Word Embeddings
Can Vectors Encode Meaning?

Katy Gero and Jeff Jacobs
Who are we?
Who are you?
Plan

1. **Theory** of using vectors to represent words (20 min)
2. **Practice** of creating embeddings (20 min)
3. **Applications** of embeddings (20 min)
4. **Pitfalls** and bias (20 min)
Theory:
Word representations
How to represent words in computing?
Dictionaries for computers, aka lexical resources

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: apricot  Search WordNet

Display Options:  (Select option to change)  F  Change

Key: "S." = Show Synset (semantic) relations, "W." = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun
- S. (n) apricot, apricot tree (Asian tree having clusters of usually white blossoms and edible fruit resembling the peach)
  - direct hypernym / full hypernym
    - S. (n) Japanese apricot, mei, Prunus mume (Japanese ornamental tree with fragrant white or pink blossoms and small yellow fruits)
    - S. (n) common apricot, Prunus armeniaca (temperate zone tree bearing downy yellow to rosy fruits)
    - S. (n) purple apricot, black apricot, Prunus dasycarpa (small hybrid apricot of Asia and Asia Minor having purplish twigs and white flowers following by inferior purple fruit)
  - member holonym
    - direct hypernym / inherited hypernym / sister term
    - S. (n) apricot (downy yellow to rosy-colored fruit resembling a small peach)
    - S. (n) yellowish pink, apricot, peach, salmon pink (a shade of pink tinged with yellow)
Visualization of ConceptNet

https://blog.conceptnet.io/tag/conceptnet/
Problems with lexical resources

1. Requires skilled people → time consuming to create
2. Personal judgements → prejudiced to the views of the creators
3. Representations are discrete → hard to share info between words

"...the WordNet team relied on existing lexicographic sources as well as on introspection.” Fellbaum 2010.
How do we know what a word means?

litofar
Does this help?

The hairy little litofar hid behind a tree.
The distributional hypothesis

The meaning of words can be discovered purely by the context in which they are used.

“You shall know a word by the company it keeps.” Firth, 1957

“Here we will discuss how each language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts (ultimately sounds) relative to other parts, and how this description is complete without intrusion of other features such as history or meaning.” Harris, 1954

A simple measure of context

The hairy little *litofar hid behind* a tree.
Related to, but not the same as, n-grams

The hairy little litofar hid behind a tree.

Bigrams:

- The hairy
- hairy little
- little litofar
- ...


Can we do better than n-gram counts?

![Bigram counts for eight of the words (out of V = 1446) in the Berkeley Restaurant Project corpus of 9332 sentences. Zero counts are in gray.](image)
Pointwise Mutual Information (PMI)

Compares the joint probability of seeing word_1 and word_2 together with the probability they are seen together by chance (based on how frequently they are seen separately)

\[ I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \]

Meaning is **distributed** across context

<table>
<thead>
<tr>
<th>Word</th>
<th>Context→</th>
<th>tree</th>
<th>hairy</th>
<th>...</th>
<th>plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td></td>
<td>5</td>
<td>3</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td></td>
<td>7</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>litofar</td>
<td></td>
<td>5</td>
<td>6</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

It’s like this but millions of rows and columns
Embeddings are vectors

- Continuous
- Represents meaning, not just uniqueness
- Can calculate similarity as the cosine (normalized distance)
- Can do other vector operations

<table>
<thead>
<tr>
<th></th>
<th>tree</th>
<th>hairy</th>
<th>...</th>
<th>plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>5</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>7</td>
<td>2</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>litofar</td>
<td>5</td>
<td>6</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

dog = [5,3,...,1]
Dimensionality reduction

SVD, PCA, LSA/I:

We can get to embeddings from PMI and other count-based measures with dimensionality reduction.
Now for something completely different...

Neural networks
Get your embeddings for free!

Try to predict this word:

- show
- me
- the
- litofar

Embedding layer (one big shared matrix) -> Hidden layer -> All possible words

Probability that a word is the next word
Just give me a good embedding...

Dear neural net,
please make this useful.
200 dimensions. Thanks.
What's your context?

<table>
<thead>
<tr>
<th>Word</th>
<th>Context→</th>
<th>?</th>
<th>?</th>
<th>...</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>5</td>
<td>3</td>
<td>...</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>7</td>
<td>2</td>
<td>...</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>litofar</td>
<td>5</td>
<td>6</td>
<td>...</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Count-based**: context is interpretable; other words or selected features

**Neural network**: context is learned
Practice:
The embedding layer of neural networks
Language model: predict the next word

show → embedding
me → embedding
the → embedding
money → embedding

Probability that a word is the next word

Try to predict this
embedding layer (one big shared matrix)
hidden layer all possible words
Embedding layer is learned implicitly

<table>
<thead>
<tr>
<th></th>
<th>me</th>
<th>show</th>
<th>litofar</th>
</tr>
</thead>
<tbody>
<tr>
<td>show</td>
<td>.5</td>
<td>.7</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.2</td>
<td>.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.1</td>
<td>.2</td>
<td>.0</td>
</tr>
</tbody>
</table>
Can we make this simpler?

More on efficiency: negative sampling

Don’t predict which word among all words in the vocabulary; instead predict from a set of words that includes the true words and some number of ‘noise’ words you draw from a distribution.

word2vec draws from a distribution of unigrams to the 3/4rd power. A similar thing goes on on GloVe. Magic? Or smoothing?

Both NCE and NEG have the noise distribution $P_n(w)$ as a free parameter. We investigated a number of choices for $P_n(w)$ and found that the unigram distribution $U(w)$ raised to the 3/4rd power (i.e., $U(w)^{3/4}/Z$) outperformed significantly the unigram and the uniform distributions, for both NCE and NEG on every task we tried including language modeling (not reported here).

Wait, but why neural networks and not the count-based stuff? (n-grams, PMI, etc.)
The popularity of word2vec

- Works better than deterministic methods
- Has a catchy name (and slogan) and feels kind of magical
- Pre-trained embeddings are made available
- Method is an exciting research area
- Improves the performance of neural network applications
Word embedding relations in 2D

Word embeddings as starbursts

Off the shelf or train yourself

Others have trained embeddings that you can download! e.g. Stanford has embeddings trained on 840 billion words from a web crawl (vocab size of 2.2 million)

You can get a subset of the word2vec embeddings from Python’s Natural Language Toolkit directly. (Or the whole thing from a web download.)

Or, you can download the code to train them yourself. Data and context matter.

- word2vec (Google)
- GloVe (Stanford)
- fastText (Facebook)
- NumberBatch (ConceptNet)
The effect of the window

The hairy little litofar hid behind a tree.

There is no reason the context window has to be a certain size, symmetric-- or rectangular.

Adding specific information

Add synonym/antonym information

Depends on your usage if you want antonyms to be similar or dissimilar.

<table>
<thead>
<tr>
<th>Before</th>
<th>east</th>
<th>expensive</th>
<th>British</th>
</tr>
</thead>
<tbody>
<tr>
<td>west</td>
<td>pricey</td>
<td>American</td>
<td></td>
</tr>
<tr>
<td>north</td>
<td>cheaper</td>
<td>Australian</td>
<td></td>
</tr>
<tr>
<td>south</td>
<td>costly</td>
<td>Britain</td>
<td></td>
</tr>
<tr>
<td>southeast</td>
<td>overpriced</td>
<td>European</td>
<td></td>
</tr>
<tr>
<td>northeast</td>
<td>inexpensive</td>
<td>England</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>After</th>
<th>eastward</th>
<th>costly</th>
<th>Brits</th>
</tr>
</thead>
<tbody>
<tr>
<td>eastern</td>
<td>pricy</td>
<td>London</td>
<td></td>
</tr>
<tr>
<td>easterly</td>
<td>overpriced</td>
<td>BBC</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>pricey</td>
<td>UK</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>afford</td>
<td>Britain</td>
<td></td>
</tr>
</tbody>
</table>

Evaluations are tricky

- Similarity rankings
- Word analogies
- Task specific
Applications in the digital humanities
“Diachronic” Word Embeddings

Multilingual Word Embeddings

Multilingual Word Embeddings

“Translation Mover’s Distance”

“Translation Mover’s Distance”

An experiment: Train a word2vec model on Jane Austen's books, then replace the nouns in P&P with the nearest word in that model. The graph shows a 2D t-SNE distance plot of the nouns in this book, original and replacement. Mouse over the blue words!

Chapter 1

It is a case universally acknowledged, that a single woman in defiance of a good sense, must be in use of a son.

However little known the feelings or views of such a woman may be on his first entering a spot, this case is so well fixed in the minds of the surrounding families, that he is considered the rightful sunderness of some one or other of their daughters.

"My dear Mr. Bennet," said his miss to him either morning, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she,"for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no apology.

"Do you not want to know who has taken it?" cried his son impatiently.
Pitfalls and bias
Man is to Computer Programmer as Woman is to Homemaker?

Doesn’t Diminish Performance!

<table>
<thead>
<tr>
<th></th>
<th>RG</th>
<th>WS</th>
<th>analogy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>62.3</td>
<td>54.5</td>
<td>57.0</td>
</tr>
<tr>
<td>Hard-debiased</td>
<td>62.4</td>
<td>54.1</td>
<td>57.0</td>
</tr>
<tr>
<td>Soft-debiased</td>
<td>62.4</td>
<td>54.2</td>
<td>56.8</td>
</tr>
</tbody>
</table>
Why Does it Matter? (Downstream Tasks)

But Training Data! (The Cop-Out)

**Ratio of Masculine to Feminine Pronouns in U.S. Books, 1900-2008**
Changes parallel increases in women’s labor force participation, education, age at first marriage, etc.

The ratio of masculine pronouns (“he,” “him,” “his,” “himself”) to feminine pronouns (“she,” “her,” “hers,” “herself”) peaked at over 4:1 in 1968. By 2000 the ratio dropped dramatically to 2:1 (Twenge et al., 2012).

Data from American English corpus of the Google Books database (~1.2 million books).
Reproduced from Twenge et al., 2012.

http://genderedinnovations.stanford.edu/case-studies/nlp.html
The Sapir-Whorf Hypothesis.

“The 'real world' is to a large extent unconsciously built upon the language habits of the group ... The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached... We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation.” -Sapir, 1958

“The world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds - and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way - an agreement that holds throughout our speech community and is codified in the patterns of our language.” -Whorf, 1940
Linguistic Battle Royale

THROUGH THE LANGUAGE GLASS
WHY THE WORLD LOOKS DIFFERENT IN OTHER LANGUAGES

GUY DEUTSCHER
AUTHOR OF THE UNFOLDING OF LANGUAGE

THE LANGUAGE HOAX
WHY THE WORLD LOOKS THE SAME IN ANY LANGUAGE

JOHN H. McWHORTER
“Man is to Computer Programmer...” Redux

“Debiased word embeddings can hopefully contribute to reducing gender bias in society. At the very least, machine learning should not be used to inadvertently amplify these biases.” (15)