CS224C: NLP for CSS

Word Embedding

Diyi Yang

Stanford CS

Announcements

Project Pitch Session (this Thursday)

- One-Page Slides [here]
- 5 mins for each team

Homework2 due May 6th

Project proposal grades will be released tonight

Overview

- Before word embedding
- Introduction to Word2vec
- Using Embeddings in Social Sciences
- Emoji2vec
- Contextualized Word Embeddings
- Using Contextualized Word Representations in Social Sciences

Many slides credit to Kaitlyn Zhou and CS224N

How did we deal with words before?

LIWC: Linguistic Inquiry and Word Count

Positive Emotion	Negative Emotion	Insight	Inhibition	Family	Negate
appreciat*	anger*	aware*	avoid*	brother*	aren't
comfort*	bore*	believe	careful*	cousin*	cannot
great	cry	decid*	hesitat*	daughter*	didn't
happy	despair*	feel	limit*	family	neither
interest	fail*	figur*	oppos*	father*	never
joy*	fear	know	prevent*	grandf*	no
perfect*	griev*	knew	reluctan*	grandm*	nobod*
please*	hate*	means	safe*	husband	none
safe*	panic*	notice*	stop	mom	nor
terrific	suffers	recogni*	stubborn*	mother	nothing
value	terrify	sense	wait	niece*	nowhere
wow*	violent*	think	wary	wife	without

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

How should we find *informative* words?

Train a classifier based on supervised data

Predict: human-labeled connotation of a document

From: all the words and bigrams in it

Use the regression coefficients as the weights

Log odds ratio

Log likelihood ratio: does “horrible” occur more % in corpus i or j?

$$
\begin{aligned}
\mathrm{llr}(horrible) &= \log \frac{P^i(horrible)}{P^j(horrible)} \\
&= \log P^i(horrible) - \log P^j(horrible) \\
&= \log \frac{\mathrm{f}^i(horrible)}{n^i} - \log \frac{\mathrm{f}^j(horrible)}{n^j}
\end{aligned}
$$

Log odds ratio

Log odds ratio: does "horrible" have a higher odds in corpus i or j?

$$
\begin{aligned}
\operatorname{lor}(horrible) &= \log\left(\frac{P^i(horrible)}{1-P^i(horrible)}\right) - \log\left(\frac{P^j(horrible)}{1-P^j(horrible)}\right) \\
&= \log\left(\frac{\frac{\mathrm{f}^i(horrible)}{n^i}}{1-\frac{\mathrm{f}^i(horrible)}{n^i}}\right) - \log\left(\frac{\frac{\mathrm{f}^j(horrible)}{n^j}}{1-\frac{\mathrm{f}^j(horrible)}{n^j}}\right) \\
&= \log\left(\frac{\mathrm{f}^i(horrible)}{n^i-\mathrm{f}^i(horrible)}\right) - \log\left(\frac{\mathrm{f}^j(horrible)}{n^j-\mathrm{f}^j(horrible)}\right)
\end{aligned}
$$

Log odds ratio with a prior

The Dirichlet intuition is to use a large background corpus to get a prior estimate of what we expect the frequency of each word w to be.

Now with prior

$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (f_w^i + \boxed{\alpha_w})}\right) - \log\left(\frac{f_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (f_w^j + \boxed{\alpha_w})}\right)$$

n^i= size of corpus *i*, n^j = size of corpus *j*, f_w^i= count of word *w* in corpus i, f_w^j= count of word *w* in corpus *j*, α_0 is the size of the background corpus, and α_w = count of word *w* in the background corpus.)

Top 50 words associated with bad (= 1-star) reviews

Class	Words in 1-star reviews	Class	Words in 5-star reviews
Negative	*worst, rude, terrible, horrible, bad, awful, disgusting, bland, tasteless, gross, mediocre, overpriced, worse, poor*	**Positive**	*great, best, love(d), delicious, amazing, favorite, perfect, excellent, awesome, friendly, fantastic, fresh, wonderful, incredible, sweet, yum(my)*
Negation	*no, not*	**Emphatics/ universals**	*very, highly, perfectly, definitely, absolutely, everything, every, always*
1Pl pro	*we, us, our*	**2 pro**	*you*
3 pro	*she, he, her, him*	**Articles**	*a, the*
Past verb	*was, were, asked, told, said, did, charged, waited, left, took*	**Advice**	*try, recommend*
Sequencers	*after, then*	**Conjunct**	*also, as, well, with, and*
Nouns	*manager, waitress, waiter, customer, customers, attitude, waste, poisoning, money, bill, minutes*	**Nouns**	*atmosphere, dessert, chocolate, wine, course, menu*
Irrealis modals	*would, should*	**Auxiliaries**	*is/'s, can, 've, are*
Comp	*to, that*	**Prep, other**	*in, of, die, city, mouth*

Jurafsky, D., V. Chahuneau, B. R. Routledge, and N. A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday, 19(4).

Overview

- Introduction to Word2vec
- Using Embeddings in Social Sciences
- Emoji2vec
- Contextualized Word Embeddings
- Using Contextualized Word Representations in Social Sciences

Problems with resources like WordNet

- A useful resource but missing nuance:
  - e.g., "proficient" is listed as a synonym for "good"
    This is only correct in some contexts
  - Also, WordNet list offensive synonyms in some synonym sets without any coverage of the connotations or appropriateness of words

- Missing new meanings of words:
  - e.g., wicked, badass, nifty, wizard, genius, ninja, bombest
  - Impossible to keep up-to-date!

- Subjective
- Requires human labor to create and adapt
- Can't be used to accurately compute word similarity (see following slides)

Representing words as discrete symbols

In traditional NLP, we regard words as discrete symbols:

hotel, conference, motel – a localist representation

Such symbols for words can be represented by one-hot vectors:

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Vector dimension = number of words in vocabulary (e.g., 500,000+)

Representing words by their context

- Distributional semantics: **A word's meaning is given by the words that frequently appear close-by**
  - *"You shall know a word by the company it keeps"* (J. R. Firth 1957: 11)
  - One of the most successful ideas of modern statistical NLP!

- When a word w appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).

- We use the many contexts of w to build up a representation of w

Word Vectors

We will build a dense vector for each word, chosen so that it is similar to vectors of words that appear in similar contexts, measuring similarity as the vector dot (scalar) product

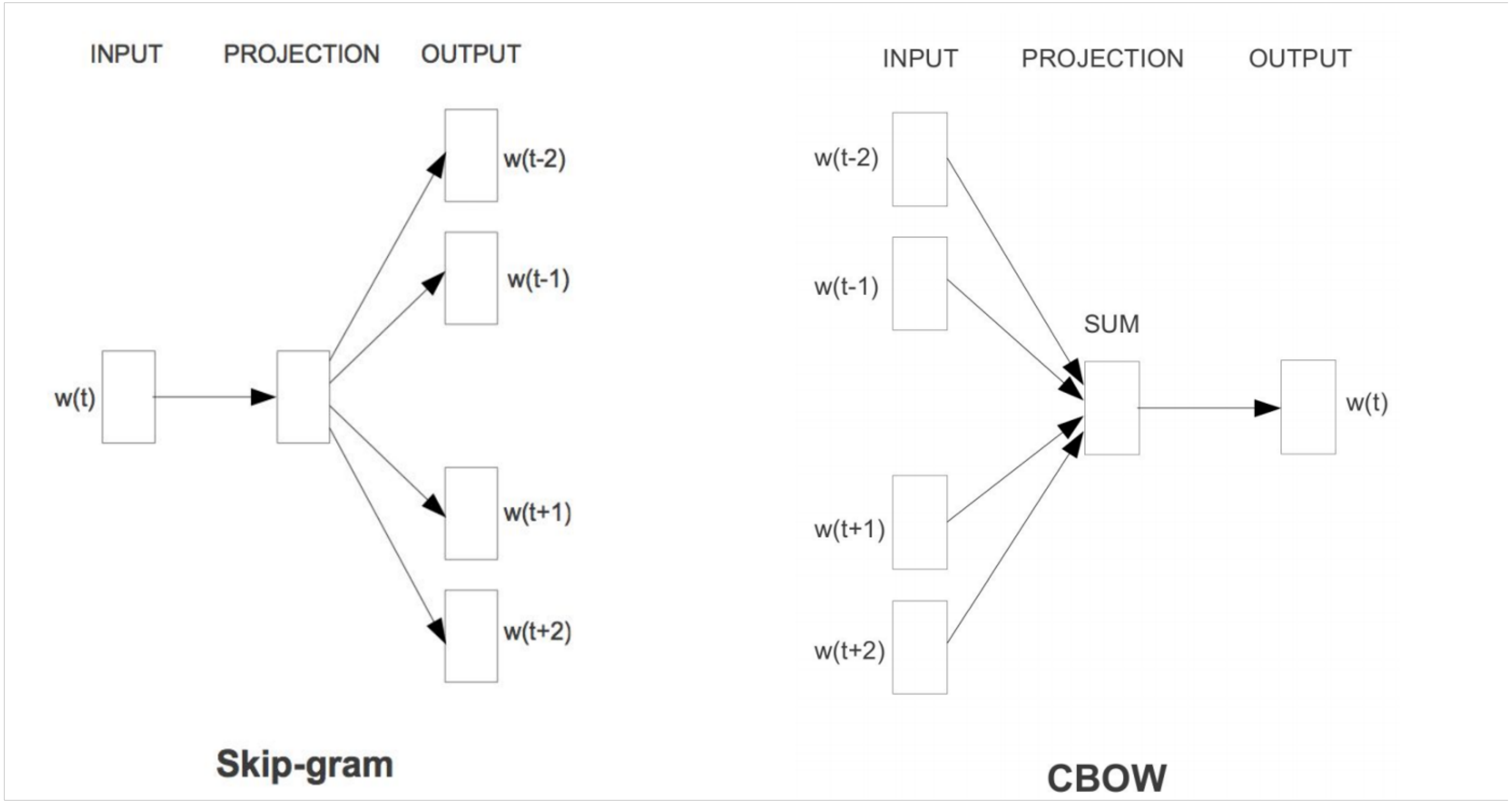
$$\textit{banking} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix} \qquad \textit{monetary} = \begin{pmatrix} 0.413 \\ 0.582 \\ -0.007 \\ 0.247 \\ 0.216 \\ -0.718 \\ 0.147 \\ 0.051 \end{pmatrix}$$

Word2Vec

Idea: words that are semantically similar often occur in similar context

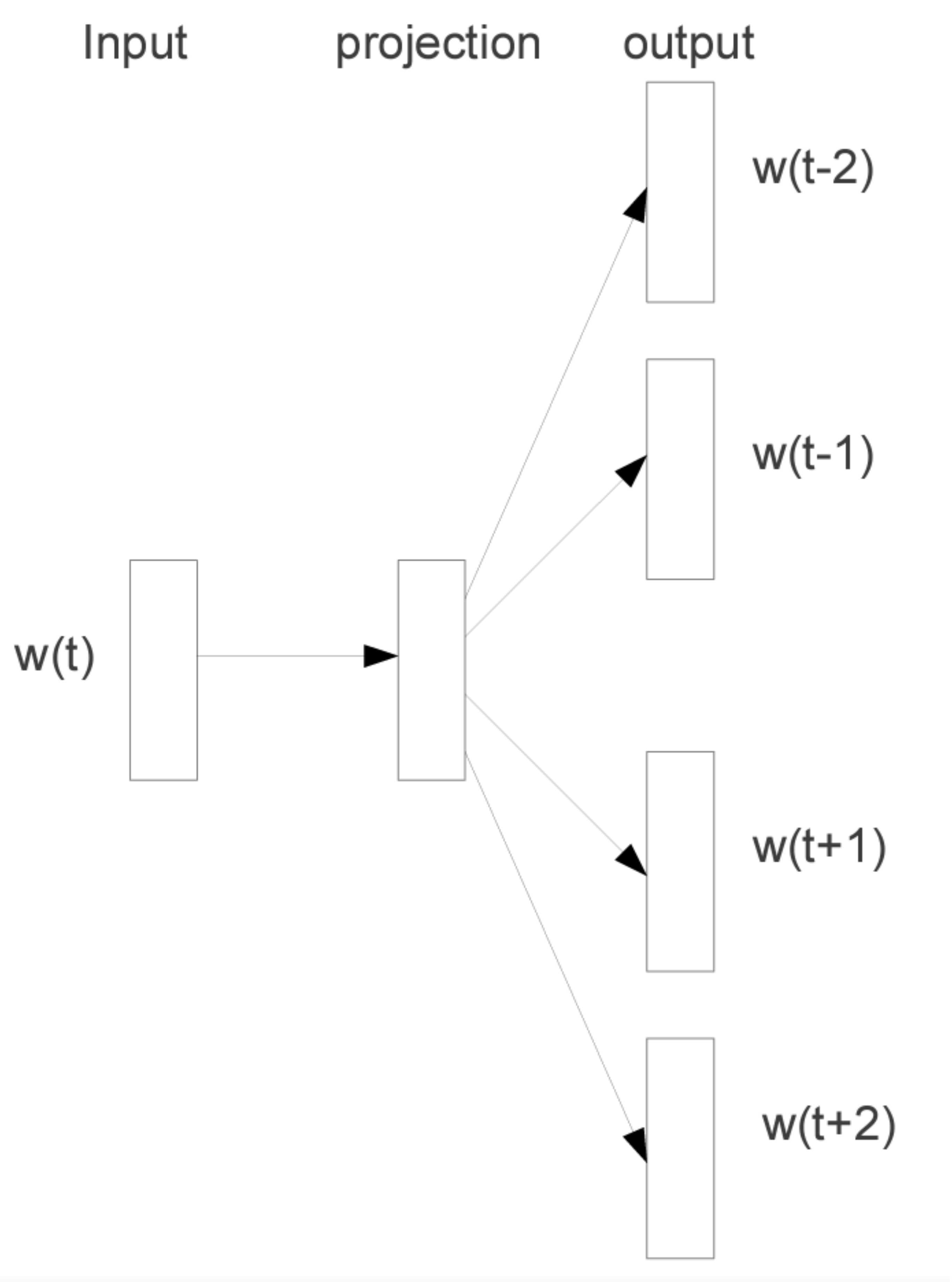
Embeddings that are good at predicting neighboring words are also good at representing similarity

Skip-gram vs. Continuous Bag of Words



Word2vec: Overview

Skip-gram model
(Mikolov et al. 2013)



We have a large corpus ("body") of text

Every word in a fixed vocabulary is represented by a vector

Go through each position *t* in the text, which has a center word *c* and context ("outside") words *o*

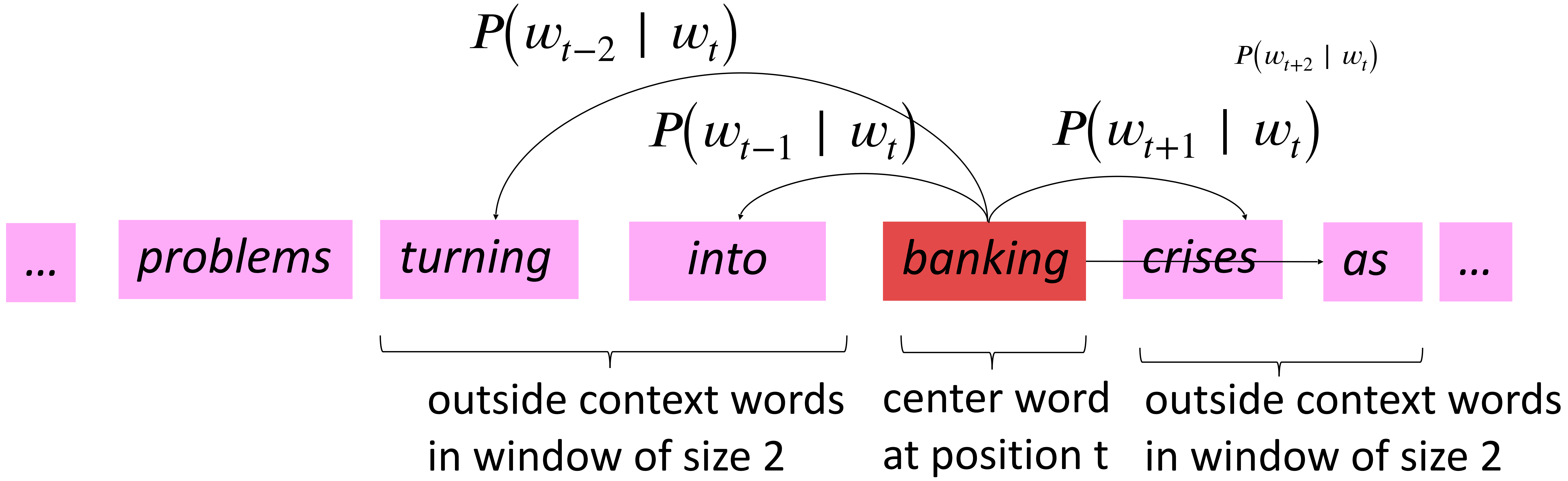
Use the similarity of the word vectors for *c* and *o* to calculate the probability of *o* given *c* (or vice versa)

Keep adjusting the word vectors to maximize this probability

Slides from CS224n

Word2Vec Overview

Example windows and process for computing $P\left(w_{t+j} \mid w_t\right)$



Slides from CS224n

Word2vec: objective function

For each position $t = 1, \ldots, T$, predict context words within a window of fixed size *m*, given center word w_t. Data likelihood:

$$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} P\left(w_{t+j} \mid w_t; \theta\right)$$

The objective function $J(\theta)$ is the (average) negative log likelihood:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log P\left(w_{t+j} \mid w_t; \theta\right)$$

Minimizing objective function $\Longleftrightarrow$ Maximizing predictive accuracy

Slides from CS224n

Word2vec: objective function

We want to minimize the objective function:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m \le j \le m \\ j \neq 0}} \log P\left(w_{t+j} \mid w_t; \theta\right)$$

Question: How to calculate $P\left(w_{t+j} \mid w_t; \theta\right)$?

Answer: We will *use two* vectors per word *w*:

v_w when *w* is a center word

u_w when *w* is a context word

Then for a center word *c* and a context word *o*:

$$P(o \mid c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Word2Vec skip-gram model with negative sampling

Instead of counting how often each word w occurs near "peach"

Train a classifier on a binary prediction task:

Is w likely to show up near "peach"?

We don't actually care about this task

But we'll take the learned classifier weights as the word embeddings

Skim-Gram Sketch

- Treat the target word and a neighboring context word as positive examples
- Randomly sample other words in the lexicon to get negative samples
- Use logistic regression to train a classifier to distinguish those two cases
- Use the weights as the embeddings

Measuring the Semantic Similarity of Vectors

The most common similarity metric is cosine, which is the angle between the vectors

For vectors u and v, the cosine similarity is the dot product of the two vectors, divided by the product of the length of the two vectors

Other distance (Euclidean, norms) might be appropriate and meaningful for a number of other tasks

Tasks Semantic Similarity

Example 1: Automatically identifying components of parts

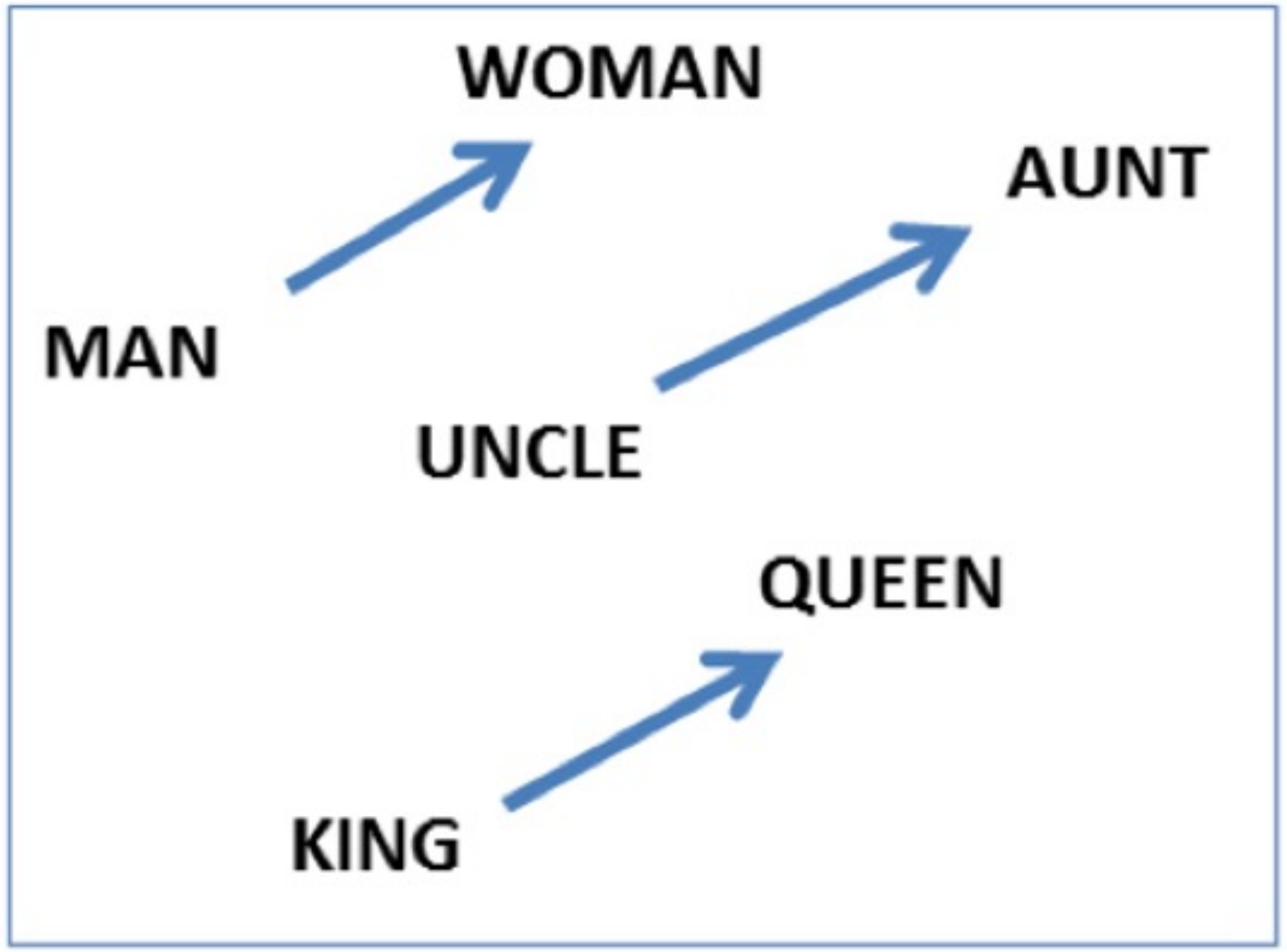
Example 2: Identifying related concepts in a historical corpus

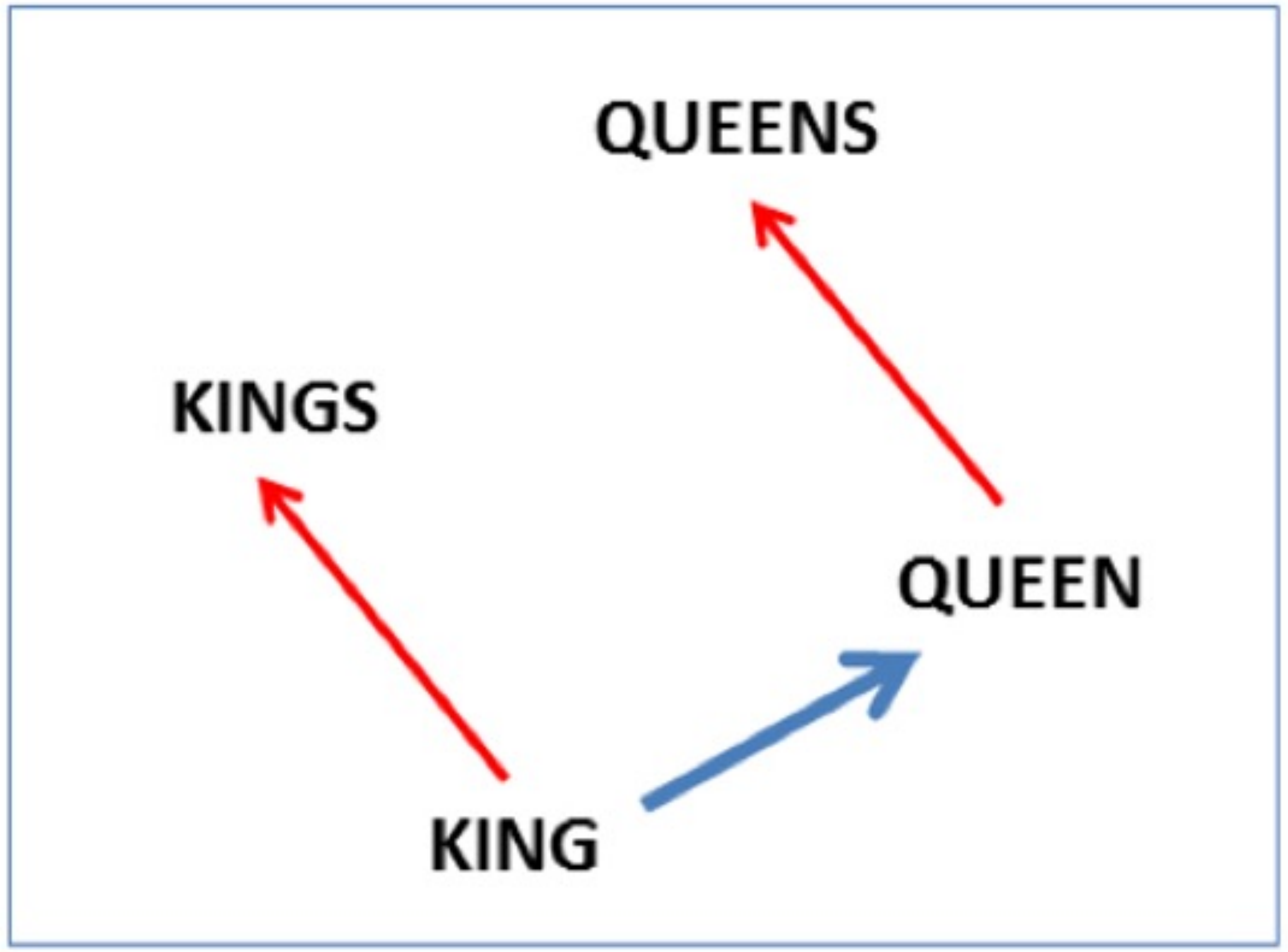
Evaluation Datasets WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015)

Analogy: Embeddings Capture Relational Meaning

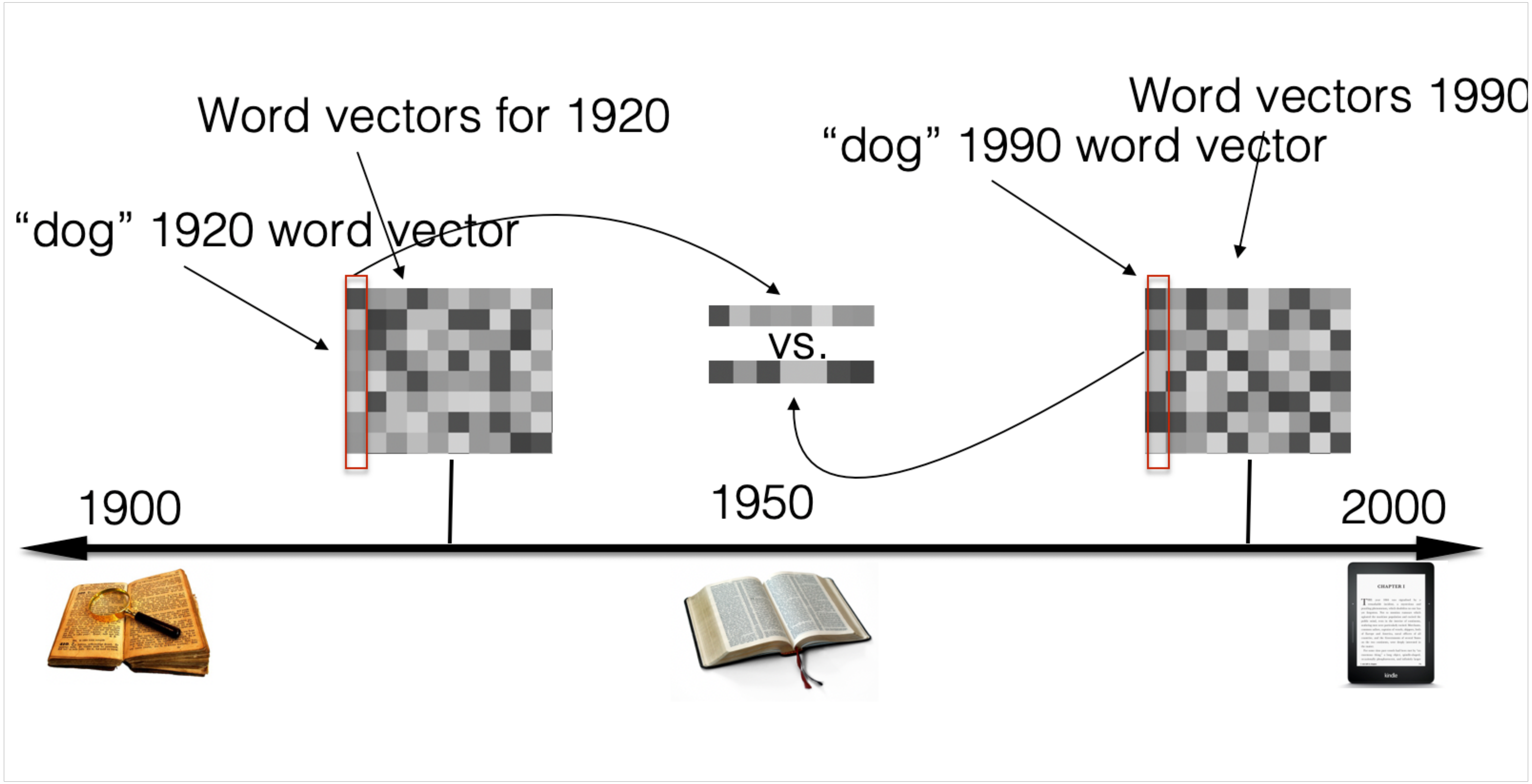
vector('king') - vector('man') + vector('woman') ≈ vector('queen')

vector('Paris') - vector('France') + vector('Italy') ≈ vector('Rome')





Diachronic word embeddings for studying language change!



Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." arXiv preprint arXiv:1605.09096 (2016).

Diachronic word embeddings for studying language change!

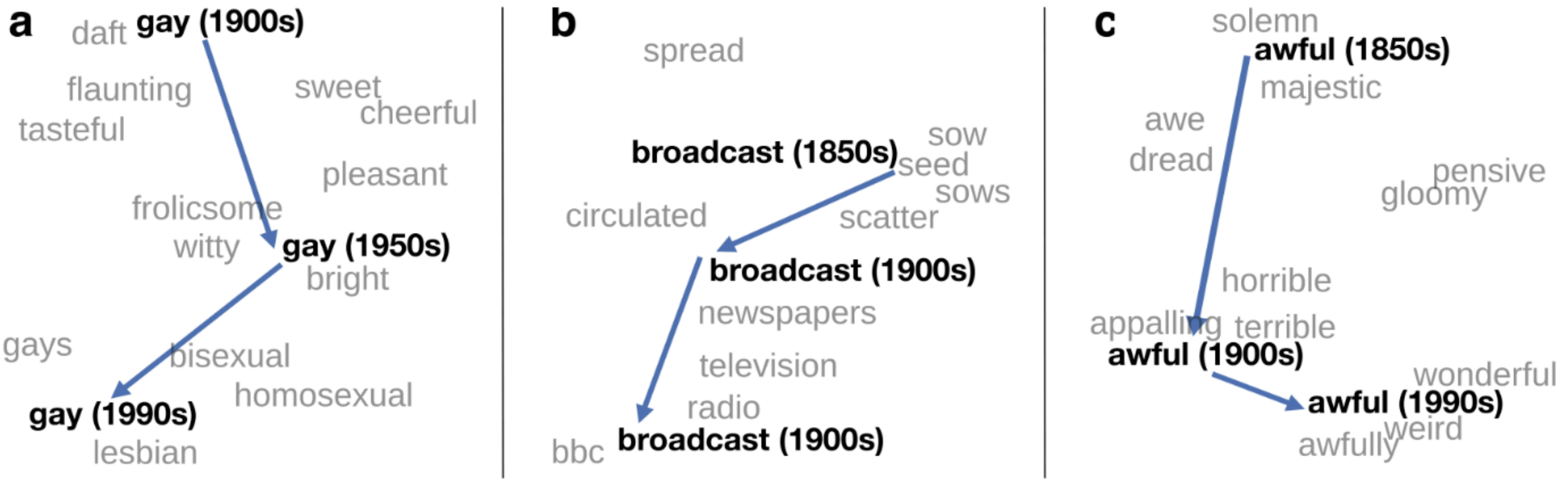


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

Diachronic word embeddings for studying language change!

Aligning historical embeddings via orthogonal Procrustes to find the best rotational alignment

$W^{(t)}$ as the matrix of word embedding learned at year t, align across time-periods while preserving cosine similarities by optimizing

$$\mathbf{R}^{(t)} = \arg \min_{\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}} \|\mathbf{W}^{(t)}\mathbf{Q} - \mathbf{W}^{(t+1)}\|_F$$

Embeddings Reflect Cultural Bias

Ask “Paris : France :: Tokyo : x”

x = Japan

Ask “father : doctor :: mother : x”

x = nurse

Ask “man : computer programmer :: woman : x”

x = homemaker

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In Advances in Neural Information Processing Systems, pp. 4349-4357. 2016.

Embedding Reflect Cultural Biases

Implicit Association test (Greenwald et al 1998): How associated are

concepts (flowers, insects) & attributes (pleasantness, unpleasantness)?

Studied by measuring timing latencies for categorization.

Trained Embeddings

Word2vec (Mikolov et al., 13)

- https://code.google.com/archive/p/word2vec/

Fasttext (Bojanowski et al., 17)

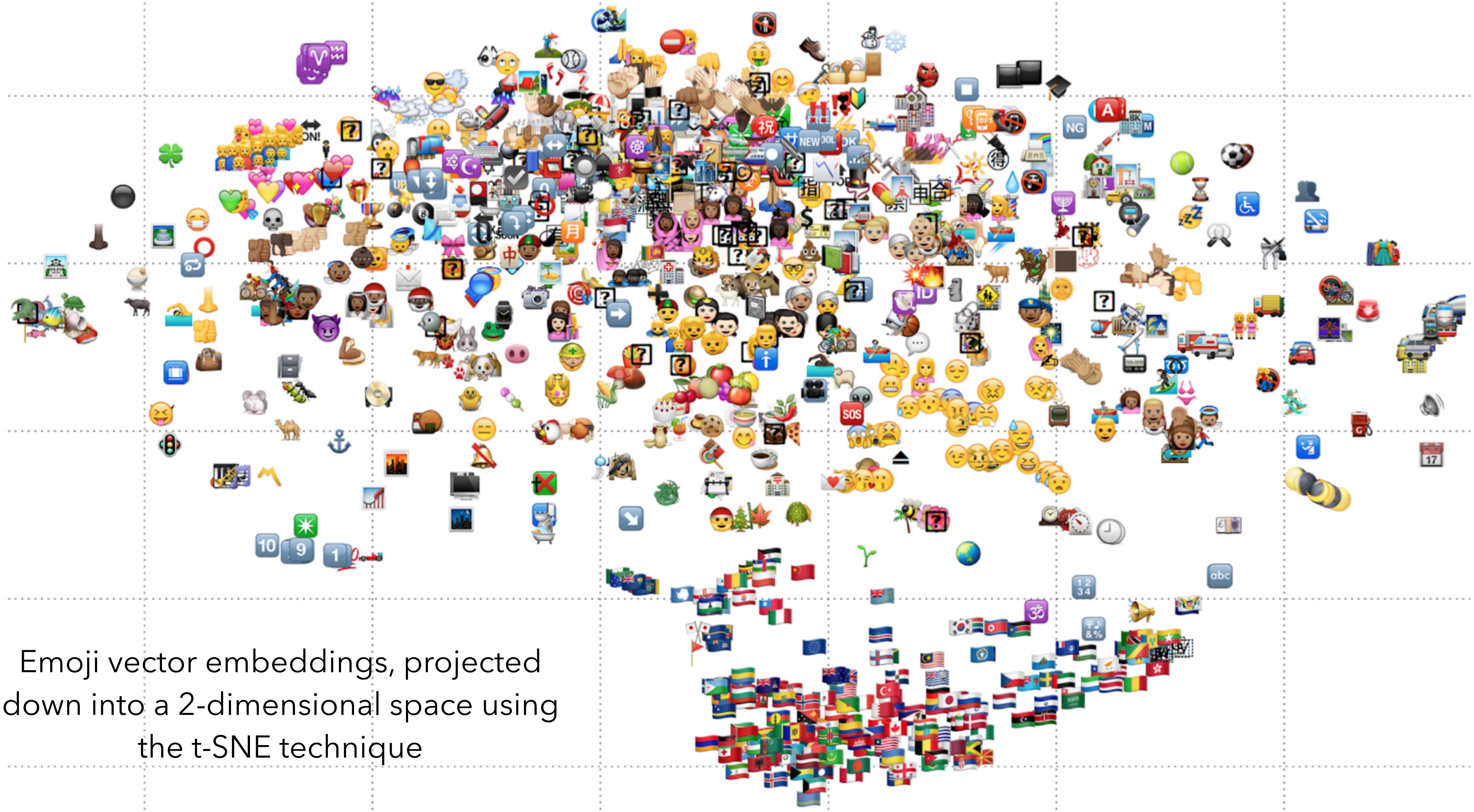
- https://fasttext.cc/

Glove (Pennington et al., 14)

- https://nlp.stanford.edu/projects/glove/

Emoji Analogy Examples

Eisner, Ben, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. "emoji2vec: Learning emoji representations from their description." arXiv preprint arXiv:1609.08359 (2016).

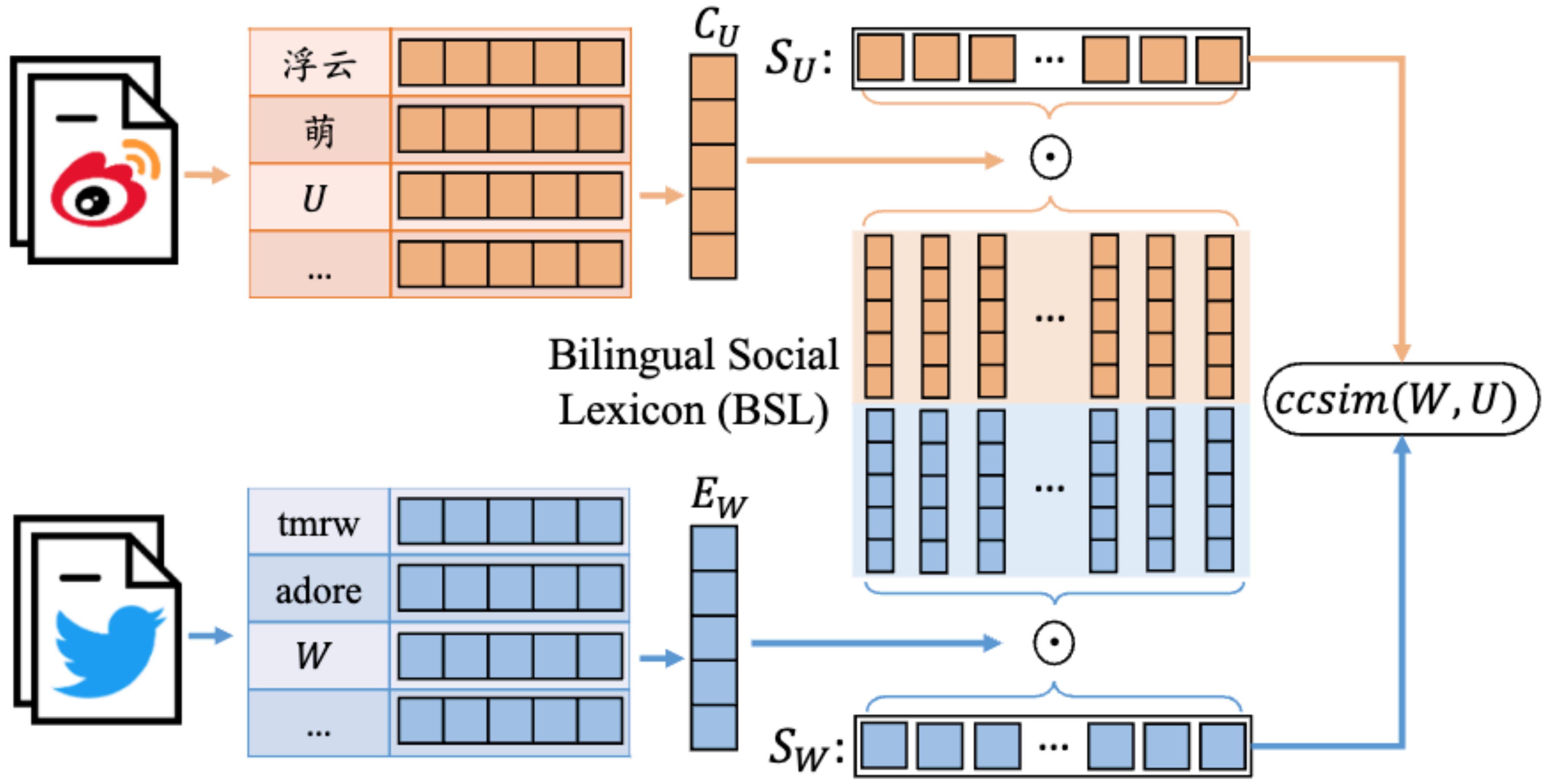


Emoji vector embeddings, projected down into a 2-dimensional space using the t-SNE technique

Eisner, Ben, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. "emoji2vec: Learning emoji representations from their description." arXiv preprint arXiv:1609.08359 (2016).

Cross-Cultural Differences via Word2Vec

Computing the cross-cultural similarity between an English word W and a Chinese word U



Lin, Bill Yuchen, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. "Mining cross-cultural differences and similarities in social media." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 709-719. 2018.

Cross-Cultural Differences via Word2Vec

Chinese Slang	English Slang	Explanation
萌	adorbz, adorb, adorbs, tweeny, attractiveee	cute, adorable
二百五	shithead, stupidit, douchbag	A foolish person
鸭梨	antsy, stressy, fidgety, grouchy, badmood	stress, pressure, burden

Slang	Explanation	Google	Bing	Baidu	Ours
浮云	something as ephemeral and unimportant as "passing clouds"	clouds	nothing	floating clouds	nothingness, illusion
水军	"water army", people paid to slander competitors on the Internet and to help shape public opinion	Water army	Navy	Navy	propaganda, complicit, fraudulent

Lin, Bill Yuchen, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. "Mining cross-cultural differences and similarities in social media." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 709-719. 2018.

Issues of Static Word Embedding

Typically ignores that one word can have different senses.

I went to the river **bank** yesterday.
I had been to the **bank** to withdraw some money.

Solution: contextualized word embedding

Give words different embeddings based on the context of the sentence (e.g. ELMo, BERT).

Contextualized Word Embeddings

Contextualized embeddings are pre-trained using context and additionally, embed words with their contexts to get a contextualized representation of word tokens

- Deep contextualized word representations (Peters et al.) (ELMo)
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.) (BERT)

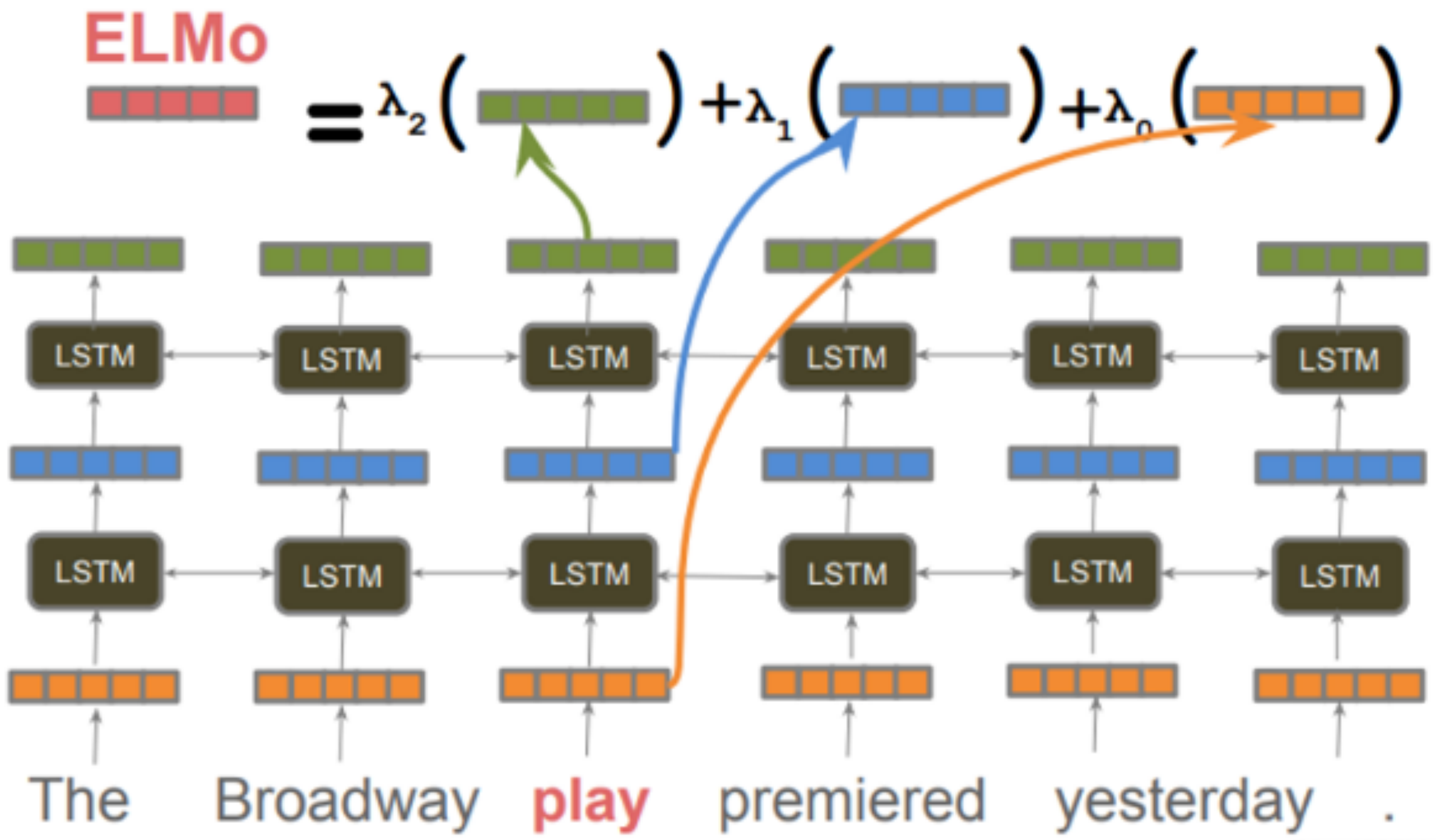
Elmo

Deep contextualized word representations

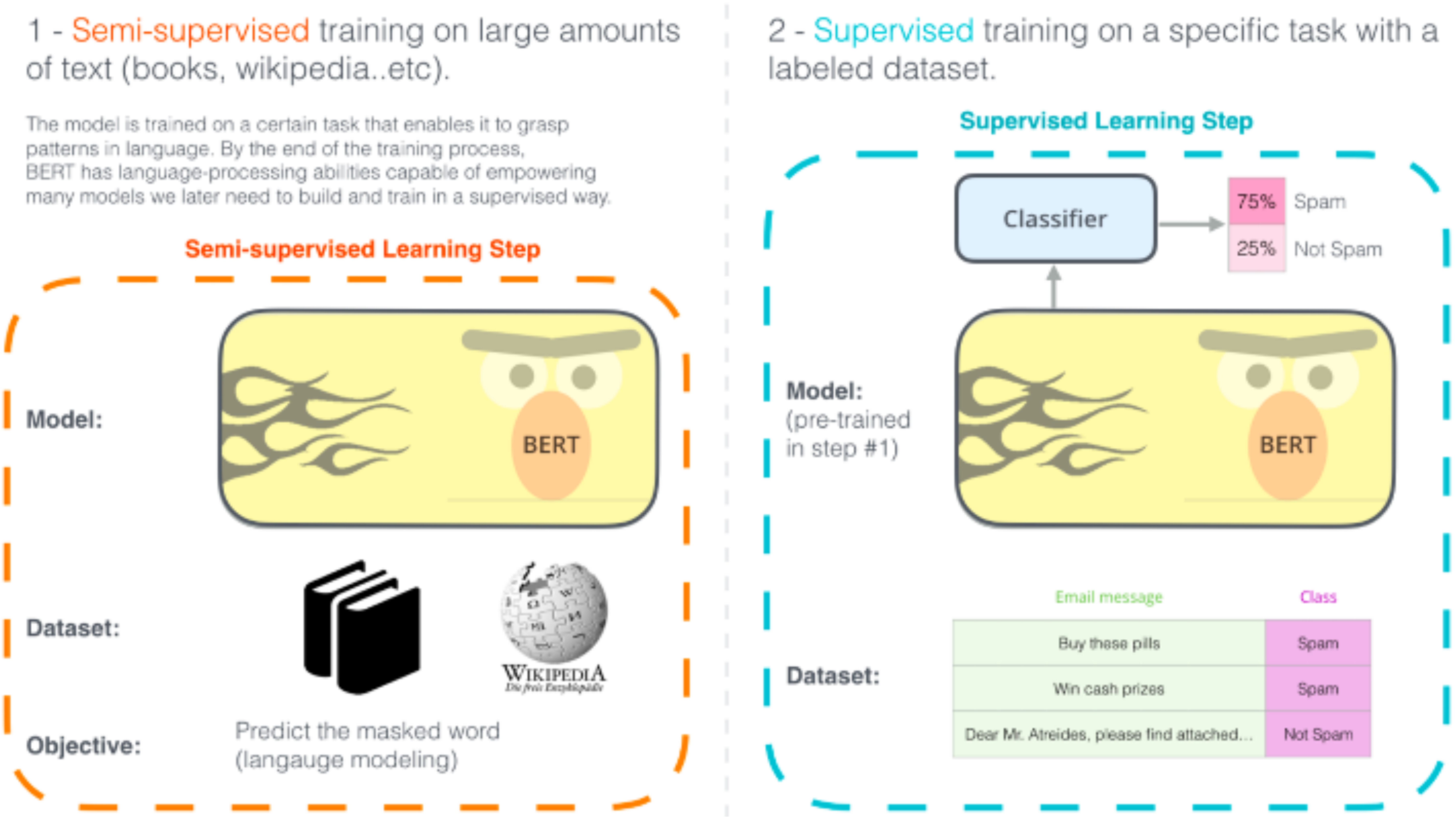
Matthew E. Peters[†], Mark Neumann[†], Mohit Iyyer[†], Matt Gardner[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], Kenton Lee[*], Luke Zettlemoyer[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

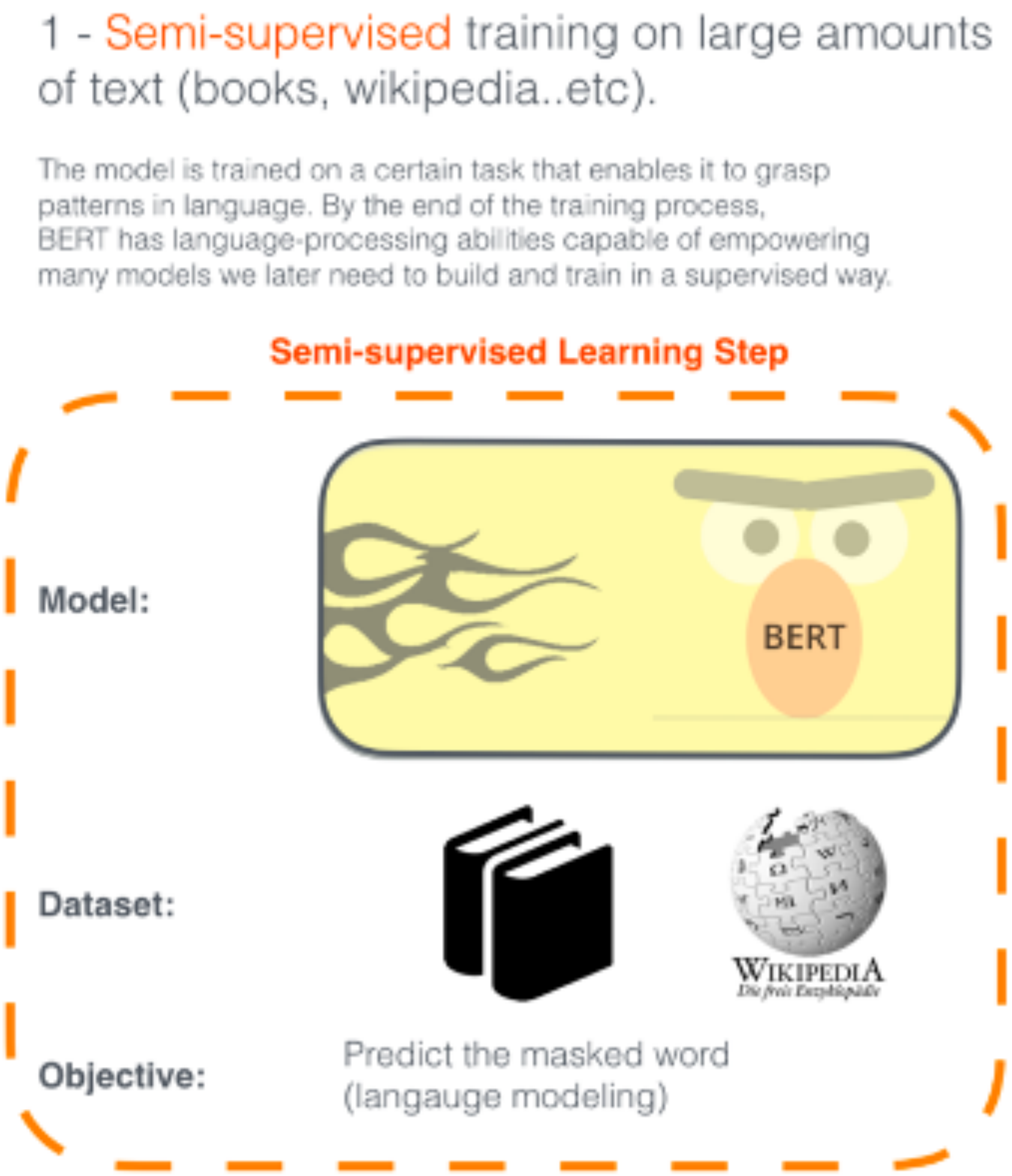


BERT



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

BERT



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

Pretraining:
Train transformer-alike models on a large dataset (e.g. books, or the entire web).

This step learns **general structure** and meaning of the text (e.g. "good" is an adjective), similar to word embedding; the knowledge is reflected by the model parameter (hence really large models).

Contextualized Word Embeddings

- For BERT, to create word embeddings, feed the model a sentence with the target word, “I went to the bank.”
- Extract the last few hidden layers from the model corresponding to the target word
- Take the average (or concatenation) of the hidden layers

Contextualized Word Embeddings for CSS

We can perform the analysis discussed above but at a more granular level!

In the diachronic sense change example, we needed to train two separate models to extract pre-trained embeddings from two different time intervals

With contextualized word embeddings, we simply have to pass in two different contexts of the word

This is done, without needing to retrain the model

Applications of Contextualized Word Embeddings

We can also examine how contemporary speakers use the same word differently

- Card et al. (2022) examines how use of the word immigrant has changed over time and how the word is used differently across political parties
- Lucy et al. (2022) examines how the representation of people varies across online communities

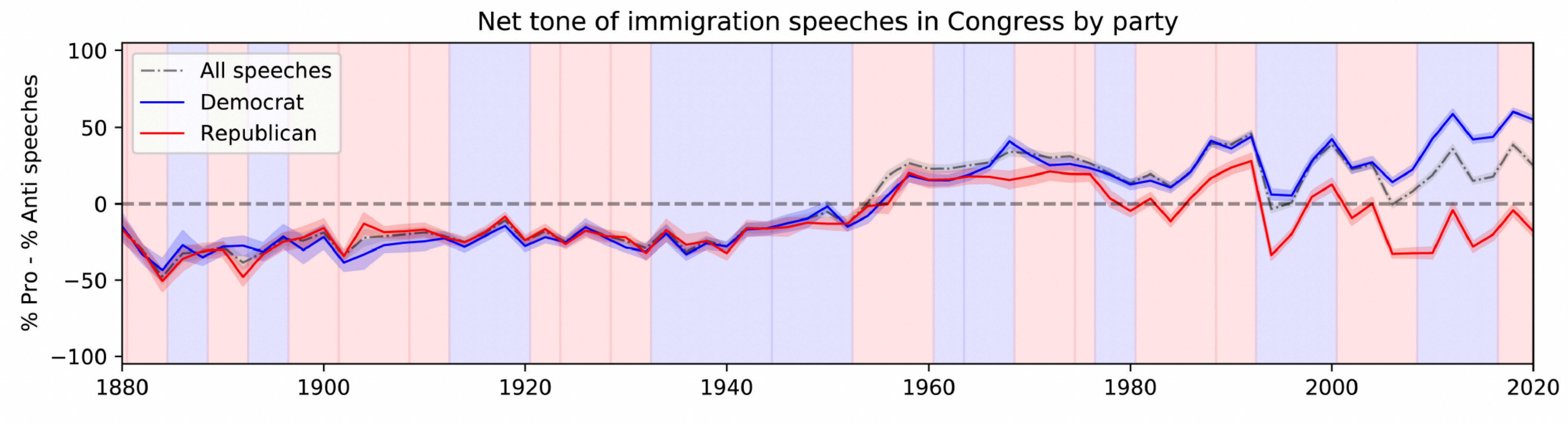
Lucy, Li, Divya Tadimeti, and David Bamman. "Discovering Differences in the Representation of People using Contextualized Semantic Axes." EMNLP 2022

Card, Dallas et al. “Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration.” Proceedings of the National Academy of Sciences of the United States of America 119 (2022)

Increasingly Polarized Framing of Immigration

Quantitative analysis of 140 years of US congressional and presidential speech about immigration

Find a rise in pro-immigration attitudes beginning in the 1940s, followed by a steady decline among Republicans (relative to Democrats)



Method for Measuring Implicit Dehumanizing Metaphors

- For each sentence that mentions “immigrant”, remove the mention (e.g., “foreigners”) from the sentence, replacing it with a special <MASK> (e.g., “the tendency of [MASK] to flock together”)
- Feed the sentence through the model and examine the words the model is predicting for the <MASK> token
- Over the predictions, sum together the probability that was placed on dehumanizing terms like “animal” or “cargo”
- The lists dehumanizing terms were selected ahead of time and are sorted into categories

14 Frames used by Republicans compared to Democrats

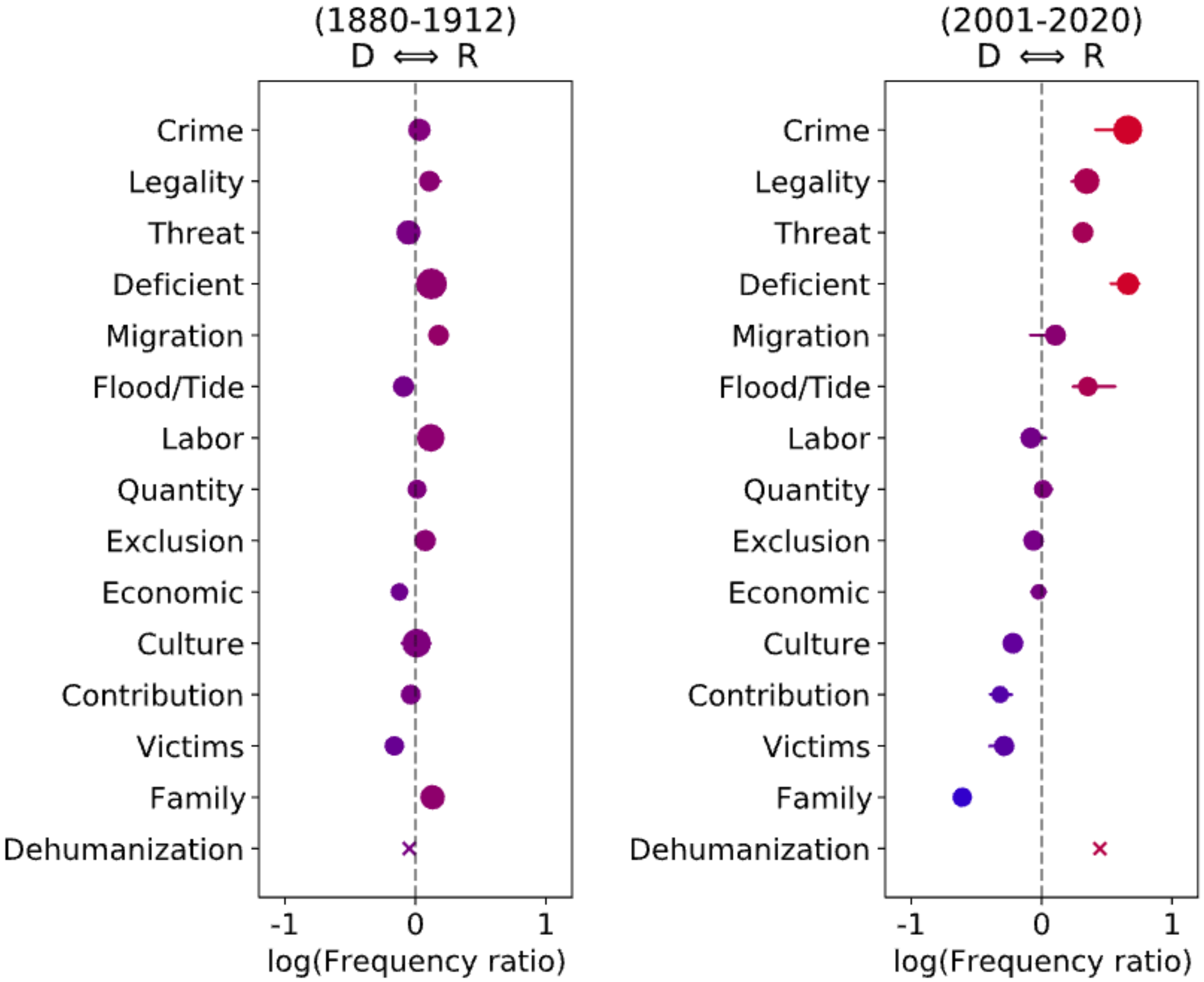


Fig. 3. Relative usage frequency for each of 14 frames by Republicans compared to Democrats, both for the late 19th/early 20th century (*Left*) and the past 2 decades (*Right*). Farther to the left on each plot represents more frequent usage by Democrats and vice versa (plotted as log frequency ratio). Circle size represents the overall prominence of the frame in speeches about immigration, relative to all speeches. To ensure the robustness of these findings, we leave out each word in turn from each frame and show the full range of possible values obtained using horizontal lines (not visible when the full range is contained within the circle). "Dehumanization" is an aggregation of metaphorical categories (see *Measuring Dehumanization*). Compared to the absence of polarization a century ago, certain frames today are disproportionately used by Republicans ("crime," "legality," "threats," "deficiency," and "flood/tide") and Democrats ("family," "victims," "contributions," and "culture"). Republicans also show significantly higher use of implicit dehumanizing metaphors like "animals" and "cargo."

Contextualized Word Embeddings Aren't Free From Biases

- Static embeddings are heavily biased by frequency based on their training (words that occur more frequently are going to be represented more closely together)
- Wolfe and Caliskan (2021) illustrate how BERT embeddings also associate minority names more likely with unpleasantness
- Zhou et al. (2022) shows how the names of low frequency (typically poorer) countries are seen as less distinct than those from high frequency (typically richer countries)

Wolfe, R., & Caliskan, A. (2021). Low frequency names exhibit bias and overfitting in contextualizing language models. arXiv preprint arXiv:2110.00672.
Zhou, Kaitlyn, Kawin Ethayarajh, and Dan Jurafsky. "Richer countries and richer representations." ACL Findings 2022

Naming these Harms

Allocation Harms: where systems unfairly allocate resources

- Imagine a recommendation system that more closely associates doctors with masculine names --- resulting in fewer opportunities for those with feminine names

Representation harms: where systems represent a group of people in an unpleasant, harmful, or demeaning manner

- Certain groups of people being represented in stereotypical or limiting ways

Some of these harms are a result of the training data, but these harms are at times further exacerbated by the algorithms and systems we build

Crawford, K. 2017. The trouble with bias. Keynote at NeurIPS.
Blodgett, S. L., S. Barocas, H. Daume III, and H. Wallach. 2020. ´ Language (technology) is power: A critical survey of "bias" in NLP. ACL.

Looking ahead

- The improvement in our ability to represent words has been the foundational to the transformative progress in NLP
- As methods and techniques improve on how words are represented, as computational social scientists, we are better able to conduct accurate and fine-grained analysis of language use
- This analysis reveals to us how words use changes over time, how concepts are connected, and where there are systematic biases and stereotypes to overcome