

Word Embedding Evaluation



Evaluating Word Embeddings

► What properties of language should word embeddings capture?

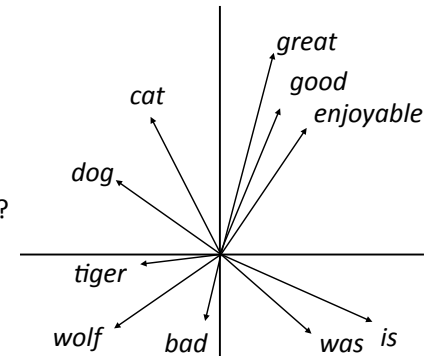
► Similarity: similar words are close to each other

► Analogy:

good is to best as smart is to ???

Paris is to France as Tokyo is to ???

► Bias?



Similarity

Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk	Luong et al. Rare Words	Hill et al. SimLex
PPMI	.755	.697	.745	.686	.462	.393
SVD	.793	.691	.778	.666	.514	.432
SGNS	.793	.685	.774	.693	.470	.438
GloVe	.725	.604	.729	.632	.403	.398

- SVD = singular value decomposition on PMI matrix
- GloVe does not appear to be the best when experiments are carefully controlled, but it depends on hyperparameters + these distinctions don't matter in practice

Levy et al. (2015)



Analogies

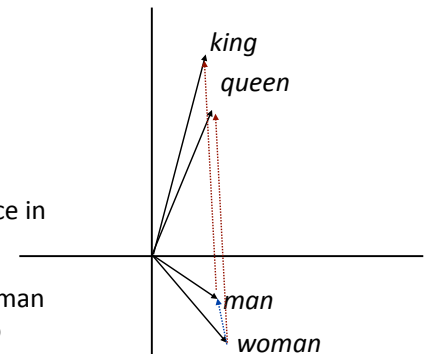
$(king - man) + woman = queen$

$king + (woman - man) = queen$

► Why would this be?

► woman - man captures the difference in the contexts that these occur in

► Dominant change: more "he" with man and "she" with woman — similar to difference between king and queen





Bias in Word Embeddings

- Identify *she* - *he* axis in word vector space, project words onto this axis

Extreme <i>she</i> occupations		
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme <i>he</i> occupations		
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Bolukbasi et al. (2016)

- Nearest neighbor of ($b - a + c$)

Racial Analogies	
black → homeless	caucasian → servicemen
caucasian → hillbilly	asian → suburban
asian → laborer	black → landowner

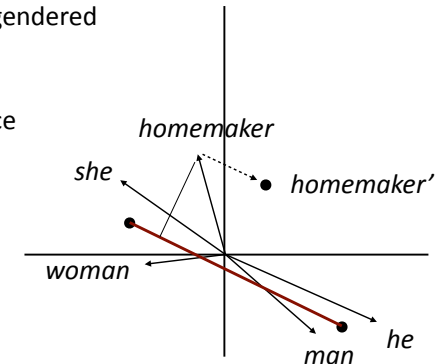
Religious Analogies	
jew → greedy	muslim → powerless
christian → familial	muslim → warzone
muslim → uneducated	christian → intellectually

Manzini et al. (2019)



Debiasing

- Identify gender subspace with gendered words
- Project words onto this subspace
- Subtract those projections from the original word

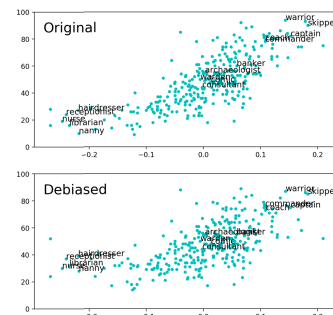


Bolukbasi et al. (2016)



Hardness of Debiasing

- Not that effective...and the male and female words are still clustered together
- Bias pervades the word embedding space and isn't just a local property of a few words



(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.

Gonen and Goldberg (2019)



Using Word Embeddings

- Approach 1: learn embeddings as parameters from your data
 - Often works pretty well, especially if data is large
- Approach 2: initialize using GloVe, keep fixed
 - Faster because no need to update these parameters
- Approach 3: initialize using GloVe, fine-tune
 - Usually works the best



Takeaways

- ▶ Continuous bag-of-words, Skip-gram, and Skip-gram with negative sampling are all similar ways to learn embeddings
- ▶ Matrix factorization approaches like GloVe are most standard
- ▶ Averaging inputs to feedforward networks can work well, will see other approaches later
- ▶ Later in the class: approaches to create “contextualized” word embeddings