How do we communicate with computers?

- If I say “What are the colors of a rainbow,” how does Siri know what I mean?
- How does Siri understand questions?

https://www.freemake.com/blog/siri-answers-20-hilarious-questions/
Can anyone guess the word?

“May the __________ be with you.”
Can anyone guess the word?

“There’s no _____ like home.”
“Don’t count the _____, make the days count.”
- Muhammad Ali
Can anyone guess the word?

“Hasta la _____, baby.”
What are word representations?
Why do we need word representations?
How do we make word representations?
• Represent each letter as a number
• How computers traditionally represent words
• Why might this not be useful?
What’s a vector?

- A vector is essentially a fixed-length list of numbers
- For example,
  - $A = [3, 1]$
  - $B = [1, -1]$
  - $A + B = [4, 0]$

https://mathinsight.org/vector_introduction
One-Hot Encodings

- Represent a word with a vector of ones and zeros
- Have a vocabulary (a set of words) called V
  - V has a size of S words
- Each vector will have one element for each word
- Elements are all 0 except for the word’s element which is one

Example:
V = {cat, dog, bird}, S = 3
cat = [1,0,0]
dog = [0,1,0]
bird = [0,0,1]
• Is there any meaning?
• Why is bird the same distance from cat as it is from dog?
• Shouldn’t cats and dogs be more related, or closer together, than birds?

Example:
\[ V = \{\text{cat, dog, bird}\} \]
\[ \text{cat} = [1,0,0] \]
\[ \text{dog} = [0,1,0] \]
\[ \text{bird} = [0,0,1] \]
• semantic definition: relating to meaning in language or logic.
• We want computers to understand words, not just represent them.
• How can we do this?
Neural Networks

- Neural networks take numbers as input and predict some output.
- There are ‘weights’ that are multiplied on each of the black lines.
- We can think of neural networks as a “function.”
- Our goal is to “teach” the neural network to give us outputs that we want.
Can we use written language to help improve representations?

The word2vec (word to vector) algorithm uses written text to learn.

Based on the distributional hypothesis:

“you shall know a word by the company it keeps” - Firth (1957)
word2vec algorithm - (Mikolov et al. 2013)
word2vec continued

https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c
word2vec algorithm - (Mikolov et al. 2013)

<table>
<thead>
<tr>
<th>Window Size</th>
<th>Text</th>
<th>Skip-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[The <strong>wide</strong> road shimmered] in the hot sun.</td>
<td>wide, the road, wide, shimmered</td>
</tr>
<tr>
<td></td>
<td>The [wide road <strong>shimmered</strong> in the ] hot sun.</td>
<td>shimmered, wide, road, shimmered, in</td>
</tr>
<tr>
<td></td>
<td>The wide road shimmered in [the hot <strong>sun</strong>].</td>
<td>sun, the sun, hot</td>
</tr>
<tr>
<td>3</td>
<td>[The <strong>wide</strong> road shimmered in ] the hot sun.</td>
<td>wide, the road, wide, shimmered in</td>
</tr>
<tr>
<td></td>
<td>[The wide road <strong>shimmered</strong> in the ] hot sun.</td>
<td>shimmered, the road, shimmered, in</td>
</tr>
<tr>
<td></td>
<td>The wide road shimmered [in the hot <strong>sun</strong>].</td>
<td>sun, in sun, the sun, hot</td>
</tr>
</tbody>
</table>

https://www.tensorflow.org/tutorials/text/word2vec
Euclidean Distance
(the distance formula)
Range: \([0, \infty)\):

\[ A = [3,1] \]
\[ B = [1,-1] \]

\[ \text{Dist}(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2} = \sqrt{(3 - 1)^2 + (1 - (-1))^2} = \sqrt{4 + 4} = \sqrt{8} \]
What’s another way to measure similarity?

Cosine Similarity
Range: \([-1, 1]\):

\[
A = [3,1] \\
B = [1,-1]
\]

\[
similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

For \(A = [3,1]\) and \(B = [1,-1]\):

\[
= \frac{3 \times 1 + 1 \times (-1)}{\sqrt{3^2 + 1^2} \sqrt{1^2 + (-1)^2}} = \frac{2}{\sqrt{10 \times 2}} = 0.447
\]


Only measures the similarity of direction!
Cosine similarity example

France and Italy are quite similar
\[ \theta \text{ is close to } 0^\circ \]
\[ \cos(\theta) \approx 1 \]

Ball and crocodile are not similar
\[ \theta \text{ is close to } 90^\circ \]
\[ \cos(\theta) \approx 0 \]

The two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)
\[ \theta \text{ is close to } 180^\circ \]
\[ \cos(\theta) \approx -1 \]

More word2vec

https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html
What can we do with these embeddings?

Question Answering

What can we do with these embeddings?

Sentiment Analysis
- e.g. Is a tweet happy or sad?

https://monkeylearn.com/sentiment-analysis/
What can we do with these embeddings?

Translation

¿Pueden las computadoras traducir el idioma? Sí.

Can computers translate the language? Yes.
What can we do with these embeddings?

Image Captioning

- A young boy is playing basketball.
- Two dogs play in the grass.
- A dog swims in the water.
- A little girl in a pink shirt is swinging.
- A group of people walking down a street.
- A group of women dressed in formal attire.
- Two children play in the water.
- A dog jumps over a hurdle.

https://github.com/danieljl/keras-image-captioning
Where are we now? BERT (Devlin et al. 2018)

Bidirectional Encoder Representations from Transformers (BERT)
BERT uses attention

https://observablehq.com/@clpuc/analyzing-the-design-space-for-visualizing-neural-attention
What’s next?

• Improving Siri’s ability to learn what you mean
  • Can Siri learn things specifically to communicate with you and the way you talk?
• Can we integrate pictures better with language?
• Can we understand entire books?
• How can we learn irony and sarcasm?
• How can we learn slang words quickly without examples?
  • Right now we need billions of examples!
Does anyone have any questions?