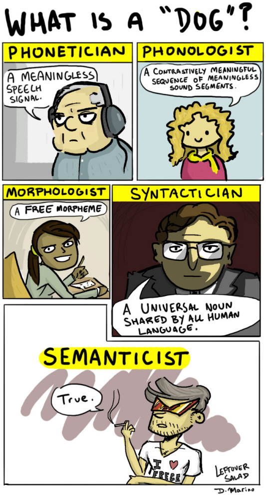
Using Word Vectors: Commit (Semantic) Crimes With Both Direction and Magnitude!

Esther Seyffarth


WHAT IS A "DOG"?
PHONETICIAN
A MEANINGLESS SPEECH SIGNAL.
PHONOLOGIST
A CONTRASTIVELY MEANINGFUL SEQUENCE OF MEANINGLESS SOUND SEGMENTS.
MORPHOLOGIST
A FREE MORPHEME
SYNTACTICIAN
A UNIVERSAL NOUN SHARED BY ALL HUMAN LANGUAGE.
SEMANTICIST
True.
I ♥ FREGE
LEFTOVER SALAD
D. Marino

Semantics!

$[\![\text{John}]\!] = \text{John}$

$[\![\text{works}]\!] = \text{f}: \text{D} \rightarrow \{0, 1\}$
For all x $\in$ D, f(x) = 1 iff x works

$[\![\text{smokes}]\!] = \text{f}: \text{D} \rightarrow \{0, 1\}$
For all x $\in$ D, f(x) = 1 iff x smokes

How about these?

- $[\![\text{smoke}]\!]$
- $[\![\text{fog}]\!]$
- $[\![\text{cloud}]\!]$

What do the meanings of these words have in common?
How do they differ?

Why do we do Semantics anyway?

- Sometimes, we want to know the ***true*** meaning of words or phrases: John means John, and nothing else.
- Other times, we look at how words interact in the real world:

(1) I can't breathe properly because of the ______ .

(2) I'm cold from all the ______ .

(3) I couldn't see anything due to the ______ .

(4) My clothes are wet from standing in the ______ for an hour.

Frames!

	smoke	**fog**	**cloud**
source:	fire	water	water
location:	anywhere	near the ground	in the sky
colour:	grey	white	grey or white
???:	...	...	...

Are some of these "more related" than others?
If so, why? How do we know?

Distributional Semantics

cloud

fog

smoke

Humans can annotate salient features of these words!

Some problems:

- Annotators must be paid.
- Annotation takes time.
- Annotators don't agree with each other.
- Annotators aren't experts for *everything*.
- Annotation is never finished.


345678

You shall know a word by the company it keeps?

Words that co-occur with our terms in a context window of 5 tokens to each side:

	smoke	**fog**	**cloud**
breathe	31	0	2
see	37	23	15
cold	0	29	11
fire	29	0	0
wet	6	24	14
white	70	19	19

But is it science if you just count words?

- No.
- The co-occurrence counts for "cloud" were lower than those for the other two terms – probably because we talk about clouds less often (in the corpus).
- We have to normalize the absolute co-occurrence counts with regard to how frequent each word is on its own. A good way to do that is *pointwise mutual information* (PMI).

You shall know a word by the relationships it commits to!

Similarity scores of the co-occurrences, where 1 is "identical" and 0 is "not at all related":

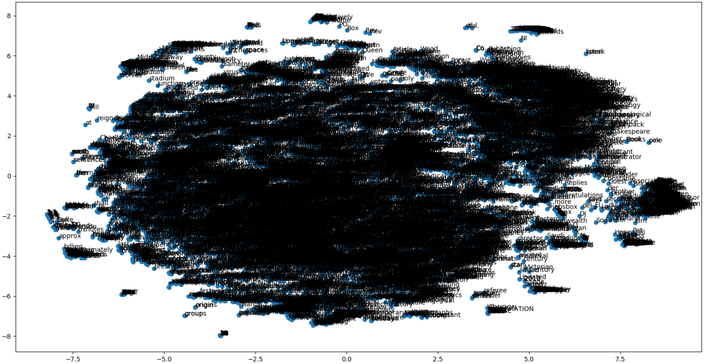
	smoke	**fog**	**cloud**
breathe	0.21	0	0.0014
see	0.40	0.39	0.39
cold	0	0.38	0.30
fire	0.40	0	0
wet	0.054	0.26	0.33
white	0.49	0.21	0.27

Why are "white" and "smoke" so similar?

- There can be conflicts between our distributional observations and the semantics that we believe to be true.
- We are fairly sure that smoke is *usually* grey. . .
- . . . but the only times people mention the color of smoke are when that color is remarkable; for instance, when electing a new pope.
- In general, "truisms" are rarely observed in the corpus, so we will miss some features of our terms!
- We say things that are unexpected more than we say things that are normal!

Vectors in the wild

- Let's look at a little demo with English word vectors that I trained on the `ukWaC` corpus for my MA thesis.
- If we have time, we can play around with the tool at `https://rare-technologies.com/word2vec-tutorial/#bonus_app` for a bit.

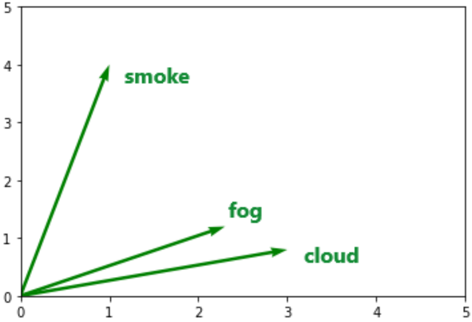

8
6
4
2
0
−2
−4
−6
−8
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

```
the -5.63 -4.51 -3.24  6.21 -0.82  0.28 -0.28 -2.90  1.39  4.43  1.93
 -2.25  0.46 -4.75  2.12  2.29 -2.66 -0.39 -0.19  2.00 -4.10 -6.15
 2.21 -7.79 -2.95 -0.14 -0.33 -1.23 -2.75  1.04  0.38  3.20 -0.60
 -0.70 -1.72 24.75 -5.54  6.52 -2.04  1.53 -0.54  0.27  6.14 -9.07

of -9.59  0.26  0.07  7.43 -1.46  1.13 -1.44 -4.10  2.77  1.02 -1.68
-0.95  1.19 -5.07  0.49 -1.34 -4.04 -4.55 -2.20 -0.14 -4.36 -4.50
-1.57 -2.03 -1.91  0.98  2.80 -3.08 -1.24  2.19 -2.42  8.83  0.20
-1.89 -1.05 22.73  7.05 -4.22 -2.91  4.65  3.26  5.98 -1.96 -4.30

and -8.41  1.02 -0.44  2.25  4.59  3.54 -0.42 -4.47  1.80  3.09
1.10 -0.26  2.43 -1.07  4.74 -4.08 -0.18  0.04  2.43  4.10 -5.34
-1.19  6.25 -2.81 -2.87 -5.20  8.16 -1.05 -2.26 -1.83  2.76  0.52
3.04 -7.30 -3.11 22.49  1.63 -3.19  1.10 -2.93 -0.79 -0.81  2.29
```

Distributional Semantics



Getting started with distributional semantics

- If you want to just use existing vectors, you can download pre-trained sets of them from the websites of the `word2vec` and `GloVe` projects.
- If you want to train your own vectors, you can download the code for `word2vec`/`GloVe` and run it on your own data – attention: you should probably run them on the HPC!
- If you just want to see some vector magic, you can check out the "Bonus App" mentioned above.
- To visualize your vectors, try this code by Vered Schwartz: `https://www.quora.com/How-do-I-visualise-word2vec-word-vectors`

What to read, what to cite

Introductory reading recommendations

- Jurafsky & Martin's *Speech and Language Processing* (`https://web.stanford.edu/~jurafsky/slp3/`) is a good, easy-to-follow introduction. Chapters 15 and 16 are especially relevant.

- Turney & Pantel's 2010 paper *From Frequency to Meaning: Vector Space Models of Semantics* (`http://jair.org/media/2934/live-2934-4846-jair.pdf`) is a more thorough primer on methods and theories around distributional semantics.

Useful practical references

- Levy, Goldberg & Dagan (2015): Improving Distributional Similarity with Lessons Learned from Word Embeddings (http://www.aclweb.org/anthology/Q15-1016)
- Bullinaria & Levy (2007): Extracting semantic representations from word co-occurrence statistics: A computational study (https://link.springer.com/content/pdf/10.3758%2FBF03193020.pdf)
- Bullinaria & Levy (2012): Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. (https://link.springer.com/content/pdf/10.3758%2Fs13428-pdf)

Other resources

- `word2vec` homepage: https://code.google.com/archive/p/word2vec/
- `GloVe` homepage: https://nlp.stanford.edu/projects/glove/
- High Performance Computing at HHU: https://www.zim.hhu.de/high-performance-computing.html

Okay, but who actually uses word vectors?

- People who do Machine Translation!
- People who do Discourse Relation Classification!
- People who do Information Retrieval!
- People who do Parsing!
- . . . and maybe you?

Questions?

Sources

What is a "Dog"? by David Marino (http://specgram.com/CLXXVI.4/03.marino.dog.html)

Heim & Kratzer (1998): Semantics in Generative Grammar

Count von Count @ IMDb (http://www.imdb.com/character/ch0000709/mediaindex)