WORDS IN SPACE

William Lowe

Hertie School

31st October 2020

THE SPACE OF WORDS

we have seen a lot of *document* spaces:

- → Rows in a *document feature matrix* as coordinates in a V space
- → Word frequency distributions, e.g. DFM rows normalized to 1, as coordinates in a (V-1) dimensional simplex
- → Topic counts as locations in a K-dimensional space
- → Topic proportions as location on a the corresponding K-1 dimensional simplex

Later, with scaling, we'll put documents into much smaller spaces, e.g. one or two dimensions.

But what about the words?

- → Do they live in a space?

DISTRIBUTIONAL HYPOTHESIS

Back in the mid-20th century *description* and *empirical data gathering* hit linguistics, arguably from Zipf (1932) but also from field linguistics (Harris, 1954).

Eventually this developed into *distributional semantics* (e.g. Cruse, 1986), which can be summarised informally as

> *You shall know a word by the company it keeps*
>
> *(Firth, 1968)*

This is the linguistics counterpart of Quine (1960) or Wittgenstein's (1958) advice: Don't look for the meaning, look for the use.

Operationally:

→ The meaning of a word is the sum total of all the ways it is used in actually existing language

CHARACTERIZING USE

A common test of language understanding: the sentence completion test. Fill in the missing word

```
All I want is a nice [       ] of tea
```

A common descriptive tool for word usage: the keyword in context

```
     that "the state shall not [discriminate] against, or grant preferential treatment
the lingering effects of racial [discrimination] against minority groups in this
 remedy the effects of societal [discrimination]. Another four Justices (Stevens
     that "the state shall not [discriminate] against, or grant preferential treatment
```

These are closely alternates:

- → By studying enough keywords in context you can learn to complete sentences
- → Words mean the same to the extent you can substitute them into the same contexts, c.f. 'bias'

Context vectors

One simple way to characterize all those lines of kwic is to just sum them up

A context-feature matrix C where C_{ij} is the number of times word j occurred in word i's kwics

Rows of C are 'context vectors' (Naturally, this depends on the window size)

Intuition: if two words are easily *intersubstitutable in context* then

- → their contexts share a lot of words
- → so their context vectors should be similar

Geometrically

- → Context vectors should point in the same direction
- → The angle between them should be small
- → the cosine of that angle should be positive

Caveats

But wait, you say. What if there is more than one sense of word *i*? e.g. 'bank' in English.

→ Mostly we just ignore it (and it's mostly fine…)

Unlike the unimportance of negation, I'm not so sure I have a good explanation for this

→ Senses do seem to be roughly power law distributed (Zipf, 1945), so maybe we just don't often need the infrequent ones for evaluation tasks?

Caveats

But wait, you say. What if there is more than one sense of word *i*? e.g. 'bank' in English.

→ Mostly we just ignore it (and it's mostly fine…)

Unlike the unimportance of negation, I'm not so sure I have a good explanation for this

→ Senses do seem to be roughly power law distributed (Zipf, 1945), so maybe we just don't often need the infrequent ones for evaluation tasks?

Some people think there are reasonable senses of similarity that are *asymmetric*

> *"No, Danish is not similar to German, German is similar to Danish"*
>
> *(Unknown Danish nationalist, circa 1998)*

We shall quietly ignore the non-symmetric senses of similarity because their 'distances' won't obey the triangle inequality.

SEMANTIC SPACE

Or as they call it these days: *word embedding*

Four approaches:

- → Just use C, possibly with transformed elements (Bullinaria & Levy, 2012)
- → Matrix factorizations of C, e.g. latent semantic indexing (LSI; Landauer & Dumais, 1997) or GloVe (Pennington et al., 2014)
- → Some intermediate representation extracted from a neural network, e.g. word2vec (Mikolov et al., 2013), though it's not so clear that the *neural network* part is doing the work

Let's dig into these after considering the kind of association we need…

Associations

The key distributional semantics concept *intersubstitutability in context* is a *second order* similarity / association

First order similarity/ direct association

'racial' and 'discrimination' are associated when they tend to occur together, e.g.

→ as measured by a *collocation* measure

Associations

The key distributional semantics concept *intersubstitutability in context* is a *second order* similarity / association

First order similarity/ direct association

'racial' and 'discrimination' are associated when they tend to occur together, e.g.

→ as measured by a *collocation* measure

Second order association

'bias' and 'discrimination' as associated because they have similar associations with 'racial'

→ two words are similar when they have the same *pattern of first order associations* with all the other words in the vocabulary

Associations

In simple context feature matrix, C_{ij} (how many times j occurred in i's contexts) is enough.

→ Maybe (as in LSI) we shrink C by Singular Value Decomposition

In fancier models, C_{ij} is replaced by a 'chance corrected' transformation like 'pointwise mutual information'

In a (unrealistic) one token window:

$$\text{PMI} = \frac{P(\text{racial}_n\ \text{discrimination}_{n+1})}{P(\text{racial})P(\text{discrimination})}$$

Note that by removing $P(\text{racial})$ from this we get an estimate of $P(\text{racial} \mid \text{discrimination})$

Associations

GloVe (Pennington et al., 2014) is a little more sophisticated and assumes that it is *relative usage* that drives association

$$\log \frac{P(\text{racial} \mid \text{discrimination})}{P(\text{racial} \mid \text{bias})}$$

but in a *proportional* way, so we log it (is this starting to seem familiar?)

Associations

GloVe (Pennington et al., 2014) is a little more sophisticated and assumes that it is *relative usage* that drives association

$$\log \frac{P(\text{racial} \mid \text{discrimination})}{P(\text{racial} \mid \text{bias})}$$

but in a *proportional* way, so we log it (is this starting to seem familiar?)

Although GloVe would not describe it this way we can think of it as choosing context vectors β_i, β_j to model the odds ratios underlying C

$$\log C_{ij} = a_i + b_j + \beta_i \beta_j$$

This is the reduced form of a log bilinear model of C (but fit with weighted least squares)

If we choose a $K < V$-dimensional space then we are claiming that facts about relative usage are less variable than the words

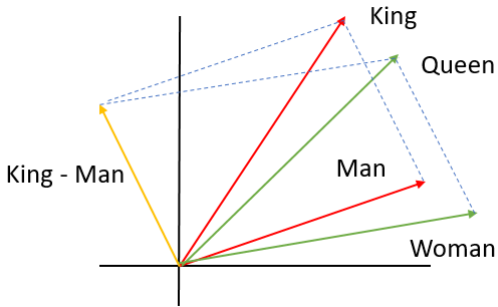
Vector similarity

OK, so we've got a bunch of word vectors one way or another

What can we do with them? Ideally

- → Learn about conceptual / semantic relationships
- → Learn about the world

Conceptual geometry



Famous vector addition example:

$$\text{King - Man + Woman} \approx \text{Queen}$$

Note this picture is (very) *schematic*. Embedding spaces are several hundred dimensional.

Bias

Caliskan et al. (2017) replicate ‘implicit bias’ using word embeddings

→ Experimental result: subjects pair two concepts they find similar faster than two concepts they find different. (Greenwald et al., 1998)

Operationalization in word embeddings: closer in space / more similar = shorter reaction time

Bias

Caliskan et al. (2017) replicate ‘implicit bias’ using word embeddings

→ Experimental result: subjects pair two concepts they find similar faster than two concepts they find different. (Greenwald et al., 1998)

Operationalization in word embeddings: closer in space / more similar = shorter reaction time

Implicit bias is controversial (bias, less so) but we don’t need to take a stand on the psychological theory

→ This is the usage!

Bias

Their measure ‘WEAT’ is computed in two steps (not how they describe it):

- → math average: cos(math words, male terms) - cos(math words to female terms)
- → arts average: cos(math words, male terms) - cos(math words to female terms)
- → math average - arts average

A ‘difference of differences’

Bias

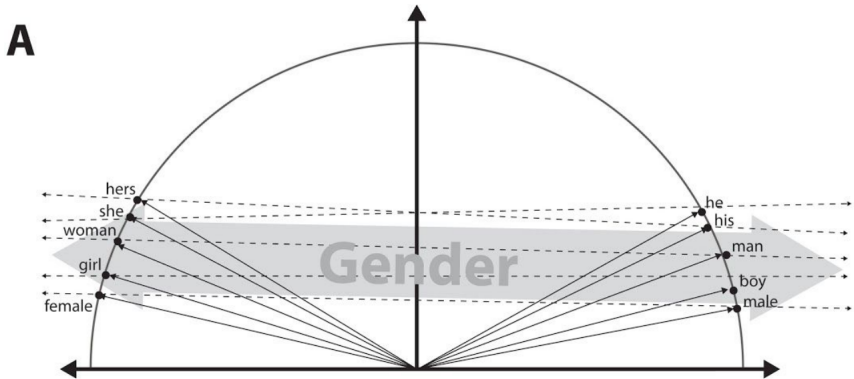
Their measure 'WEAT' is computed in two steps (not how they describe it):

- → math average: cos(math words, male terms) - cos(math words to female terms)
- → arts average: cos(math words, male terms) - cos(math words to female terms)
- → math average - arts average

A 'difference of differences'

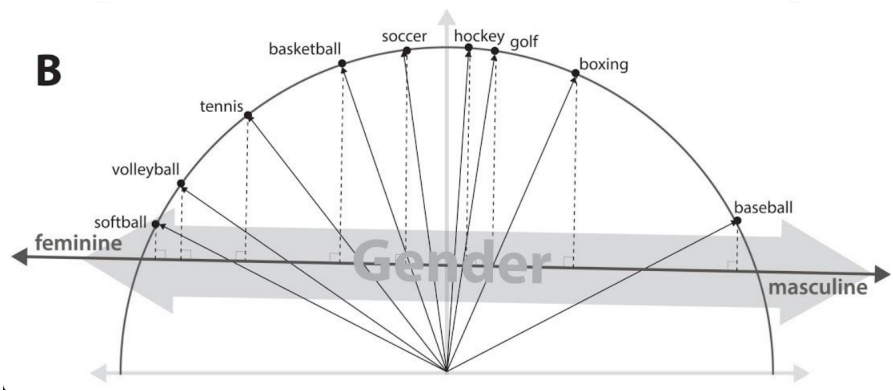
Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(5)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(5)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(5)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (5)	(7)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (9)	(7)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(9)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(9)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(9)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Gender projection



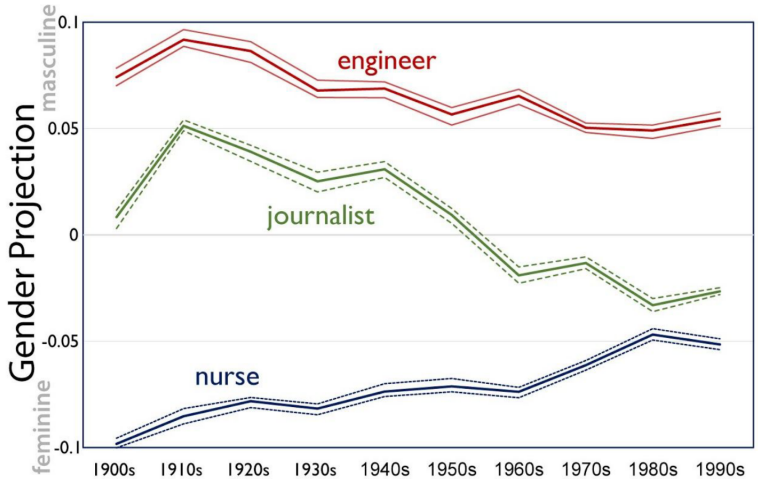
from Kozlowski et al. (2019)

Gender projection



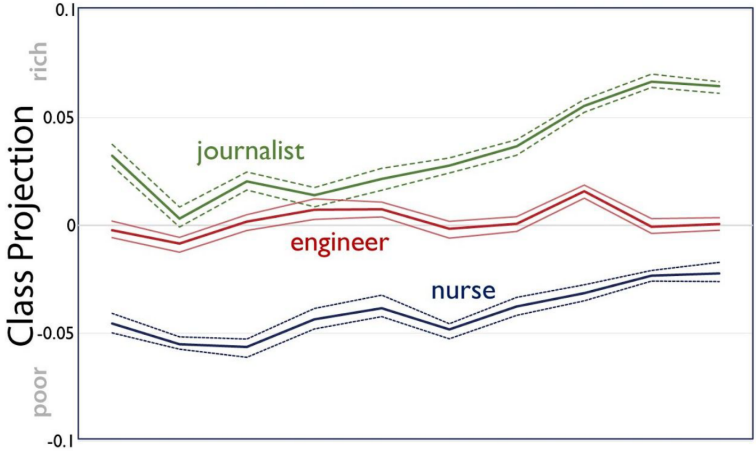
from Kozlowski et al. (2019)

Reflecting society



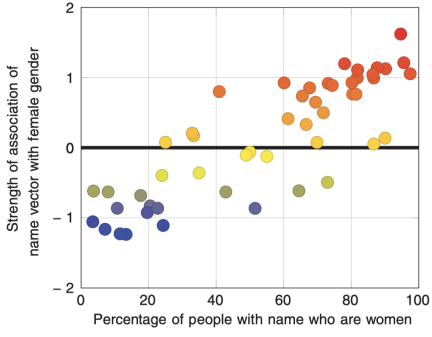
from Kozlowski et al. (2019)

Reflecting society



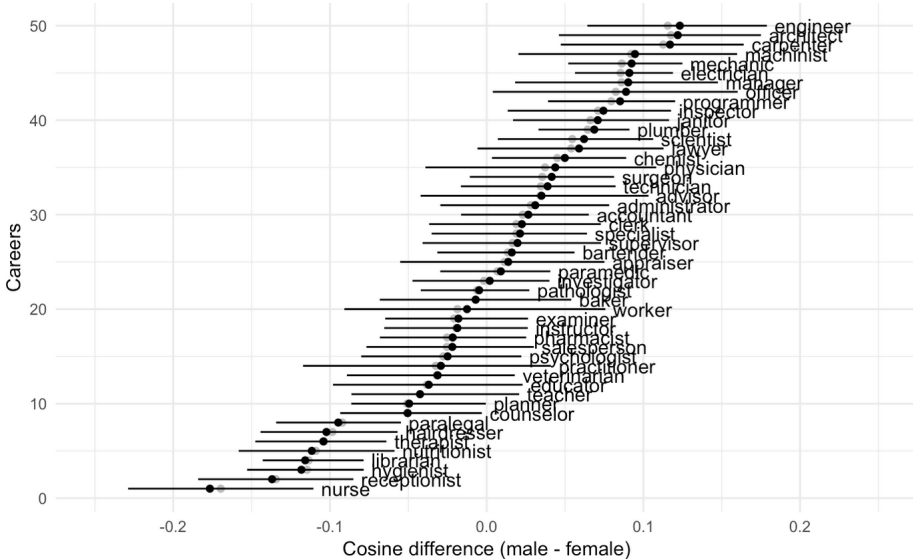
from Kozlowski et al. (2019)

Reflecting society

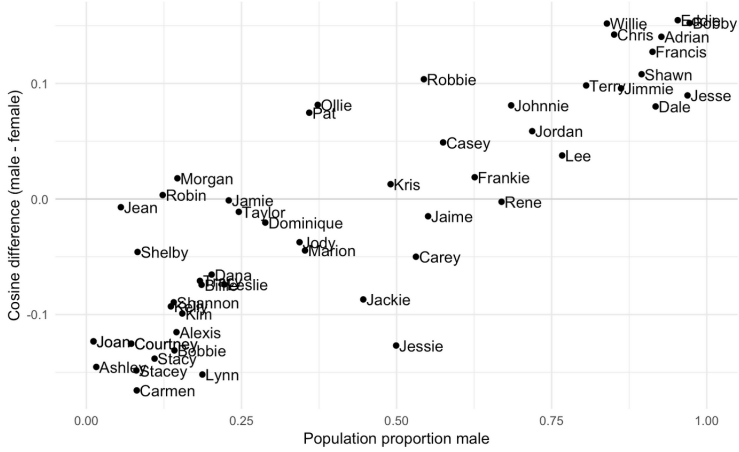


from Caliskan et al. (2017)

Reflecting society



Reflecting society



redrawn from from Caliskan et al. (2017)

Practicalities

Word embeddings can be computationally intensive

- → for serious work: gensim (in python; Hurek 2020 [link]) or text2vec (in R; Selivanov 2020 [link])
- → for testing things out: LSA [link]

or you can just use some *pre-computed* embeddings, e.g. from Google

- → GloVe makes available

Rodriguez and Spirling (2020) argue that this is usually just fine.

References

Bullinaria, J. A. & Levy, J. P. (2012). 'Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and svd'. *Behavior Research Methods*, *44*(3), 890–907.

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). 'Semantics derived automatically from language corpora contain human-like biases'. *Science*, *356*(6334), 183–186.

Cruse, D. A. (1986). 'Lexical semantics'. Cambridge University Press.

Firth, J. R. (1968). A synopsis of linguistic theory. In F. R. Palmer (Ed.), *Selected papers of j. r. firth: 1952-1959*. Longman.

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. (1998). 'Measuring individual differences in implicit cognition: The implicit association test. - psycnet'. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Harris, Z. S. (1954). 'Distributional structure'. *WORD*, *10*(2-3), 146–162.

References

Kozlowski, A. C., Taddy, M. & Evans, J. A. (2019). 'The geometry of culture: Analyzing meaning through word embeddings'. *American Sociological Review*, *84*(5), 905–949.

Landauer, T. K. & Dumais, S. T. (1997). 'A solution to plato's problem: The latent semantic analysis theory of induction and representation of knowledge'. *Psychological Review*, (104), 211–240.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013, September 6). *Efficient estimation of word representations in vector space*. Retrieved October 28, 2020, from http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R. & Manning, C. (2014). 'Glove: Global vectors for word representation'. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Quine, W. v. O. (1960). 'Word and object'. MIT Press.

References

Rodriguez, P. L. & Spirling, A. (2020). *Word embeddings: What works, what doesn't, and how to tell the difference for applied research*.

Wittgenstein, L. (1958). 'Philosophical investigations' (G. E. M. Anscombe, Trans.). Blackwell.

Zipf, G. K. (1932). 'Selected studies of the principle of relative frequency in language'. Oxford University Press.

Zipf, G. K. (1945). 'The meaning-frequency relationship of words'. *The Journal of General Psychology*, *33*(2), 251–256
_eprint: https://doi.org/10.1080/00221309.1945.10544509.