Geometries of Word Embeddings

Pramod Viswanath

University of Illinois
Natural language processing is widely used in daily life.
Natural language processing pipeline

- Segment sentences to sequences of **words**
  - **Tokenizer**

- **Morphological analysis**
  - Understand the meaning of **words**

- **Syntactic and Semantic Parser**
  - Understand the structures of **word** sequence

- ......

**Word is the basic unit of natural language.**
Representing Words

- **Atomic** symbols
  - Large vocabulary size (~1,000,000 words in English)
  - Joint distributions impossible to infer

Words could be represented by vectors.
Word Vector Representations

• **Word2Vec** (2013)
  - Google
  - Publicly available

• **GloVe** (2014)
  - Stanford NLP Pipeline
  - Publicly available
Principle of Word Vector Representations

“A word is characterized by the company it keeps.”

— Firth ‘57

She tried to speak write to him about his drinking.

Similar words should have similar vector representations.
Cooccurrence matrix

A series of many genres, including fantasy, drama, coming of age,…

(series, genres)
(of, genres)
(many, genres)
(including, genres)
(fantasy, genres)
(drama, genres)
PMI matrix is low rank

word2vec (Mikolov ’13) and GloVe (Pennington ’14)

target word $u(w)$  context word $v(c)$

$$u(w)^T v(c) \approx \log \left( \frac{p_{W,C}(w,c)}{p_W(w)p_C(c)} \right)$$
Word Similarity

\[ \text{sim}(w_1, w_2) \overset{\text{def}}{=} \frac{u(w_1)^T u(w_2)}{\|u(w_1)\| \|u(w_2)\|} \]
Powerful Representations

Lexical

✓ Word Similarity
✓ Concept Categorization
✓ Vector differences encode rules

\[
\text{talk} - \text{talking} = \text{eat} - \text{eating}
\]

\[
\text{man} - \text{king} = \text{woman} - \text{queen}
\]

\[
\text{France} - \text{Paris} = \text{Italy} - \text{Rome}
\]
This talk: Geometry of Word Vectors

- isotropy of word vectors
  - projection towards isotropy

- subspace representations of sentences/phrases
  - polysemy (prepositions)
  - idiomatic/sarcastic usages
Isotropy and Word Vectors

• Start with off-the-shelf vectors
  • Word2Vec and GloVe
  • Publicly available

• Postprocessing
  • Simple
  • Universally improves representations
Non-zero mean may affect the similarity between words.
Spectrum of word vectors
Postprocessing

- **Remove the non-zero mean**

  \[ \mu \leftarrow \frac{1}{|V|} \sum_{w \in V} v(w); \quad \tilde{v}(w) \leftarrow v(w) - \mu \]

- **Null the dominating \( D \) components**

  \[ u_1, \ldots, u_d \leftarrow \text{PCA}(\{\tilde{v}(w), w \in V\}) \]

  \[ v'(w) \leftarrow \tilde{v} - \sum_{i=1}^{D} (u_i^T v(w)) u_i \]

Renders off-the-shelf representations even stronger
Lexical-level Evaluation

✓ Word Similarity

✓ Concept Categorization
Word Similarity

Assign a similarity score between a pair of words

(stock, phone) -> 1.62
(stock, market) -> 8.08

Datasets: RG65, wordSim-353, Rare Words, MEN, MTurk, SimLex-999, SimVerb-3500.
Concept Categorization

Group words into different semantic categories.

Datasets: ap, ESSLLI, battig

Bear allocation airstream bull cat allotment blast cow drizzle credit puppy quota clemency

avg. improvement

word2vec 7.5%

GloVe 0.6%
Sentence-level Evaluation

✓ Sentential Textual Similarity (STS) 2012-2016

- 21 Different datasets: pairs of sentences
  - algorithm rates similarity
  - compare to human scores

- Average improvement of 4%
Postprocessing Generalizes

- Multiple dimensions, different hyperparameters
  - Word2Vec and GloVe
  - TSCCA and RAND-WALK
- Multiple languages
  - Spanish, German datasets
- Universally improves representations
Top Dimensions Encode Frequency

SKIP-GRAM

GLOVE

CBOW

$\alpha_1(w)$

$\alpha_2(w)$

$10^7$

$10^6$

$10^5$

$10^4$

$10^3$

$10^2$
RAND-WALK model

\[ p_{W,C}(w, c) = \frac{1}{Z_0} \exp \left( \|v(w) + v(c)\|^2 \right) \]

vectors \( v(w) \) are isotropic (Arora et al, ’16)

PMI matrix is low-rank

\[ \log \frac{p_{W,C}(w, c)}{p_W(w)p_C(c)} \propto v(w)^T v(c) \]
Post-processing and Isotropy

Measure of isotropy

\[
\frac{\min_{\|x\|=1} \sum_w \exp(x^T v(w))}{\max_{\|x\|=1} \sum_w \exp(x^T v(w))}
\]

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec</td>
<td>0.7</td>
<td>0.95</td>
</tr>
<tr>
<td>GloVe</td>
<td>0.065</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Rounding to Isotropy

- **First order** approximation of isotropy measure
  - subtract the mean
- **Second order** approximation of isotropy measure
  - project away the top dimensions [S. Oh]
- Inherently different
  - recommendation systems, [Bullinaria and Levy, ‘02]
  - CCA, Perron-Frobenius theorem
Summary

- Word Vector Representations
  - Off-the-shelf — Word2Vec and GloVe

- We improve them universally
  - Angular symmetry

- Other geometries?
Sentence Representations
What to preserve?

- Syntax information
  - grammar, parsing
- Paraphrasing
  - machine translation
- Downstream applications
  - text classification

This movie was funny and witty

Classifier

Classifier
Representation by Vectors

• Bag-of-words
  • frequency, tf-idf weighted frequency

• Average of word vectors:

• Neural networks:
Low rank Subspace

“A piece of bread, which is big, is having butter spread upon it by a man.”

Sentence word representations lie in a low-rank subspace
rank $N = 4$
Sentence as a Subspace

• **Input:** a sequence of words $\{v(w), w \in s\}$

• Compute the first $N$ principal components

  $$u_1, \ldots, u_N \leftarrow \text{PCA}(v(w), w \in s),$$
  $$S \leftarrow [u_1, \ldots, u_N].$$

• **Output:** orthonormal basis [Mu, Bhat and V, ACL ’17]
Similarity between Sentences

\[ \text{CosSim}(s_1, s_2) = \frac{1}{N} d(S_1, S_2) \]

\[ \overset{\Delta}{=} \frac{1}{N} \sqrt{\text{tr}(S_1 S_1^T S_2 S_2^T)} \]
<table>
<thead>
<tr>
<th>sentence pair</th>
<th>Ground Truth</th>
<th>Predicted Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man is doing exercises.</td>
<td>0.78</td>
<td>0.82</td>
</tr>
<tr>
<td>The man is training.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two men are hugging.</td>
<td>0.28</td>
<td>0.38</td>
</tr>
<tr>
<td>Two men are fighting.</td>
<td>0.4</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Examples
Collaborators

Hongyu Gong  Jiaqi Mu  Suma Bhat