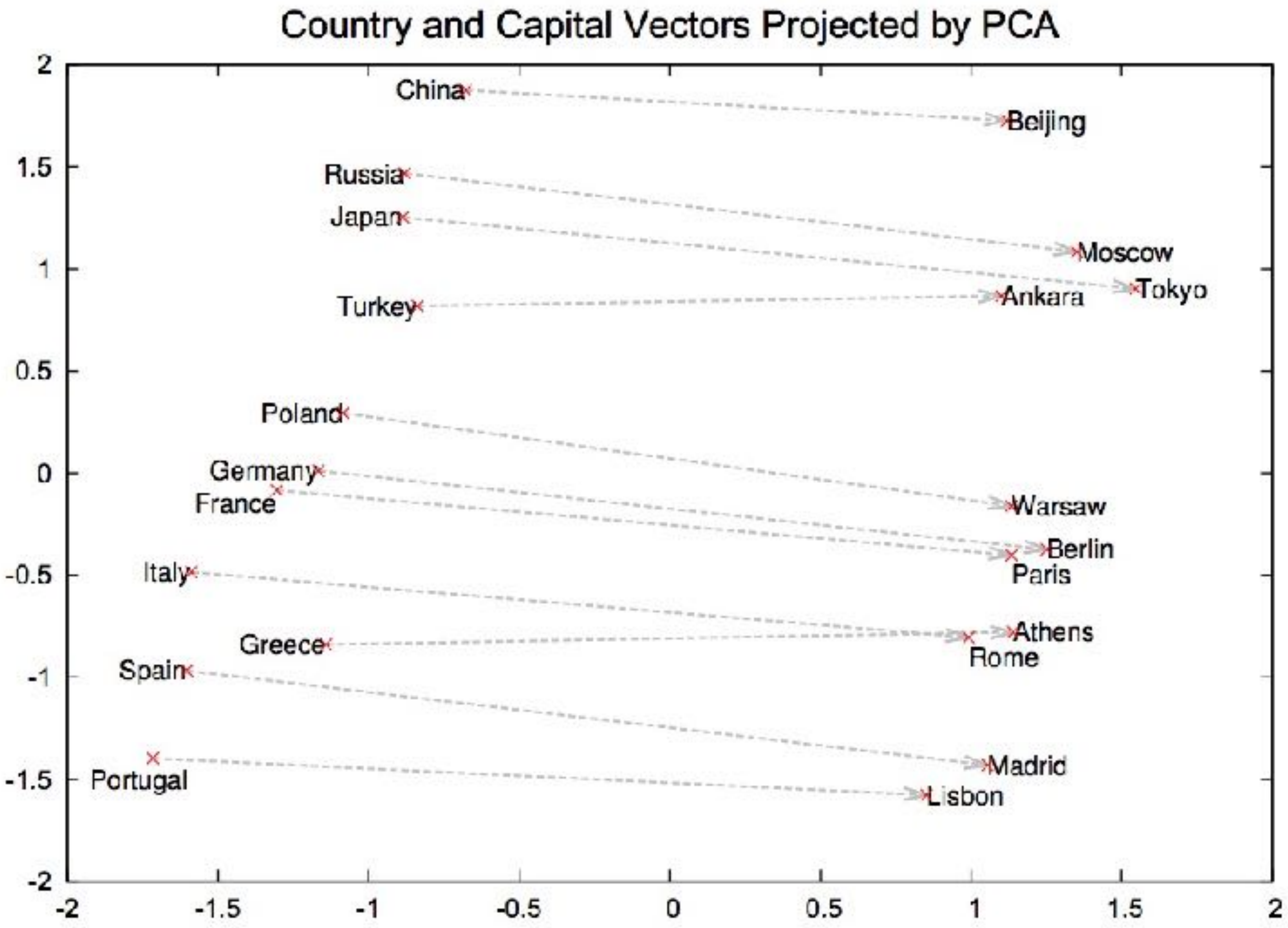
- Stochastic gradient descent and backpropagation
- It is useful to sub-sample the frequent words (e.g., *the, is, a*)
- Words are thrown out proportional to their frequency (makes things faster, reduces importance of frequent words like IDF)
- Non-linearity does not seem to improve performance of these models, thus the hidden layer does not use activation function
- **Problem:** very large output layer -size equal to vocabulary size, can easily be in order of millions (too many outputs to evaluate)
- **Solution:** negative sampling (also Hierarchical softmax)

Learning skip-gram

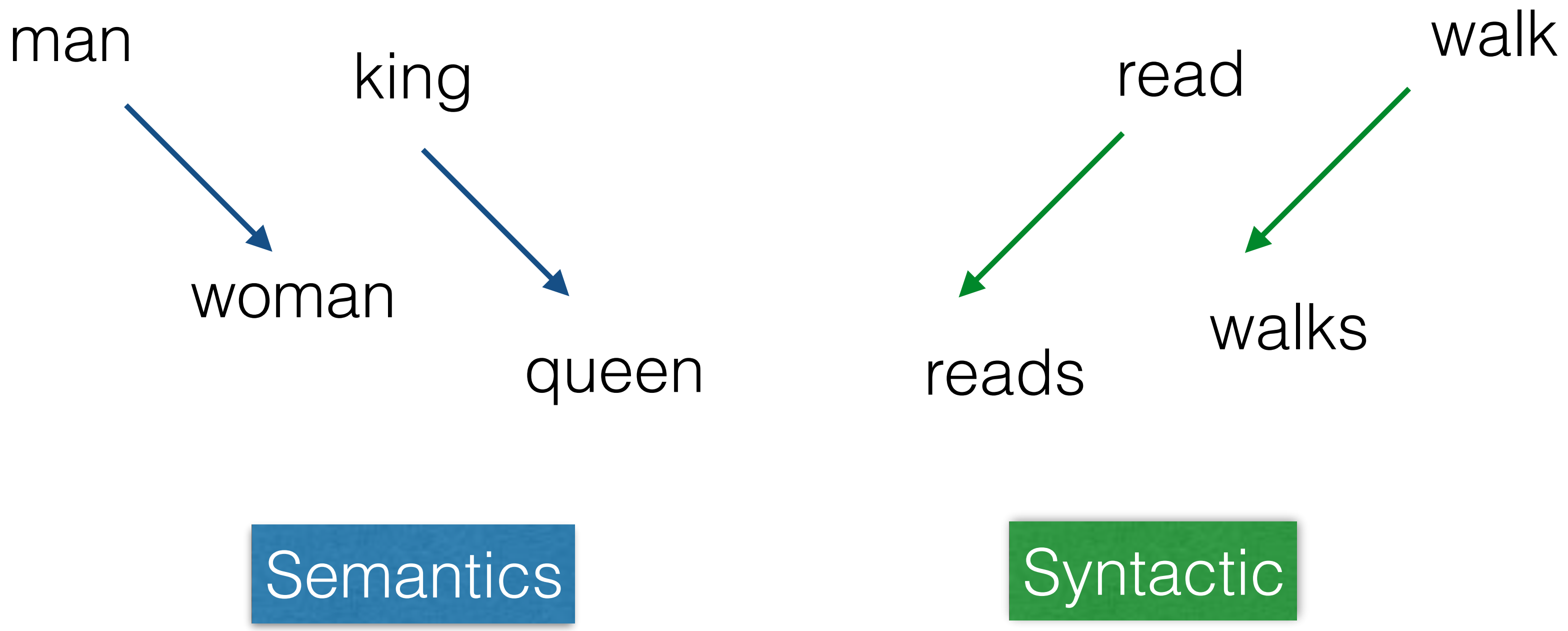


- Instead of propagating signal from the hidden layer to the whole output layer, only the output neuron that represents the positive class + few randomly sampled neurons are evaluated
- The output neurons are treated as independent logistic regression classifiers
- This makes the training speed independent of the vocabulary size (can be easily parallelized)

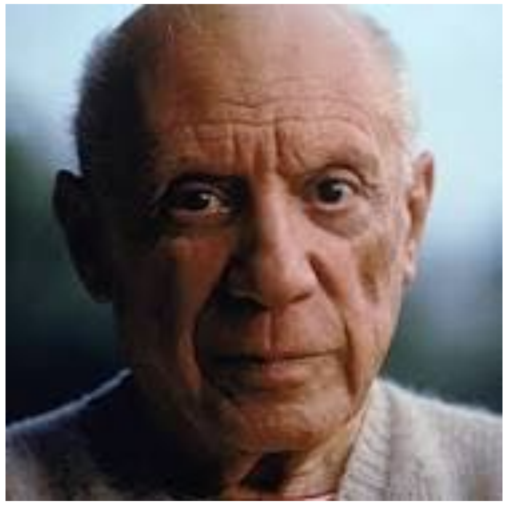
Word representations capture *some* world knowledge

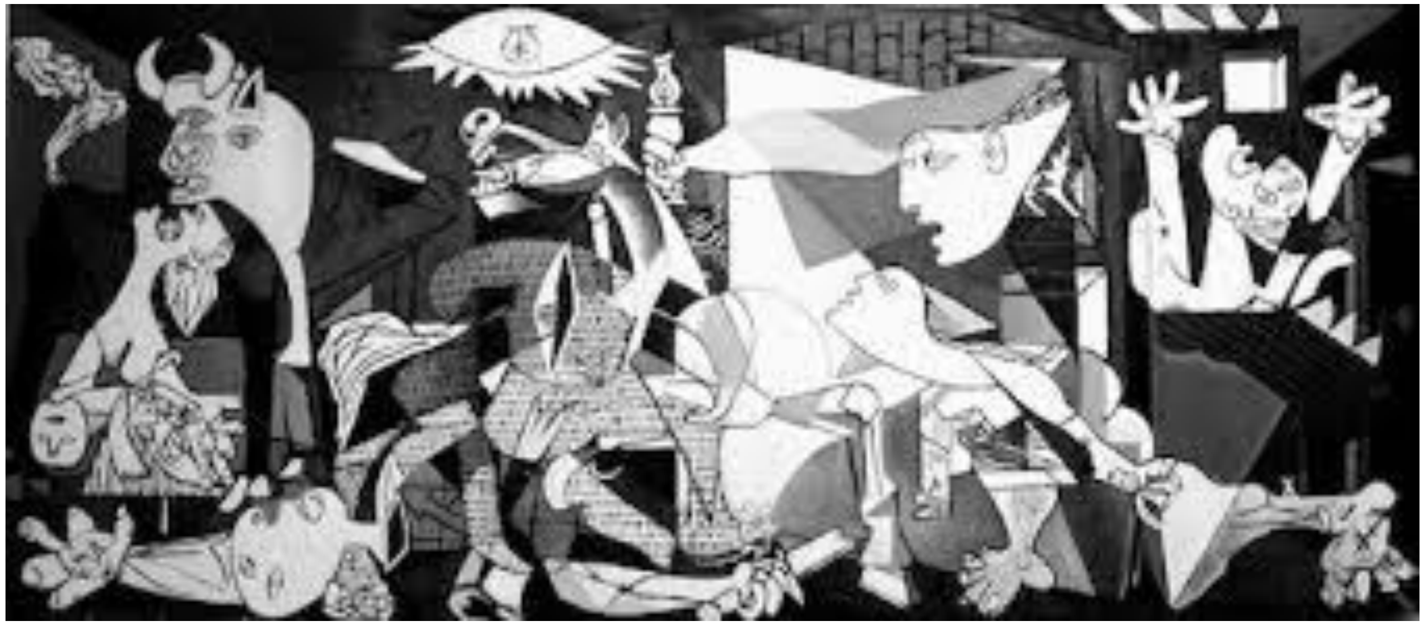


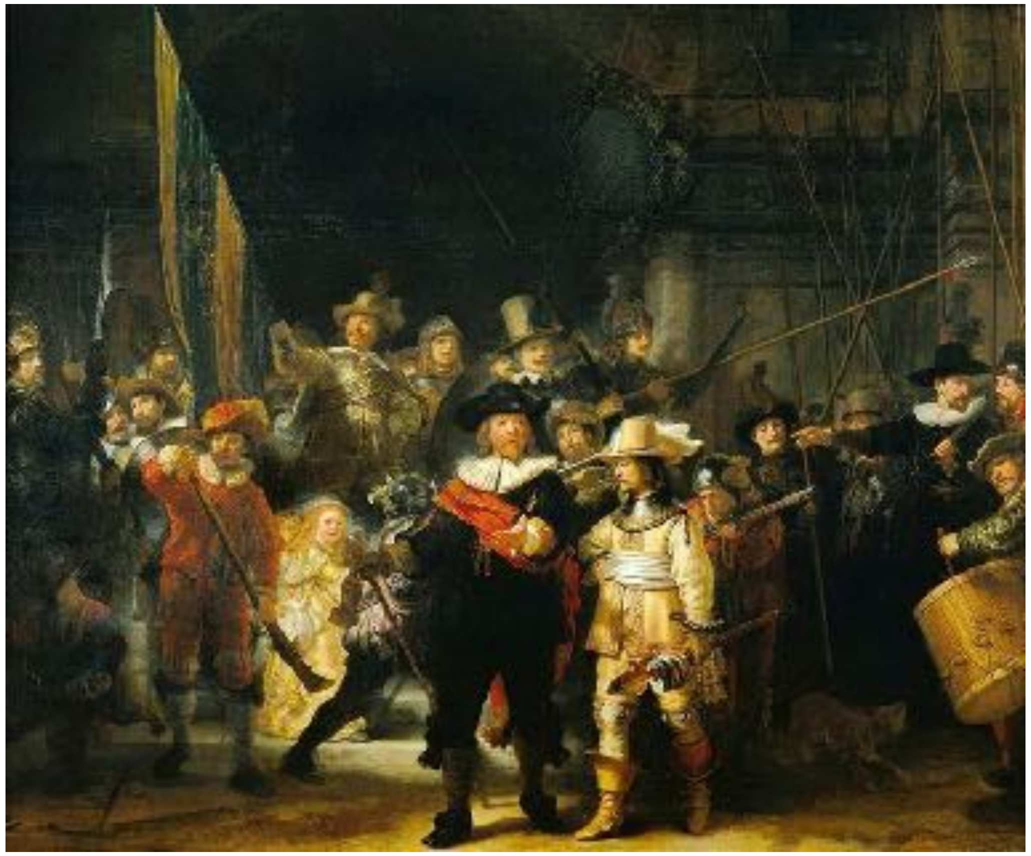
Continuous Word Representations



Will it learn this?



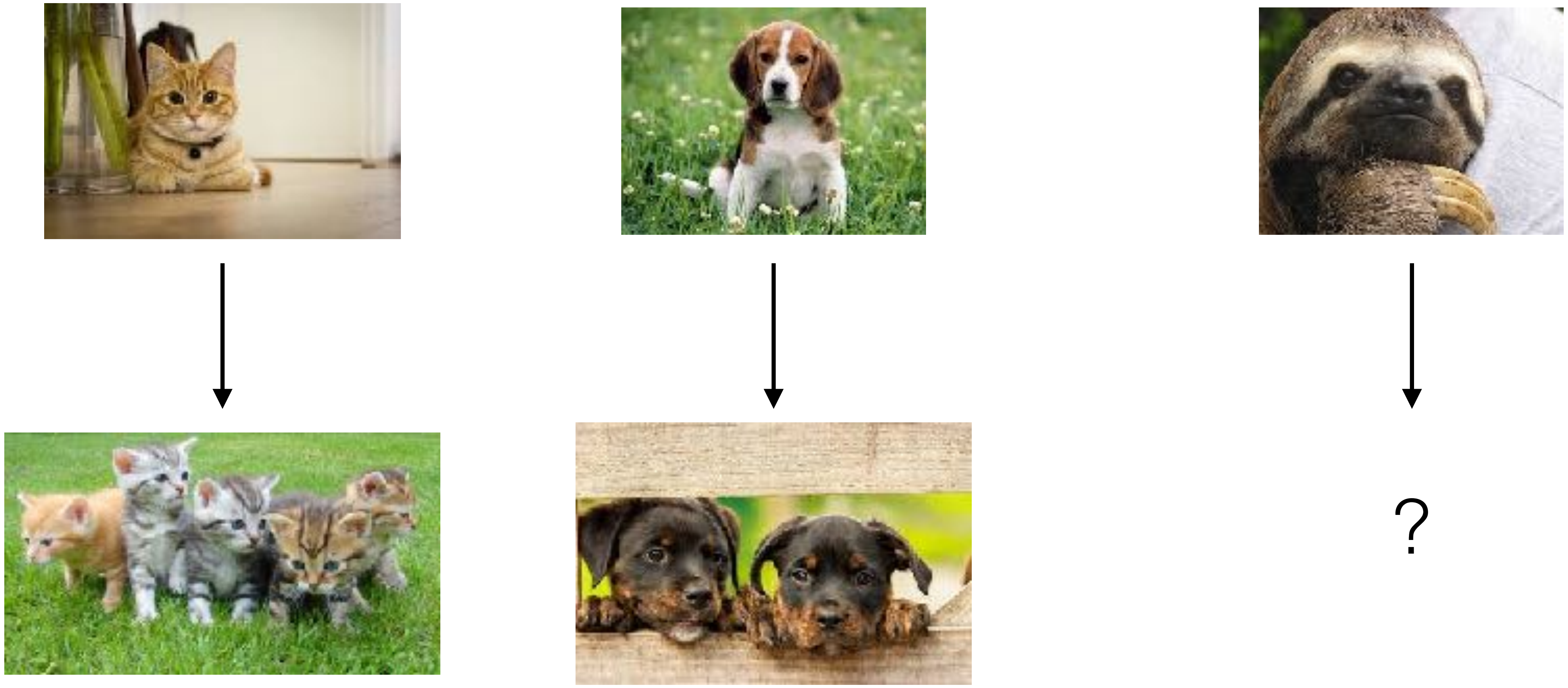




(Additional) limitations of word2vec

- Closed vocabulary assumption

- Cannot exploit **functional relationships** in learning



Is this language?

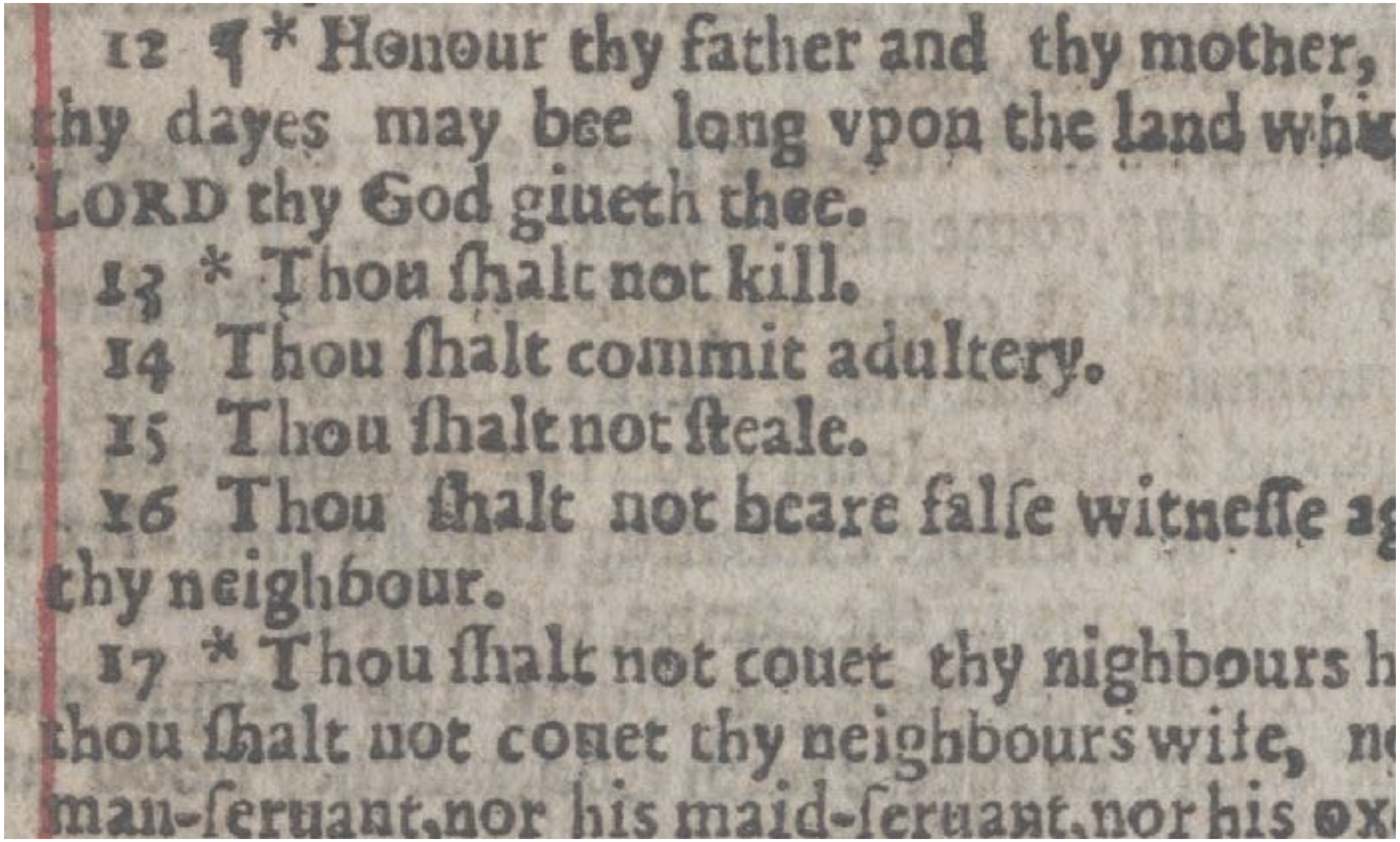
What our data contains:

A Lorillard spokeswoman said, “This is an old story.”

What word2vec thinks our data contains:

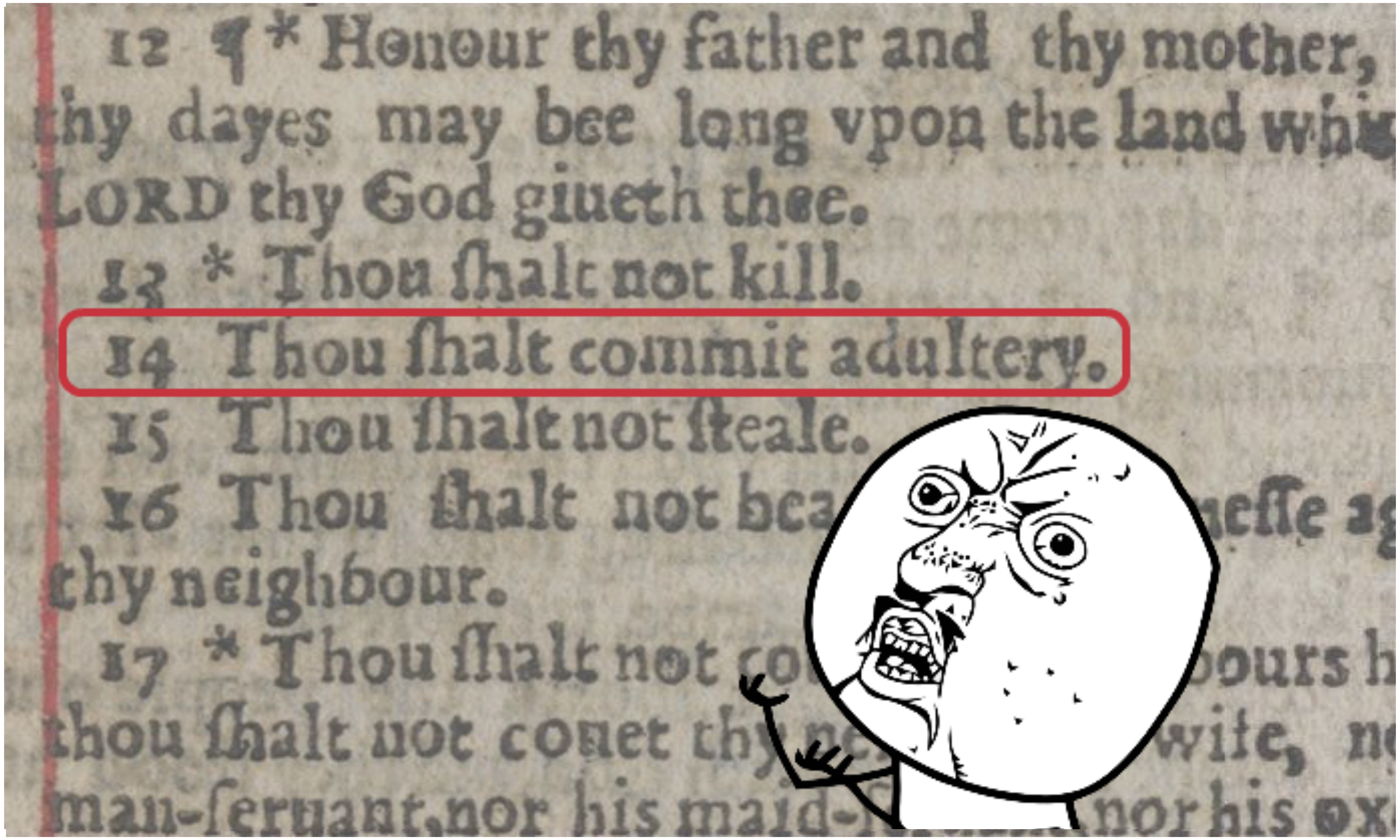
A UNK UNK said, “This is an old story.”

Is it ok to ignore words?



12 ¶ * Honour thy father and thy mother,
thy dayes may bee long vpon the land whi
LORD thy God giueth thee.
13 * Thou ſhalt not kill.
14 Thou ſhalt commit adultery.
15 Thou ſhalt not ſteale.
16 Thou ſhalt not beare falſe witneſſe ag
thy neighbour.
17 * Thou ſhalt not couet thy nighbours h
thou ſhalt not couet thy neighbours wife, n
man-ſeruant,nor his maid-ſeruant,nor his ox

Is it ok to ignore words?



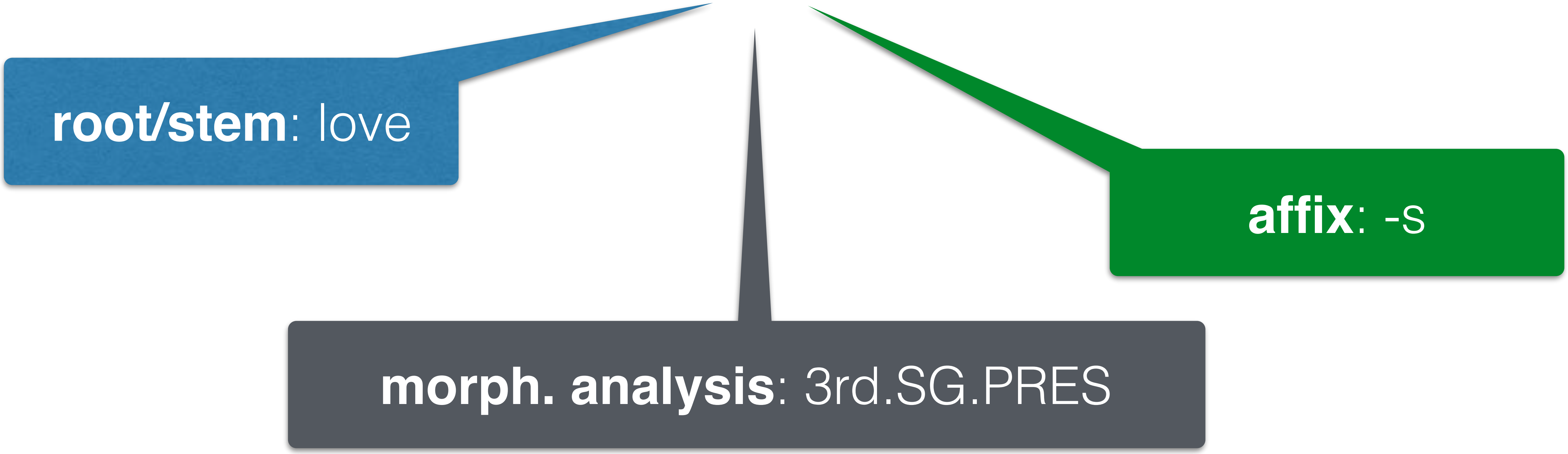
12 ¶ * Honour thy father and thy mother,
thy dayes may bee long vpon the land whi
LORD thy God giueth thee.
13 * Thou ſhalt not kill.
14 Thou ſhalt commit adultery.
15 Thou ſhalt not ſteale.
16 Thou ſhalt not bea neſſe ag
thy neighbour.
17 * Thou ſhalt not co bours h
thou ſhalt not couet thy ne wife, n
man-ſeruant, nor his maid- nor his ox

What **we** know about linguistic structure

Morpheme: the smallest meaningful unit of language

“loves”

love +s

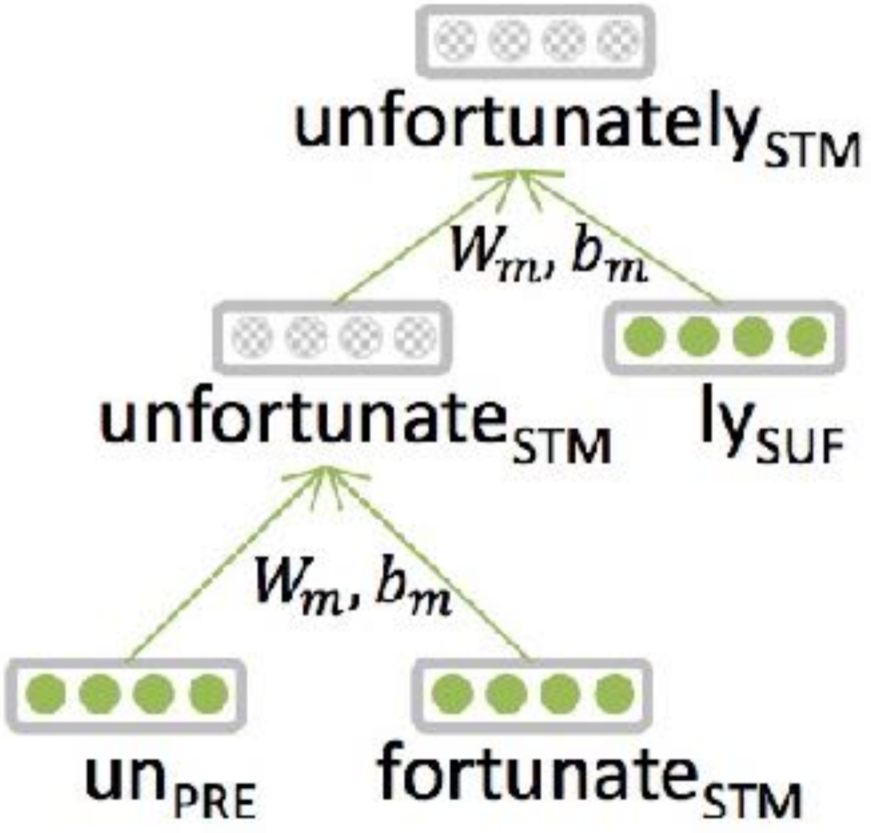


root/stem: love

affix: -s

morph. analysis: 3rd.SG.PRES

What if we embed morphemes rather than words?



Basic idea: compute representation recursively from children

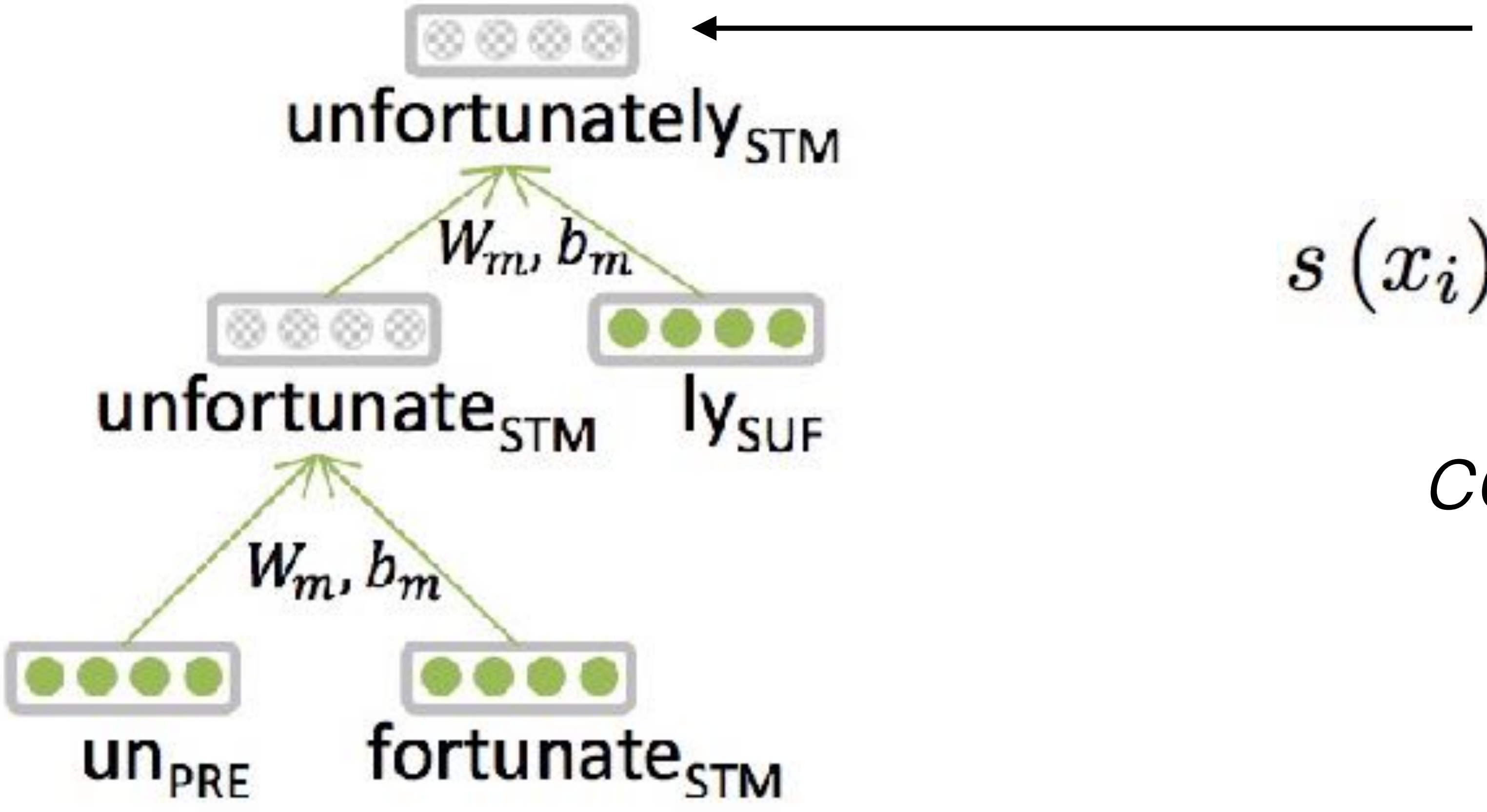
$$\boldsymbol{p} = f(\boldsymbol{W_m}[\boldsymbol{x}_{\texttt{stem}}; \boldsymbol{x}_{\texttt{affix}}] + \boldsymbol{b_m})$$

f is an activation function (e.g. tanh)

Vectors in green are morpheme embeddings (parameters)

Vectors in grey are computed as above (functions)

Train compositional morpheme model by minimizing distance to reference vector



Target output:
reference vector p_r

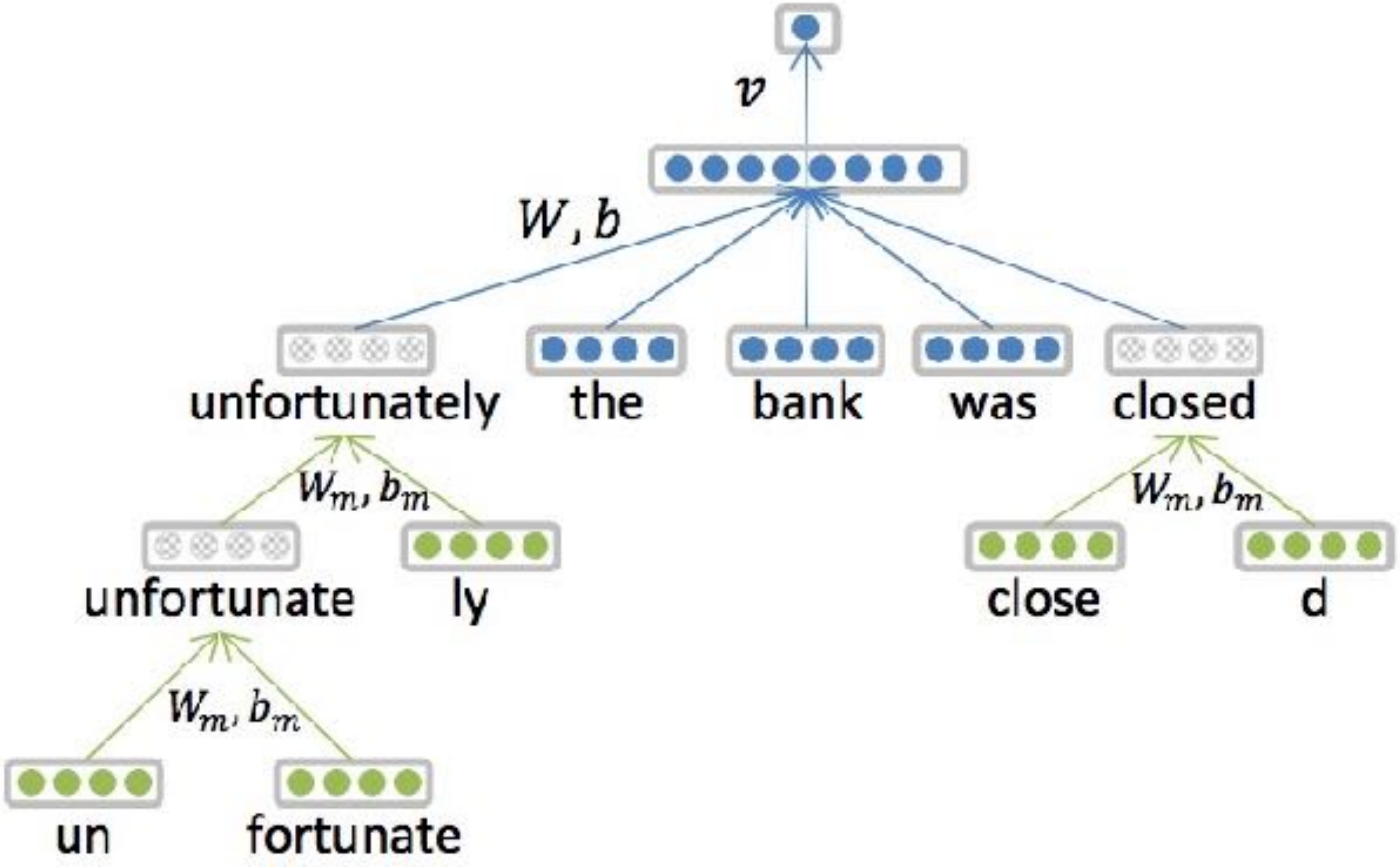
$$s\left(x_i\right) = \|\boldsymbol{p}_c(x_i) - \boldsymbol{p}_r(x_i)\|_2^2$$

contructed vector is p_c

Minimize:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} s\left(x_i\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

Or, train in context using backpropagation



(Basically a feedforward LM)

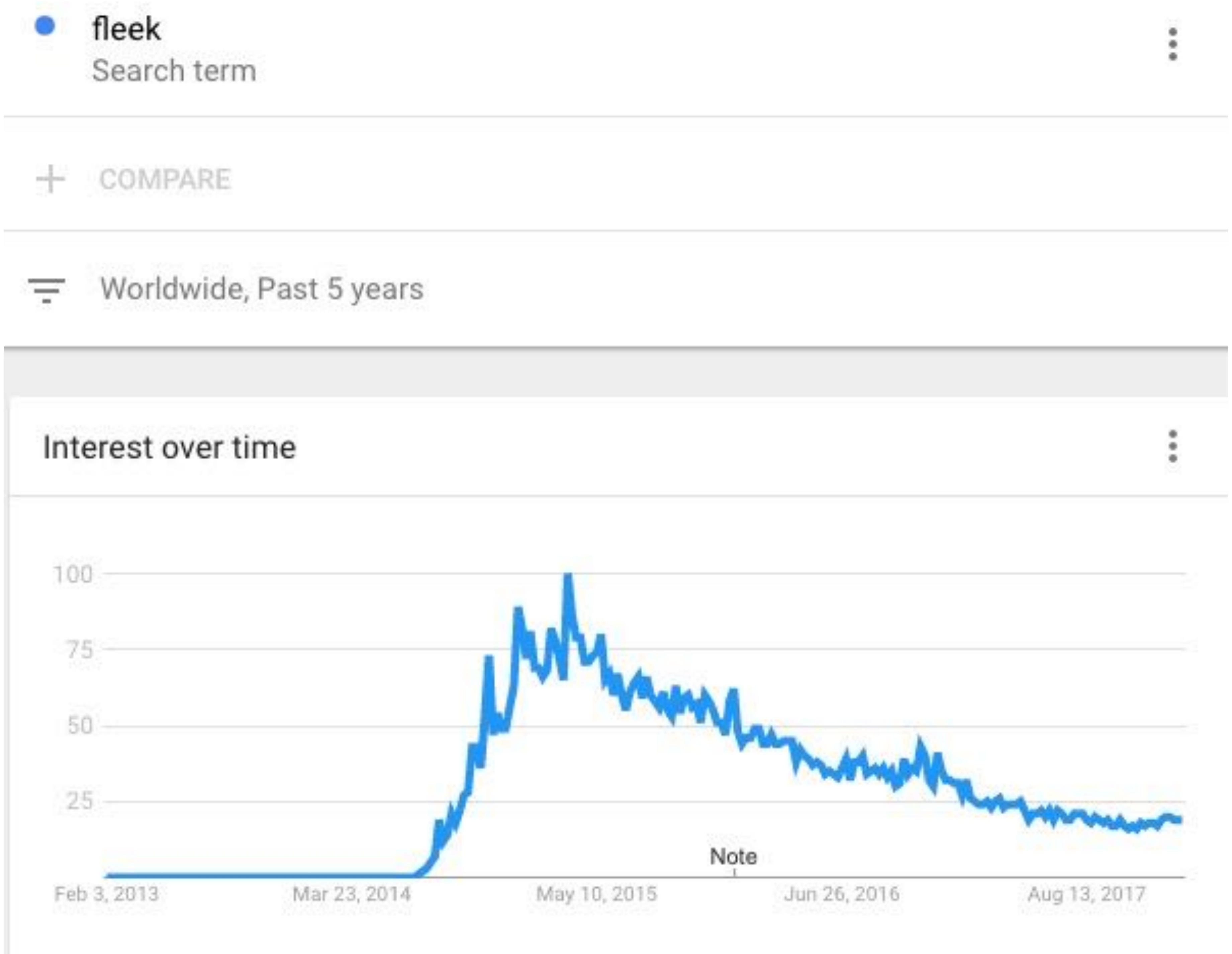
Vectors in blue are word or n-gram embeddings (parameters)
Vectors in green are morpheme embeddings (parameters)
Vectors in grey are computed as above (functions)

Where do we get morphemes?

- Use an unsupervised morphological analyzer (we’ll talk about unsupervised learning later on).

- How many morphemes are there?

New stems are invented every day!



fleek

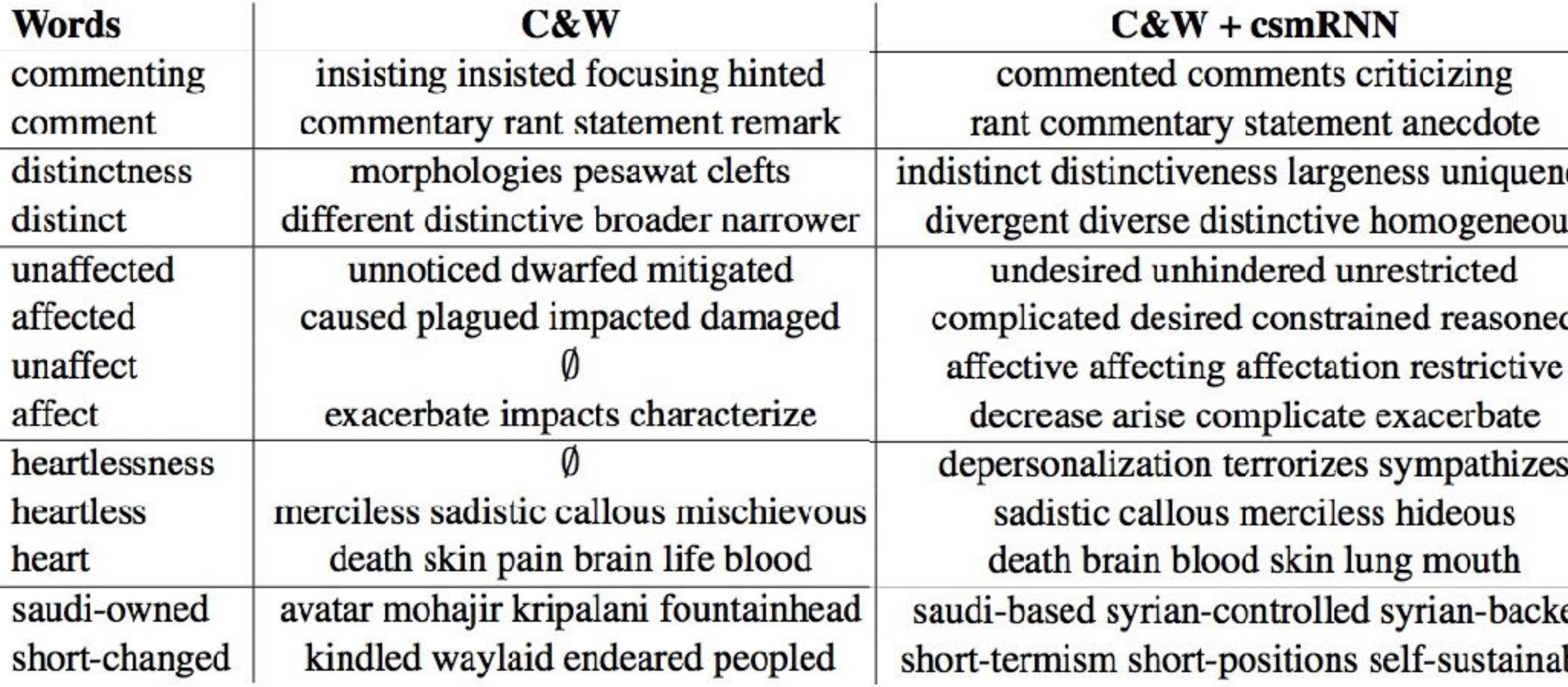
Search term

Worldwide, Past 5 years

Interest over time

fleeking, fleeked, and fleeker are all attested…

Representations learned by compositional morphology model



Words	C&W	C&W + csmRNN
commenting	insisting insisted focusing hinted	commented comments criticizing
comment	commentary rant statement remark	rant commentary statement anecdote
distinctness	morphologies pesawat clefts	indistinct distinctiveness largeness uniquen
distinct	different distinctive broader narrower	divergent diverse distinctive homogeneou
unaffected	unnoticed dwarfed mitigated	undesired unhindered unrestricted
affected	caused plagued impacted damaged	complicated desired constrained reasoned
unaffect	∅	affective affecting affectation restrictive
affect	exacerbate impacts characterize	decrease arise complicate exacerbate
heartlessness	∅	depersonalization terrorizes sympathizes
heartless	merciless sadistic callous mischievous	sadistic callous merciless hideous
heart	death skin pain brain life blood	death brain blood skin lung mouth
saudi-owned	avatar mohajir kripalani fountainhead	saudi-based syrian-controlled syrian-backe
short-changed	kindled waylaid endeared peopled	short-termism short-positions self-sustainal

Summary

- Deep learning is not magic and will not solve all of your problems, but representation learning is a very powerful idea.
- Word representations can be transferred between models.
- Word2vec trains word representations using an objective based on language modeling—so it can be trained on unlabeled data.
- Sometimes called unsupervised, but objective is supervised!
- Vocabulary is not finite.
- Compositional representations based on morphemes make our models closer to open vocabulary.