Geometries of Word Embeddings

Pramod Viswanath

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

(target words)

<table>
<thead>
<tr>
<th>context words</th>
<th>…</th>
<th>genres</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>series</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>of</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>many</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>including</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>fantasy</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>drama</td>
<td>…</td>
<td>+1</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
PMI matrix is low rank

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

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

$$u(w)^T v(c) \approx \log \left( \frac{p_{W,C}(w, c)}{p_W(w)p_C(c)} \right)$$
\[
\text{sim}(w_1, w_2) \overset{\text{def}}{=} \frac{u(w_1)^T u(w_2)}{\|u(w_1)\| \|u(w_2)\|}
\]
Powerful Representations

Lexical

✓ Word Similarity
✓ Concept Categorization
✓ Vector differences encode rules

\[
talk - talking = eat - eating
\]
\[
man - king = woman - queen
\]
\[
France - Paris = Italy - Rome
\]
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

<table>
<thead>
<tr>
<th></th>
<th>avg. norm</th>
<th>norm of avg.</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WORD2VEC</strong></td>
<td>2.04</td>
<td>0.69</td>
<td><strong>0.34</strong></td>
</tr>
<tr>
<td><strong>GLOVE</strong></td>
<td>8.30</td>
<td>3.15</td>
<td><strong>0.37</strong></td>
</tr>
</tbody>
</table>
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
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%
Sentiment Analysis

This movie was funny and witty

Classifier

Postprocessing in downstream applications
Recurrent neural networks

Output

RNN

Input

\[ v(w_1) \quad v(w_2) \quad \cdots \quad v(w_T) \]

\[ h_1 \quad h_2 \quad \cdots \quad h_T \]

nonlinear unit

softmax

\[ \hat{y} \]
Experiment results

Similar Performance
Retrograde Analysis

before

\[ \hat{y}_{\text{before}} \]

\[
\{ v(w_t) - b_{\text{before}} \}_{t=1}^{T}
\]

\[
\{ \nu(w_t) \}_{t=1}^{T}
\]

\[
\text{RNN}
\]

\[
\text{linear bias}
\]

after

\[ \hat{y}_{\text{after}} \]

\[
\{ v(w_t) - \mu - b_{\text{after}} \}_{t=1}^{T}
\]

\[
\{ v(w_t) - \mu \}_{t=1}^{T}
\]

\[
\text{RNN}
\]

\[
\text{linear bias}
\]
Linear bias == mean vector

\[ b_{\text{before}} + \mu \quad \text{and} \quad b_{\text{after}} \]

highly correlated

avg. cosine similarity 0.69

Neural networks automatically learn to postprocess
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
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
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**

\[
\begin{align*}
  u_1, \ldots, u_N & \leftarrow \text{PCA}(v(w), w \in s), \\
  S & \leftarrow [u_1, \ldots, u_N].
\end{align*}
\]

- **Output**: orthonormal basis \([\text{Mu, Bhat and V, ACL ’17}]\)
Similarity between Sentences

sentence $s_1$ -> subspace $S_1$ -> similarity

sentence $s_2$ -> subspace $S_2$

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

$\triangleq \frac{1}{N} \sqrt{\text{tr}(S_1S_1^TS_2S_2^T)}$
## Examples

<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>The man is doing exercises.</td>
<td>0.28</td>
<td>0.38</td>
</tr>
<tr>
<td>Two men are hugging.</td>
<td>0.4</td>
<td>0.43</td>
</tr>
<tr>
<td>Two men are fighting.</td>
<td>0.4</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Semantic Textual Similarity Task

- Skip-thought
- Siamese CBOW
- doc2vec
- average
- subspace

2015.answ5-studetns
Sense Disambiguation
Polysemous Nature of Words

“crane”
Sense Representation

- **supervised**: aided by hand-crafted lexical resources
  - example: WordNet

- **unsupervised**: by inferring the senses directly from text
Disambiguation via Context

- (machine) The little prefabricated hut was lifted away by a huge crane.

- (bird) The sandhill crane (``Grus canadensis'') is a species of large crane of North America and extreme northeastern siberia.
Context Representation by Subspaces
Monosemous Intersection Hypothesis

The target word vector should reside in the intersection of all subspaces
Recovering the Intersection

- **Input**: a set of context \( \{c\} \), the target word \( w \)
- **context representations** \( \{S(c \setminus w)\} \)
- **Output**: recover the vector that is “closest” to all subspaces

\[
\hat{u}(w) = \arg \min_{\|u\| = 1} \sum_{w \in c} d(u, S(c \setminus w))^2
\]

\[
= \arg \min_{\|u\| = 1} \sum_{w \in c} \sum_{n=1}^{N} (u^T u_n(c \setminus w))^2
\]

rank-1 PCA of \( \{u_n(c \setminus w)\}_{c,n=1,...,N} \)
Polysemous Intersection Hypothesis

“crane”

The context subspaces of a polysemous word intersect at different directions for different senses.
Sense Induction

- **Input:** Given a target polysemous word \( w \)
  - contexts \( c_1, \ldots, c_M \)
    - number indicating the number of senses \( K \)

- **Output:** partition the \( M \) contexts into \( K \) sets \( S_1, \ldots, S_K \)

\[
\min_{u_1, \ldots, u_K, S_1, \ldots, S_K} \sum_{k=1}^{K} \sum_{c \in S_k} d^2(u_k, S(c \setminus w)).
\]
K-Grassmeans

- **Initialization:** randomly initialize $K$ unit-length vectors $u_1, \ldots, u_K$

- **Expectation:** group contexts based on the distance to each intersection direction

  $$S_k \leftarrow \{c_m : d(u_k, S(c_m \setminus w)) \leq d(u_{k'}, S(c_m \setminus w)) \forall k'\}, \forall k.$$ 

- **Maximization:** update the intersection direction for each group based on the contexts in the group.

  $$u_k \leftarrow \arg\min_u \sum_{c \in S_k} d^2(u, S(c \setminus w))$$
Sense Disambiguation

• **Input**: Given a new context instance for a polysemous word

• **Output**: identify which sense this word means in the context.

Can you hear me? You're on the **air**. One of the great moments of live television, isn't it?
Soft & Hard Decoding

- **Soft Decoding**: output a probability distribution

  \[
  P(w, c, k) = \frac{\exp(-d(u_k(w), S(c \setminus w)))}{\sum_{k'} \exp(-d(u_{k'}(w), S(c \setminus w)))}
  \]

- **Hard Decoding**: output a deterministic classification

  \[
  k^* = \arg \min_k d(u_k(w), S(c \setminus w))
  \]
SemEval Share Tasks

V-measure

F-score

[Mu, Bhat and V, ICLR ’17]
Sense Representation
Sense-specific Embeddings

- Our approach:
  - first disambiguate senses on a large Wiki corpus
  - successive decoding
  - lexical-specific
  - re-run word2vec

- [Arora, et. al. ’16]: dictionary learning
  - global fit for all words; hierarchical topic discovery
Polysemy Dataset
Arora et al. 2016

Sense No.1
main argument points sketchy
book summary index synopsis

abstract

 Sense No.5
mammal primates tail short
jungle animal primate gorilla

affect

 Sense No.7
every aspect uncritical behavior
mimic follow mime imitator

crane

 Sense No.77
slow human whole shoulder leg finger
limb biceps

ape

 Sense No.83
following preceding eighth year
month summer july calendar

appendix

 Sense No.110
separated embodiment mind
existing abstruse concept idea
profound

...
Police Lineup Task

Given a target word (e.g. abstract) and 20 senses, select K (1-6) senses to cover as many true senses of the target word as possible.
Performance

**Soft Decoding**

**Hard Decoding**
Two Applications

• Rare Senses
  • Idiomaticity

• Frequent Senses
  • Prepositions
Big Fish
There are many living big fish species in the ocean.
He enjoys being a **big fish**, playing with politicians.
Non-Compositionality

• (English) He enjoys being a big fish, playing with the politicians.

• (Chinese) 在當時人們看來，有文化，有墨水的人，就是知識分子。

• (German) In Bletchley Park gab es keinen Maulwurf – mit einer Ausnahme, John Cairncross, aber der spionierte für Stalin.
Motivation

• Non-compositionality in natural language
  • very frequent
  • embodies the creative process
  • applications: information retrieval, machine translation, sentiment analysis, etc.

• Question: Detect idiomaticality

• Challenge: context dependent
Previous Works

• Linguistic resources
• Wiktionary: list definitions
• WordNet: lexical supersenses
• Psycholinguistic database: infer feelings conveyed

• Our contribution: integrate with polysemy

View idiomaticity as a rare sense
Compositional or Not

• (Compositional) Knife has a cutting edge, a sharp side formed by the intersection of two surfaces of an object

• (Idiomatic) Utilize his vast industry contacts and knowledge while creating a cutting edge artworks collection
Geometry of Context Words

- "cutting edge"
- all words -- compositional
- all words -- idiomatic
Geometry of Context Subspace

- “cutting edge”
- sentence subspace -- compositional
- sentence subspace -- idiomatic
Geometry of Context Subspace

- “cutting edge”
- sentence subspace — compositional
- sentence subspace — idiomatic

Idiomaticity score:
- distance between target phrase and context
Subspace-based Algorithm

- Principal Component Analysis (PCA) of sentence word vectors\[^1\]
  
  - Subspace representation

- Compositionality: distance between target word/phrase and subspace

- Test: Idiomatic if distance > threshold
Subspace-based Algorithm

- NO linguistic resources
- Multilingual: English, German and Chinese
- Context sensitive
- Accurate detection in extensive experiments
Ironic

I Love going to the dentist! Looking forward to it all week.

Non-ironic

Love to hear that youthcamp was so awesome!
Subspace-based Algorithm

- “glad”
- sentence subspace -- non-irony
- sentence subspace -- irony

**Irrony detection:** distance from target phrase to context space
Metaphor

- Figurative speech that refers to one thing by mentioning another

• **Metaphor**  
  They often wear an attitude that says – 'I can get away with anything'

• **Non-Metaphor**  
  We always wear helmets when we are riding bikes
Geometry of Metaphor

- "wear"
- sentence subspace -- non-metaphor
- sentence subspace -- metaphor

- Metaphor detection: distance from target phrase to context space
Common Umbrella of Compostionality

- Idiomaticity Detection
- Irony Detection
- Metaphor Detection

- **Context dependent** [Gong, Bhat and V, AAAI ’17]
Experiments: Idioms

- **Given**: bigram phrase and context
- **Goal**: decide idiomatic or not

- **Standard Datasets**:
  - English: English Noun Compounds, e.g., cash cow
    English Verb Particle Compounds, e.g., fill up
  - GNC: German Noun Compounds, e.g., maulwurf
  - Chinese: Chinese Noun Compounds, e.g., 墨水
# Idiomaticity Detection Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>F1 score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENC Dataset</td>
<td>State-of-art</td>
<td>75.5</td>
</tr>
<tr>
<td></td>
<td>This talk</td>
<td>84.2</td>
</tr>
<tr>
<td>EVPC Dataset</td>
<td>State-of-art</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>This talk</td>
<td>46.2</td>
</tr>
<tr>
<td>GNC Dataset</td>
<td>PMI</td>
<td>61.1</td>
</tr>
<tr>
<td></td>
<td>This talk</td>
<td>62.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Dataset</td>
<td>Baseline</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td>This talk</td>
<td>88.3</td>
</tr>
</tbody>
</table>
“in” has 15 senses:

- Manner or degree: *in all directions*
- Time frame: *in 2017*
- Things entered: *in the mail*
- Things enclosed: *in the United States*
- Profession aspects: *in graduate studies*
- Variable quality: *in a jacket*
His band combines professionalism with humor. (Accompanier)

She blinked with confusion. (Manner & Mood)

He washed a small red teacup with water. (Means)
Feature Selection for Disambiguation

**Left context feature**: average of left context

**Right context feature**: average of right context

**Context-interplay feature**: the vector closest to both left and right context space
Intrinsic Evaluation

• SemEval dataset\(^1\): 34 prepositions instantiated by 24,663 sentences covering 332 sense

• Oxford English Corpus (OEC) dataset\(^2\): 7,650 sentences collected from Oxford dictionary

• Spatial relation dataset\(^3\): 5 fine-grained spatial relations with 400 sentences

## Intrinsic Evaluation: SemEval

<table>
<thead>
<tr>
<th>System</th>
<th>Resources</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our system</strong></td>
<td>English corpus</td>
<td>0.80</td>
</tr>
<tr>
<td>Litkowski, 2013</td>
<td>Lemmatizer, dependency parser</td>
<td>0.86</td>
</tr>
<tr>
<td>Srikumar and Roth, 2013</td>
<td>dependency parser, WordNet</td>
<td>0.85</td>
</tr>
<tr>
<td>Gonen and Goldberg, 2016</td>
<td>multilingual corpus, aligner, dependency parser</td>
<td>0.81</td>
</tr>
<tr>
<td>Ye and Baldwin, 2007</td>
<td>chunker, WordNet dependency parser</td>
<td>0.69</td>
</tr>
<tr>
<td>System</td>
<td>Resources</td>
<td>Accuracy</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Our system</td>
<td>English corpus</td>
<td>0.40</td>
</tr>
<tr>
<td>Litkowski, 2013</td>
<td>Lemmatizer, dependency parser, WordNet</td>
<td>0.32</td>
</tr>
</tbody>
</table>
## Intrinsic Evaluation: Spatial Relation

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Spatial Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>Full Containment</td>
<td>apple in the bag</td>
</tr>
<tr>
<td></td>
<td>Partial Containment</td>
<td>finger in the ring</td>
</tr>
<tr>
<td>on</td>
<td>Adhesion to Vertical Surface</td>
<td>sign on the building</td>
</tr>
<tr>
<td></td>
<td>Support by Horizontal Surface</td>
<td>leaf on the ground</td>
</tr>
<tr>
<td></td>
<td>Support from Above</td>
<td>bat on the branch</td>
</tr>
</tbody>
</table>

Our system achieves an accuracy of **77.5%**, compared with 71% achieved by the state-of-art.
Extrinsic Evaluation

• **Light-weight** disambiguation system
  • no reliance on external linguistic resources

• **Efficient** scaling to enrich large corpus
  • train sense representations

• **Extrinsic** evaluation
  • semantic relation
  • paraphrasing of phrasal verbs
Extrinsic Evaluation: Semantic Relation

- Sense representations encode relations

\[
\begin{align*}
\text{in (Location) + Korea} & \sim \text{Korean} \\
\text{from (RangeStart) + Rome} & \sim \text{Italy}
\end{align*}
\]
Extrinsic Evaluation: Paraphrasing

to **fight for** (sense: Benefits) the first prize
~ to **win** the first prize

to **fight for** (sense: Purpose) legal rights
~ to **defend** legal rights
Conclusion

- Geometries of word vectors
  - Angular symmetry
  - better representations

- Geometry of polysemy
  - subspace representations
  - idiomaticity detection preposition vectors

Fun:

- modeling, algorithms, language
Collaborators

Hongyu Gong  Jiaqi Mu  Suma Bhat