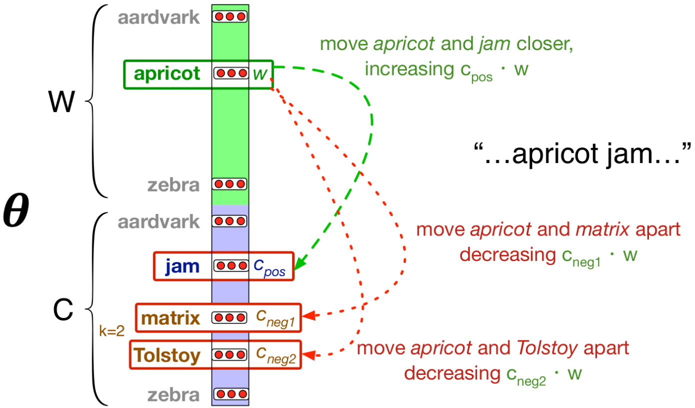
Word embeddings (part 2)

Paweł Rychlikowski

Instytut Informatyki UWr

30 kwietnia 2022

Intuition of one step of gradient descent



The skip-gram model with negative sampling (HW2)

- Notation more similar to class and HW2:

$$J_{neg-sample}(\boldsymbol{u}_o, \boldsymbol{v}_c, U) = -\log \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) - \sum_{k \in \{K\ sampled\ indices\}} \log \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)$$

- We take *k* negative samples (using word probabilities)
- Maximize probability that real outside word appears;
  minimize probability that random words appear around center word

- Sample with P(*w*)=U(*w*)$^{3/4}$/Z, the unigram distribution U(w) raised to the 3/4 power (We provide this function in the starter code).
- The power makes less frequent words be sampled more often

11

Word2Vec training details

- Linear learning rate decay
- Window size ≈ 10
  - smaller window – more syntactic relations
  - bigger window – more semantic
- 3-6 epochs
- Starting learning rate $= 0.003$

Word2Vec training details

- Linear learning rate decay
- Window size ≈ 10
  - smaller window – more syntactic relations
  - bigger window – more semantic
- 3-6 epochs
- Starting learning rate $= 0.003$

Sample negative words with $P'(w) \sim \text{cnt}(w)^{0.75}$

Not only words!

We can apply the same algorithm to different objects:

- In Wikipedia: $P(\text{pointer to doc}_j|\text{doc}_i)$
- For characters: $P(c_j|c_i)$ (discovers vowels?)
- For recomendation systems: $P(\text{product}|\text{customer})$

Not only words!

We can apply the same algorithm to different objects:

- In Wikipedia: $P(\text{pointer to doc}_j|\text{doc}_i)$
- For characters: $P(c_j|c_i)$ (discovers vowels?)
- For recomendation systems: $P(\text{product}|\text{customer})$

Quiz

What is the meaning of:

- doc2vec
- node2vec
- import2vec
- code2vec
- dna2vec
- wave2vec

Common phrases

Common phrases can be treated as words!

Common phrases

Common phrases can be treated as words!

In original paper very simple strategy was implemented:

1. Find valuable, common bigrams, replace them by a new word
   - New York $\rightarrow$ New_York
2. Repeat!

Common phrases

Common phrases can be treated as words!

In original paper very simple strategy was implemented:

1. Find valuable, common bigrams, replace them by a new word
   - New York $\rightarrow$ New_York
2. Repeat!

Bigram quality:

$$\text{score}(w_i, w_j) = \frac{count(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

where $\delta \approx 0.5$

Common phrases

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

How to use word2vec?

Best option: use **gensim** library

How to use word2vec?

Best option: use **gensim** library

1. Task 1: train vectors
2. Task 2: test vectors
3. Task 3: work with pretrained vectors

Let's test it in a notebook!

5. How to evaluate word vectors?

- Related to general evaluation in NLP: Intrinsic vs. extrinsic
- Intrinsic:
  - Evaluation on a specific/intermediate subtask
  - Fast to compute
  - Helps to understand that system
  - Not clear if really helpful unless correlation to real task is established
- Extrinsic:
  - Evaluation on a real task
  - Can take a long time to compute accuracy
  - Unclear if the subsystem is the problem or its interaction or other subsystems
  - If replacing exactly one subsystem with another improves accuracy → Winning!

GloVe introduction

- GloVe: Global Vectors for Word Representation, Jeffrey Pennington, Richard Socher, Christopher D. Manning

GloVe introduction

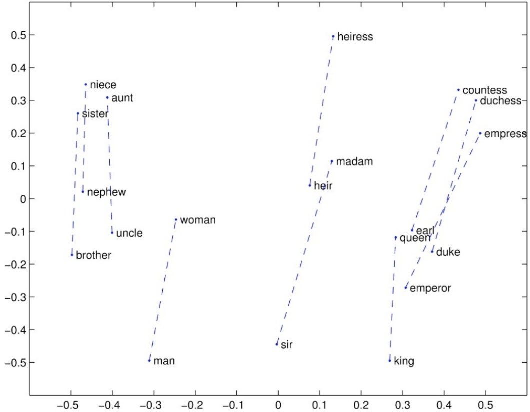
- GloVe: Global Vectors for Word Representation, Jeffrey Pennington, Richard Socher, Christopher D. Manning
- Semantic vectors via occurences statistics

GloVe introduction

- GloVe: Global Vectors for Word Representation, Jeffrey Pennington, Richard Socher, Christopher D. Manning
- Semantic vectors via occurences statistics

Quite similar results to word2vec, both methods are popular, and stil used

Glove Visualizations



28

Intrinsic word vector evaluation

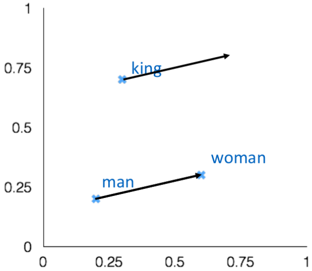
- Word Vector Analogies

a:b :: c:?

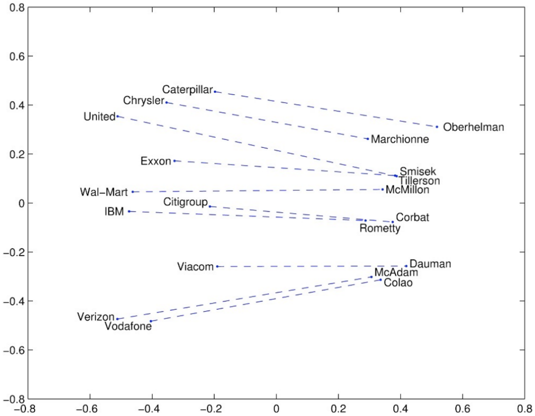
$$d = \arg\max_i \frac{(x_b - x_a + x_c)^T x_i}{||x_b - x_a + x_c||}$$

man:woman :: king:?

- Evaluate word vectors by how well their cosine distance after addition captures intuitive semantic and syntactic analogy questions
- Discarding the input words from the search (!)
- Problem: What if the information is there but not linear?

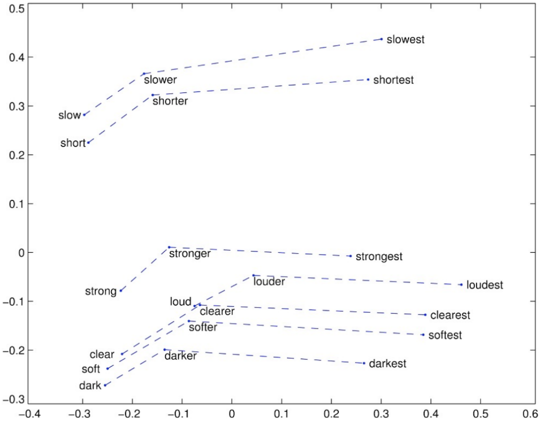


Glove Visualizations: Company - CEO



29

Glove Visualizations: Comparatives and Superlatives



30

Why it works?

Suppose we have that:

- $w_1, w_2, \ldots, w_k$ are (typical) contexts for women
- $m_1, m_2, \ldots, m_k$ are (typical) contexts for men
- $r_1, r_2, \ldots, r_k$ are (typical) contexts for medieval ruler

Why it works?

Suppose we have that:

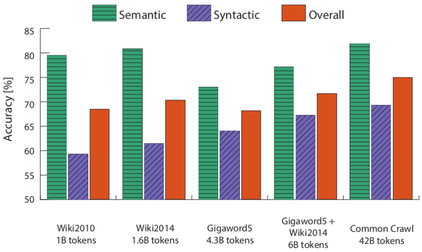
- $w_1, w_2, \ldots, w_k$ are (typical) contexts for women
- $m_1, m_2, \ldots, m_k$ are (typical) contexts for men
- $r_1, r_2, \ldots, r_k$ are (typical) contexts for medieval ruler

The training objectives

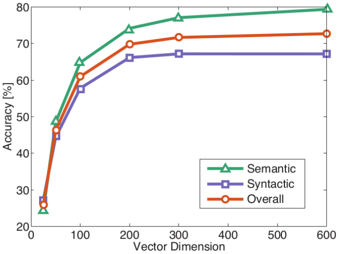
- Make vec('king') similar to $\sum_i con(m_i) + \sum_i con(r_i)$
- Make vec('queen') similar to $\sum_i con(w_i) + \sum_i con(r_i)$
- Make vec('man') similar to $\sum_i con(m_i)$
- Make vec('woman') similar to $\sum_i con(m_i)$

Analogy evaluation and hyperparameters

- More data helps
- Wikipedia is better than news text!



- Dimensionality
- Good dimension is ~300



32

Another intrinsic word vector evaluation

- Word vector distances and their correlation with human judgments
- Example dataset: WordSim353 http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10
book	paper	7.46
computer	internet	7.58
plane	car	5.77
professor	doctor	6.62
stock	phone	1.62
stock	CD	1.31
stock	jaguar	0.92

34

wordsim353

Definition

$\cos(u, v) = \frac{u \cdot v}{|u||v|}$

wordsim353

Definition

$\cos(u, v) = \frac{u \cdot v}{|u||v|}$

How to test vectors using human annotated pairs:

- Compute rank correlation between cosine similarity of vectors and mean similarities
- Do **ABX** tests

Language model and its applications

- Language model: $P(w_t | w_0 w_1 \dots w_{t-1})$
- LM can be used for sampling sentences

Some (funny) examples of a large LM by OpenAI – **GPT-2**

GPT-2 as question answerer

Q: How many wives did Henry VIII have? A: Henry VII had eight. Edward VI had two and James I had two. William and Mary each had three. Henry VIII was a bonapartist, his first wife only stayed married to him for two years and she was just

GPT-2 as question answerer

Q: How many wives did Henry VIII have? A: There were six in all, including the Church of England's Lady Jane Grey, his third wife.
Q: Who were the original Spice Girls? A: Victoria Beckham, Melanie Brown, Geri Halliwell, Emma Bunton and

GPT-2 as question answerer

Q: How many wives did Henry VIII have? A: None; but he had sixteen mistresses.
Q: What was the title given to Anne Boleyn? A: Duchess of Richmond and Lennox; and Cardinal Reginald Pole's daughter.
Q: Why did Henry VIII have so many