NLP

Topics

- Some interesting (according to me) ideas from NLP

Word Representations

Goals

- A vector for words (given a vocabulary)
- Potentially useful for a variety of tasks
  - How?

Word2Vec

- Word context
  - Surrounding words
- Similar context = similar meaning

Skip-Gram

- Key task:
  - Given word, predict word context

Context

- Say 2 words to the left + 2 words to the right
  - The quick brown fox jumps over the lazy dog

- The quick brown fox jumps over the lazy dog
- The quick brown fox jumps over the lazy dog
- The quick brown fox jumps over the lazy dog
- The quick brown fox jumps over the lazy dog
- The quick brown fox jumps over the lazy dog

Task

- Given word in the middle,
- Predict the words in the context

Ideas?

Classification Problem

- Input:
  - One hot word
- Output:
  - Yes/No over vocabulary

Network Shape

- V = Size of vocab
- 300 = word vector size (chosen)

Architecture

- `Dense(300, input_shape=(V,))`
- `Softmax(V)`

300 Dim Vector

- Called word vector
- You get some semantic properties from it
- Used heavily downstream

Speedup

- Delete words from corpus with some probability during training
- High frequency words like “the”, “and” more likely to be deleted

Speedup

- Negative sampling
- Small # of “negative” words selected during training
- Sample some “negatives” from corpus

Vector Arithmetic

Resulting word vectors show some interesting phenomena

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

King - man + woman =

Queen

Is this surprising

- I don’t think so
- https://levyomer.files.wordpress.com/2014/04/linguistic-regularities-in-sparse-and-explicit-word-representations-conll-2014.pdf
- ^ Not that big a deal

Typical Uses

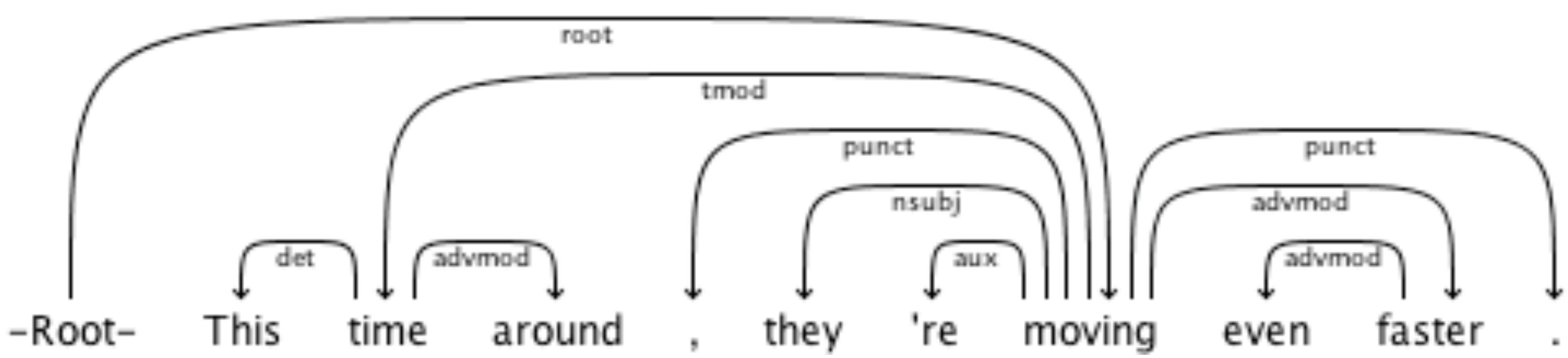
- You can take these vectors and use them elsewhere
- Instead of one-hot words, use this dense vector
- Smaller model etc. etc.

Traditional ML

- Word vectors fit in well with CRFs, SVMs etc.

Other Contexts

- Dependency parse


root
tmod
punct
nsubj
punct
advmod
det
advmod
aux
advmod
-Root- This time around , they 're moving even faster .

Interesting Research Directions

- https://arxiv.org/abs/1607.06520
- Fairness, Transparency etc.

Querying Word2Vec

Extreme *she* 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 *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Or

- Automated Inference on Criminality Using Face Images
- Rightfully panned
- https://arxiv.org/abs/1611.04135

Moral

- AI < Human stupidity
- Modeling includes data-collection, pre-processing, model choice, etc.
- Use your brain