Word embeddings (II)

CS 490A, Fall 2020

Applications of Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs490a_f20/

Brendan O’Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

including slides from Eisenstein (2019) and Jurafsky & Martin 3rd Ed.

- Back to demo - what’s in these embeddings?

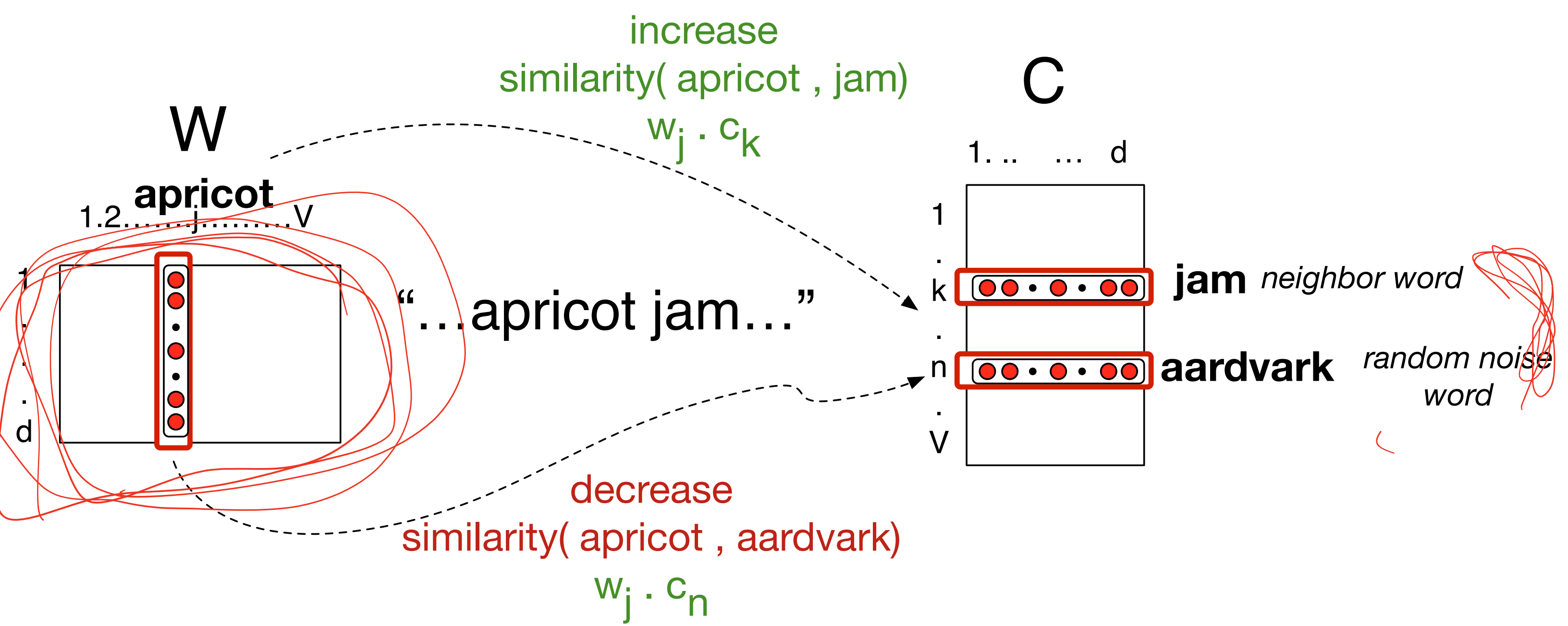
Word embeddings

- 1. Input: Large textual corpus

- Wikipedia
- News
- Books
- Social Media

- 2. Language Model: predict words from nearby words
  - GLOVE, SVD: factorize the word-context cooccurrence matrix
  - word2vec: model is viewed as predicting one word's surrounding context words

- 3. Take out the vectors the model was forced to learn; use in downstream applications in practice, we learn two different sets of embeddings (**W** for *target* words, **C** for context words), but throw away **C**



Defining contexts

Window size C affects the nature of the similarity something like…

syntax <—> basic meaning <—> topical meaning

C = ±2 The nearest words to *Hogwarts:*

- *Sunnydale*
- *Evernight*

C = ±5 The nearest words to *Hogwarts:*

- *Dumbledore*
- *Malfoy*
- *halfblood*

Defining contexts

subject   direct obj.

NSUBJ   DOBJ

*The moment one **learns** English, complications set in* (Alfau, 1999)	
Brown Clusters	$\{one\}$
WORD2VEC, $h = 2$	$\{moment, one, English, complications\}$
Structured WORD2VEC, $h = 2$	$\{(moment, -2), (one, -1), (English, +1), (complications, +2)\}$
Dependency contexts,	$\{(one, \text{NSUBJ}), (English, \text{DOBJ}), (moment, \text{ACL}^{-1})\}$

DOBJ

eat   [TARGET]

Alternate/mis- spellings

- Distributional methods are really good at this
- Brown clusters on Twitter: http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

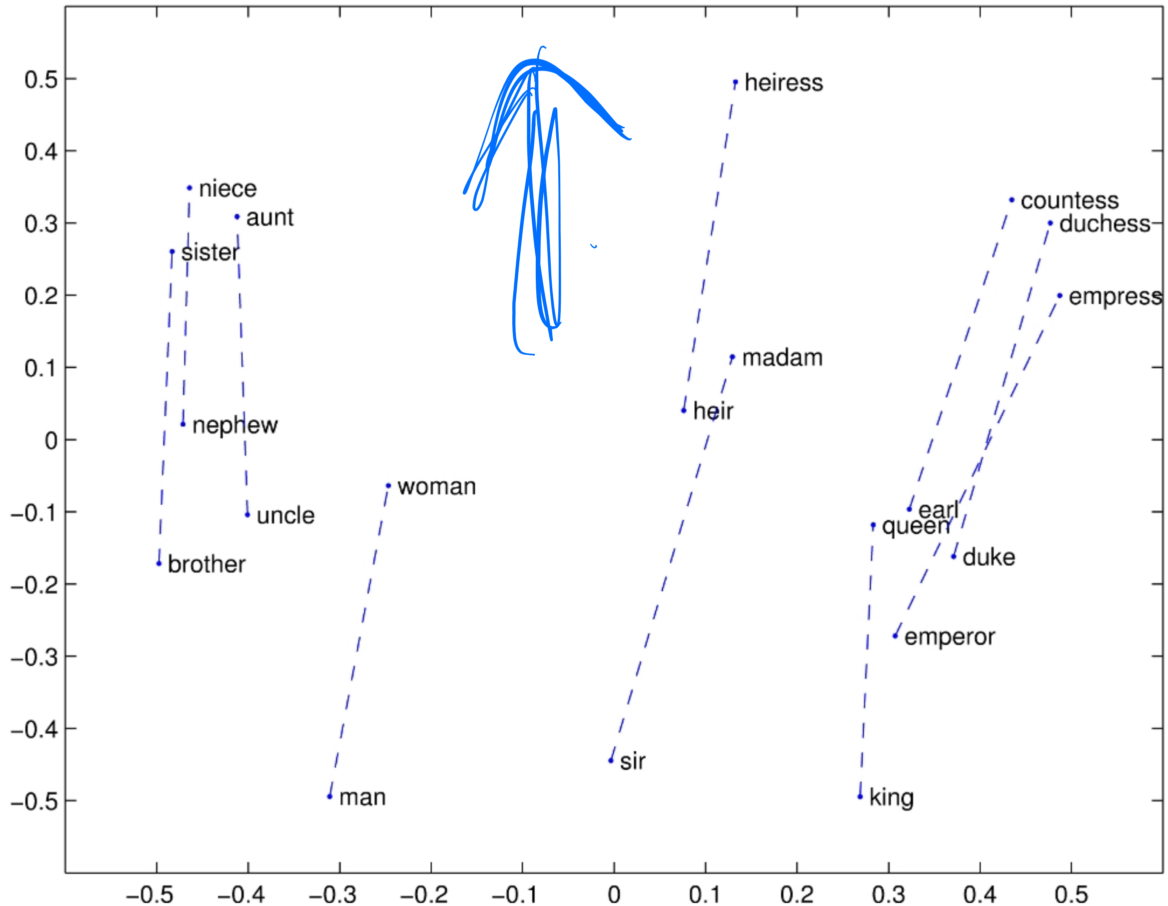
Pre-trained embeddings

- Widely useful. But make sure you know what you're getting!
  - Examples: GLOVE, fasttext, word2vec, etc.
  - Is the corpus similar to what you care about?
  - Should you care about the *data*?

Evaluating embeddings

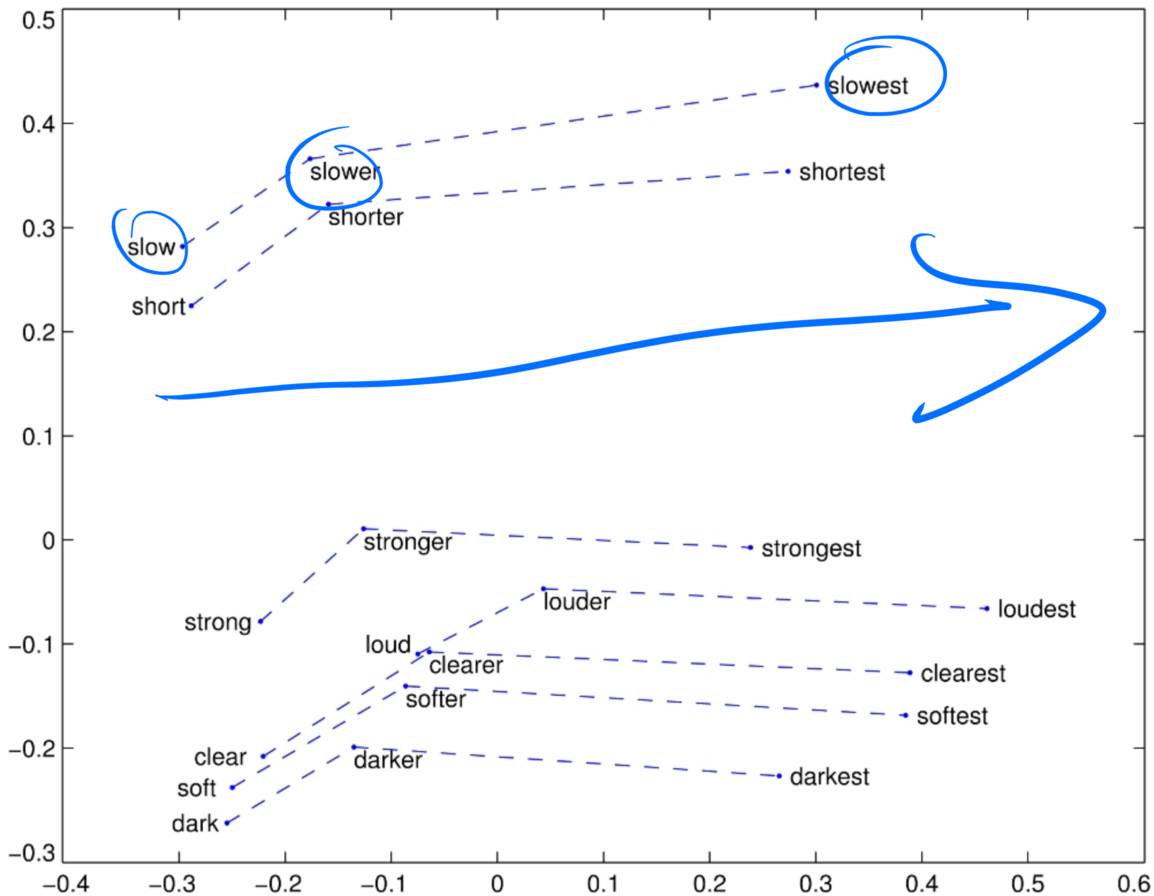
- Anecdotal inspection (not a real evaluation, but better than nothing; but seen next slides)
- Intrinsic evaluations
  - Compare embeddings' word pair similarities to human judgments
    - TOEFL: "*Levied* is closest to *imposed*, *believed*, *requested*, *correlated*"
    - Numerical similarity judgments (e.g. Wordsim-353)
  - There some other attempts at this (word analogies) but IMO not trustworthy (e.g. Linzen 2016)
- Extrinsic evaluation: use embeddings in some task

PCA dim. reduction of selected embeddings



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

PCA dim. reduction of selected embeddings



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

Application: keyword expansion

- I have a few keywords for my task. Are there any I missed?
- Automated or semi-automated new terms from embedding neighbors

dog

Canine

cat

hound

Select

is this right for my app.?

words

- Other non-embedding lexical resources can do this too (e.g. WordNet), but word embeddings typically cover a *lot* of diverse vocabulary why better??

Application: document embedding

- Instead of bag-of-words, can we derive a latent embedding of a document/sentence?

Avg them!

"Bag of Embeddings"

$w_1 = I \Rightarrow V(w_1) = [\,1st \;|\; \;|\; \;|\; \,]$

$w_2 = saw \Rightarrow V(w_2) = [\,2 \;|\; \,]$

$w_3 = the$

w_4 movie

Avg

Doc Embed = $[\;|\;|\;|\;|\;]$

① $\vec{x} = \frac{1}{N_{tok}} \sum_{i=1}^{N_{tok}} V(w_i)$

② Use $\vec{x}$ for LogReg

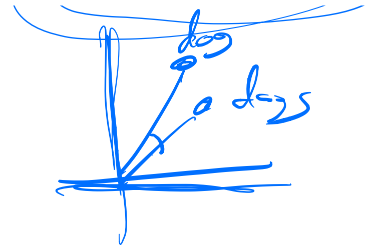
$p(y=1|x) = \frac{1}{1+e^{-\beta^T \vec{x}}}$

See: Arora et al. 2017

$dog = [5, 3, -2]$

$dogs = [5.1, 2.8, -1.2]$

$\beta = [2, 2, 2,]$



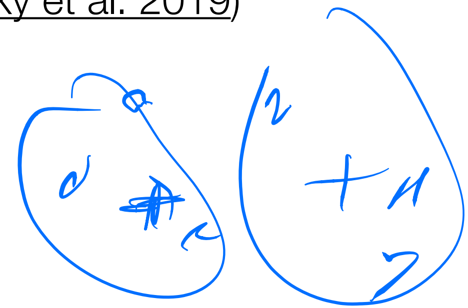
inv freq weighting

$$\tilde{x} = \frac{1}{Z} \sum_i \frac{\overbrace{a}^{10^{-3}}}{a + \underbrace{p(w)}_{\text{freq. in corpus}}} v(w_i)$$

$Z =$

Exploratory usage

- Example: tweets about mass shootings (Demszky et al. 2019)
  1. Average word embeddings => tweet embeddings
  2. Cluster tweets (kmeans)
  3. Interpret clusters' words (closest to centroid)



Topic	10 Nearest Stems
news (19%)	break, custodi, #breakingnew, #updat, confirm, fatal, multipl, updat, unconfirm, sever
investigation (9%)	suspect, arrest, alleg, apprehend, custodi, charg, accus, prosecutor, #break, ap
shooter's identity & ideology (11%)	extremist, radic, racist, ideolog, label, rhetor, wing, blm, islamist, christian
victims & location (4%)	bar, thousand, california, calif, among, los, southern, veteran, angel, via
laws & policy (14%)	sensibl, regul, requir, access, abid, #gunreformnow, legisl, argument, allow, #guncontolnow
solidarity (13%)	affect, senseless, ach, heart, heartbroken, sadden, faculti, pray, #prayer, deepest
remembrance (6%)	honor, memori, tuesday, candlelight, flown, vigil, gather, observ, honour, capitol
other (23%)	dude, yeah, eat, huh, gonna, ain, shit, ass, damn, guess

Table 1: Our eight topics (with their average proportions across events) and nearest-neighbor stem embeddings to the cluster centroids. Topic names were manually assigned based on inspecting the tweets.