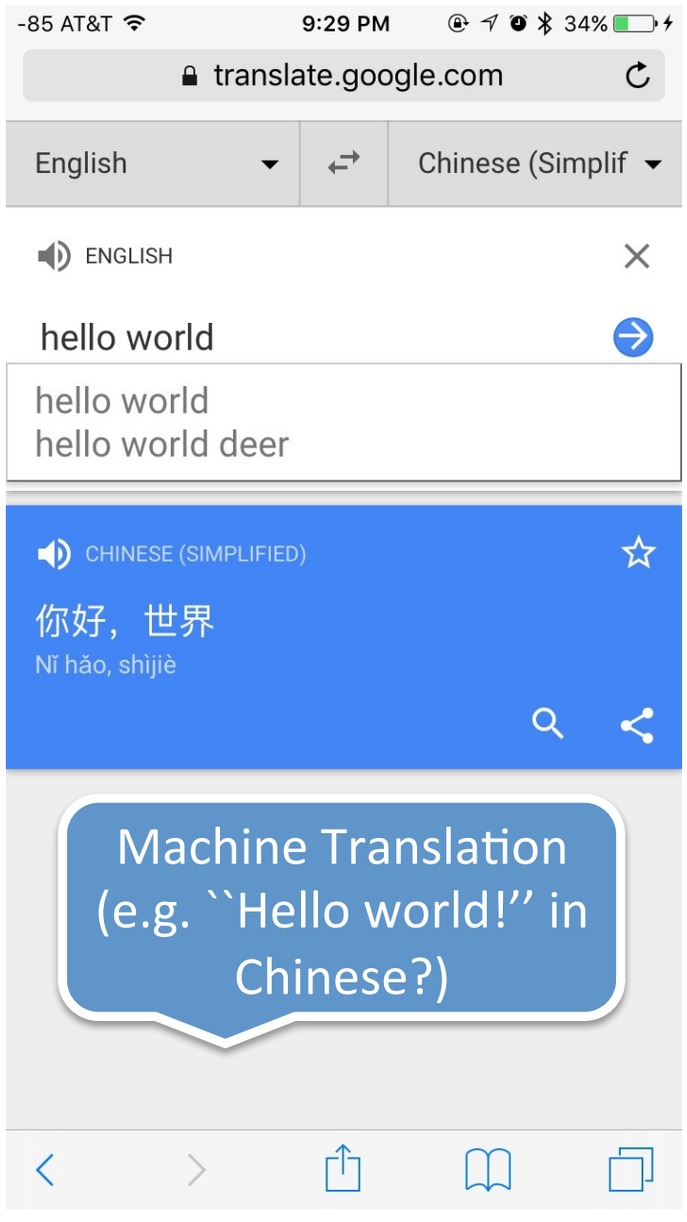
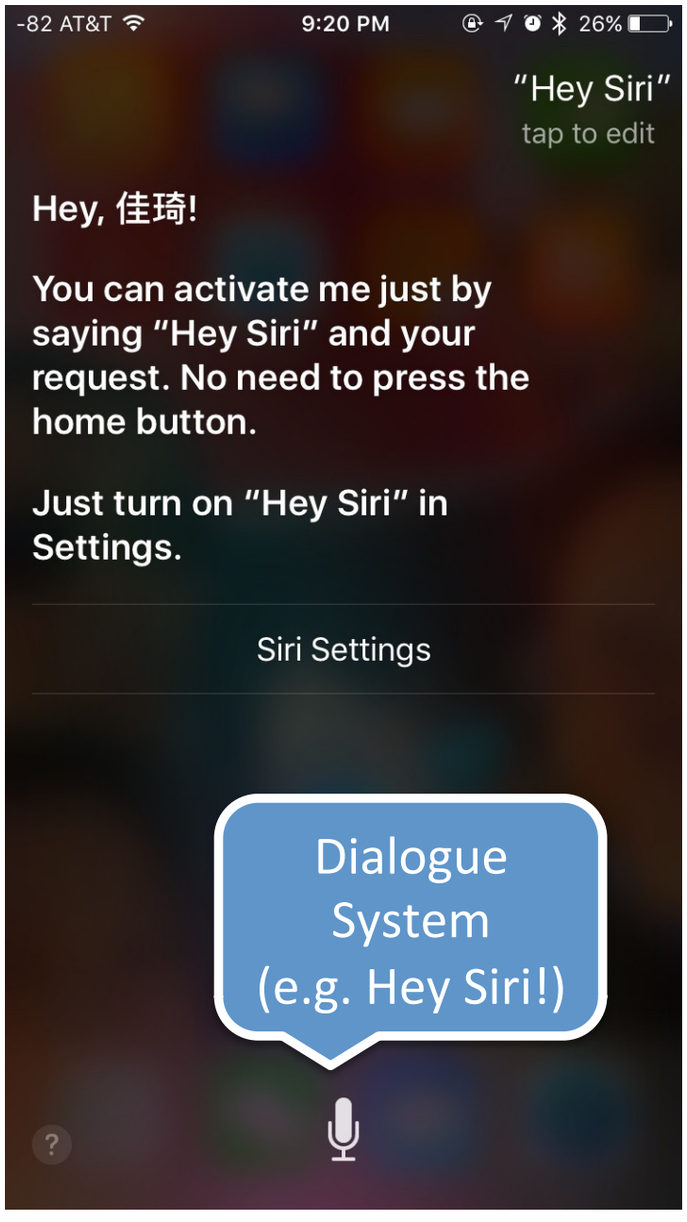
# Geometries of Word Embeddings

Pramod Viswanath

University of Illinois

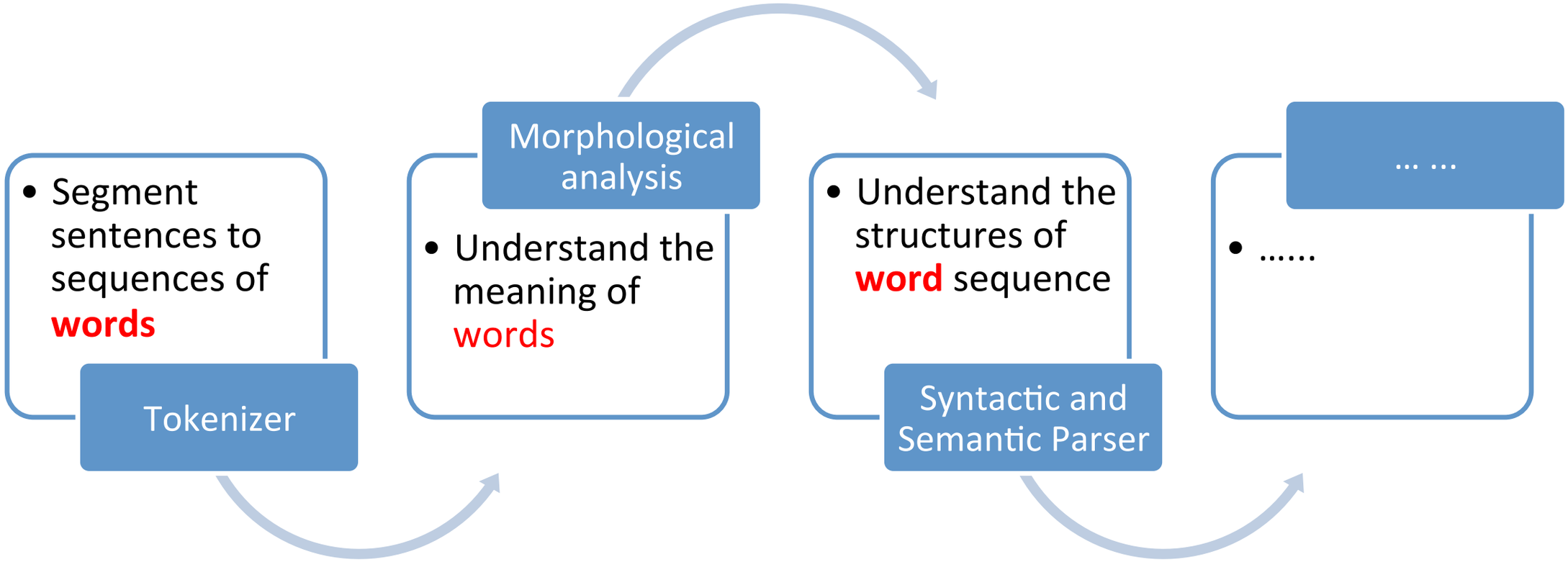



**natural language processing**




**Natural language processing is widely used in daily life.**

# Natural language processing pipeline




**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.**

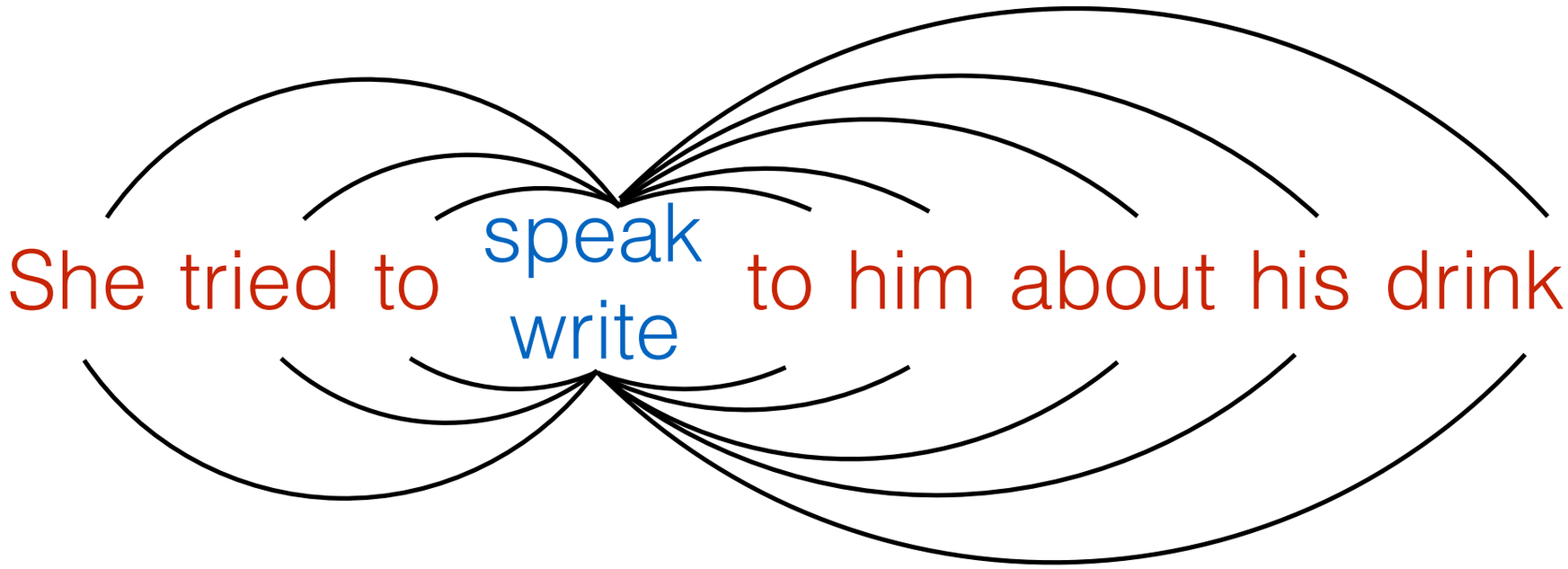
# Word Vector Representations

- **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




**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)

**target words**

**context words**

| | … | genres | … |
|---|---|---|---|
| … | … | … | … |
| series | … | +1 | … |
| of | … | +1 | … |
| many | … | +1 | … |
| including | … | +1 | … |
| fantasy | … | +1 | … |
| drama | … | +1 | … |
| … | … | … | … |

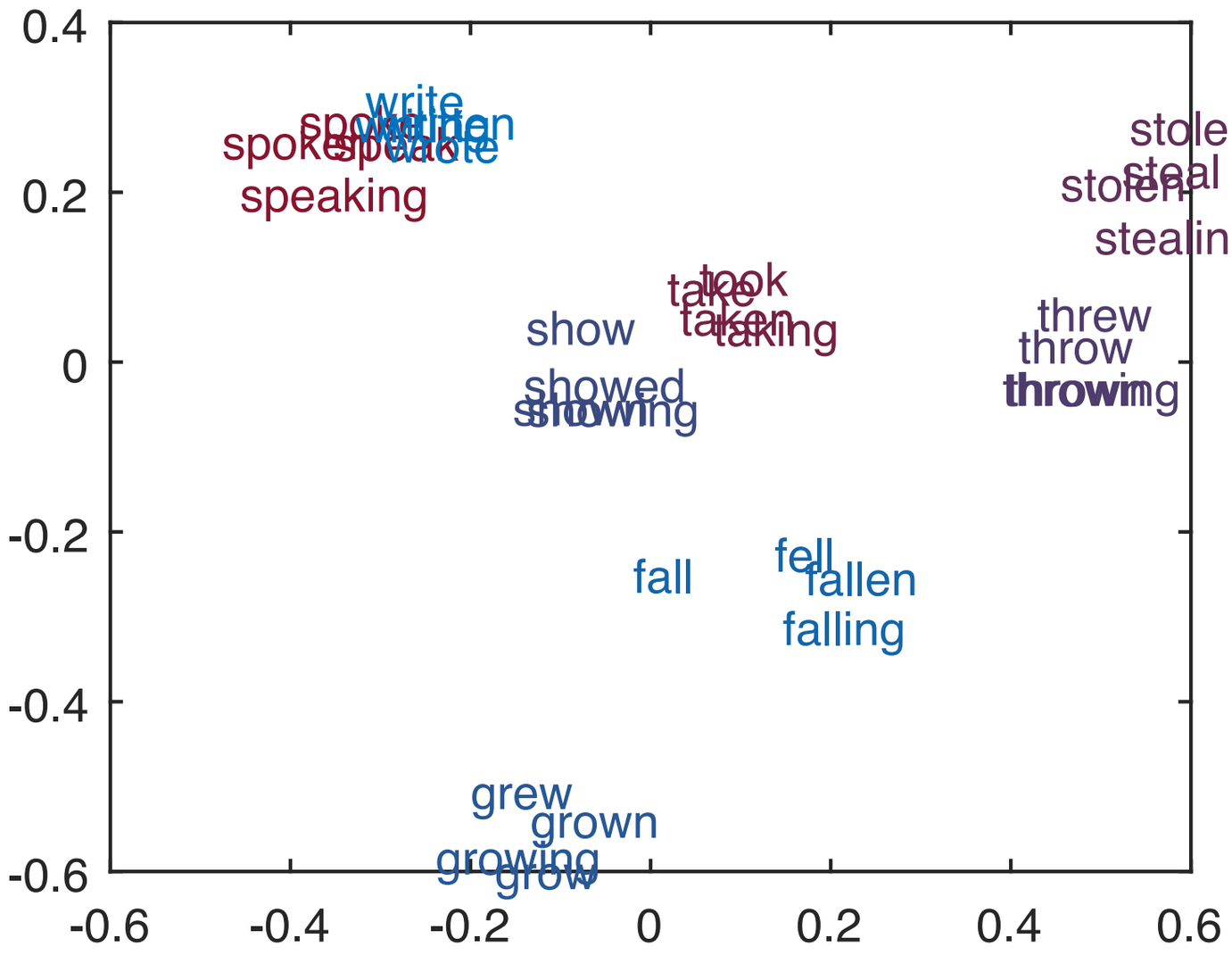
# PMI matrix is low rank

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

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

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

# Word Similarity




$$\text{sim}(w_1, w_2) \stackrel{\text{def}}{=} \frac{u(w_1)^{\text{T}} u(w_2)}{\|u(w_1)\| \|u(w_2)\|}$$

# 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

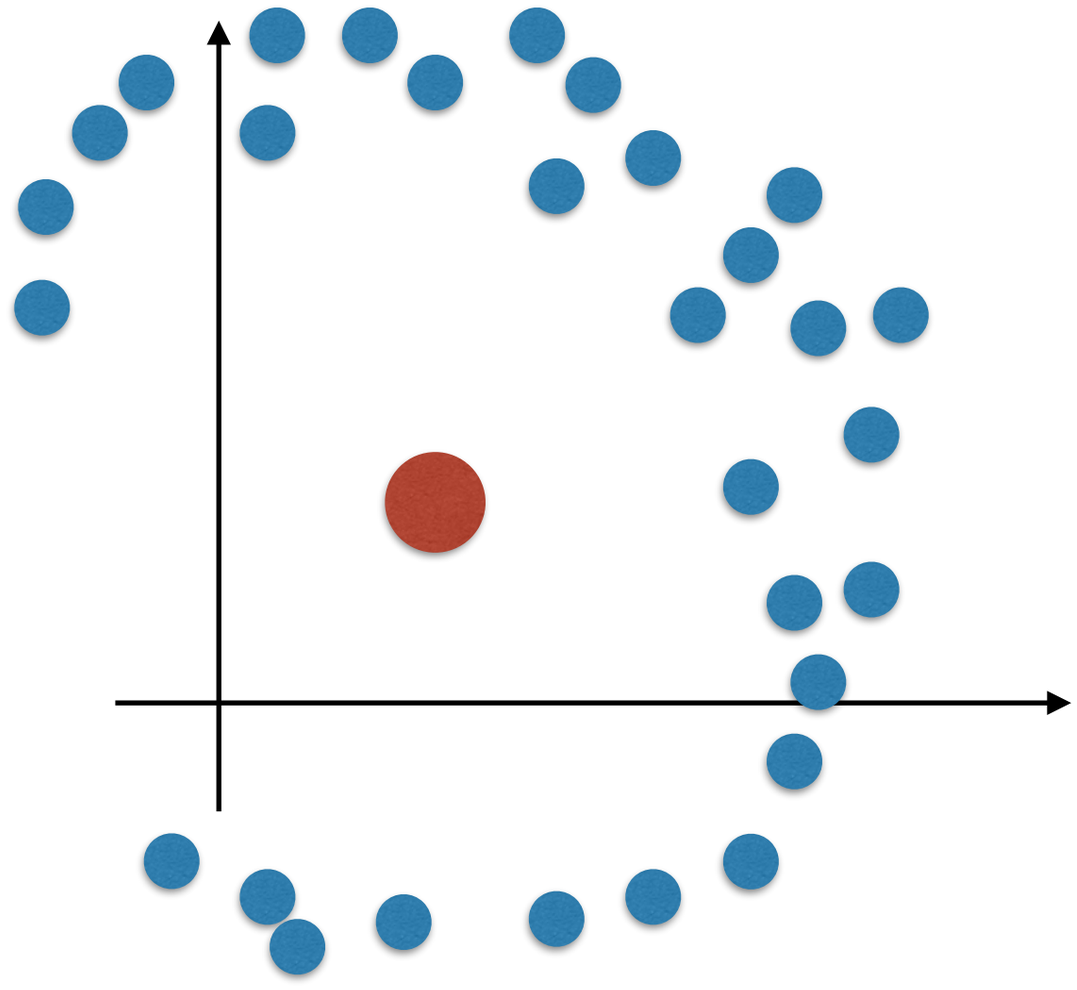
# Isotropy and Word Vectors

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

- **Postprocessing**
  - Simple
  - Universally improves representations

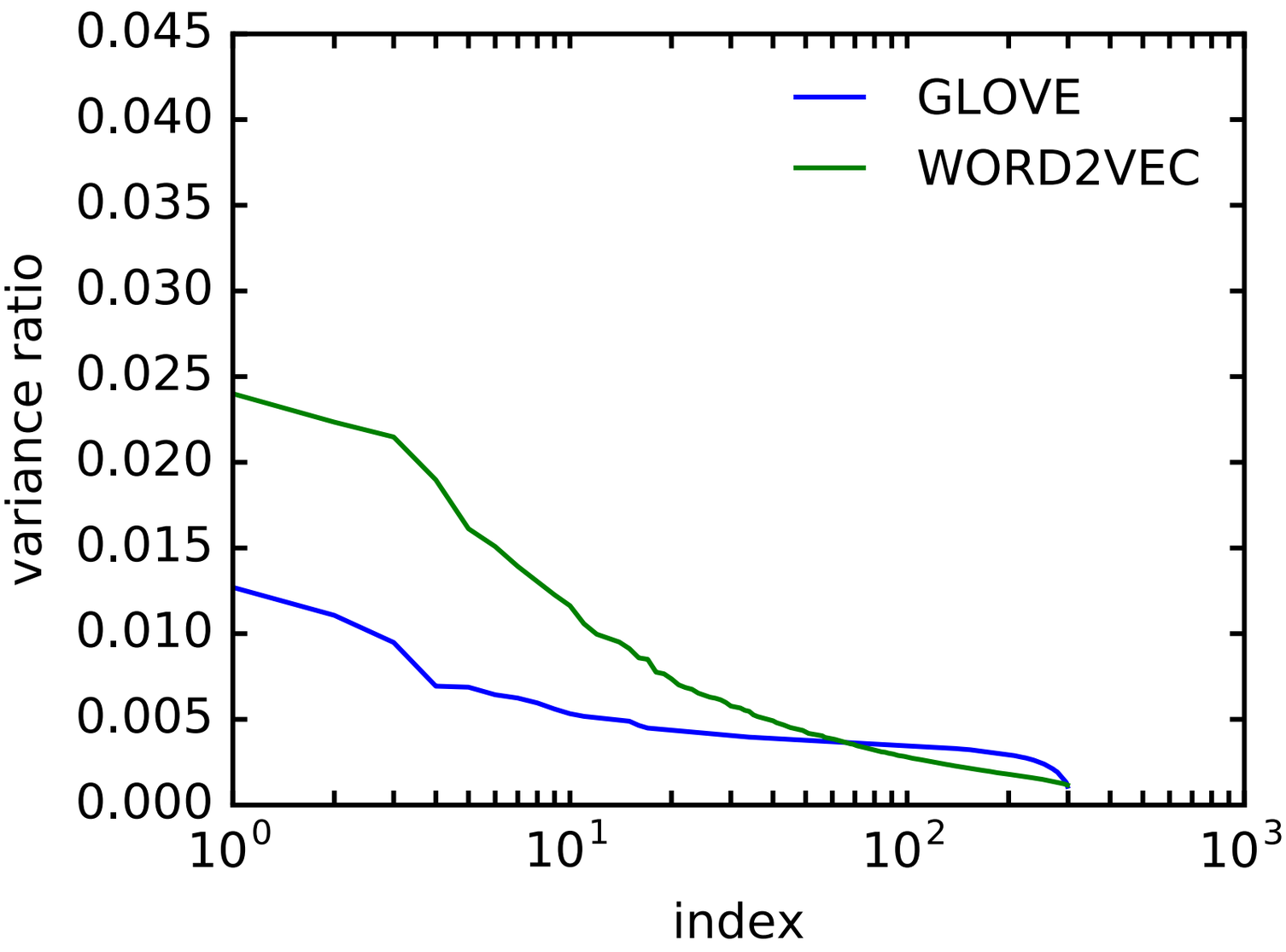
# Geometry of word vectors

| | avg. norm | norm of avg. | ratio |
|---|---|---|---|
| WORD2VEC | 2.04 | 0.69 | **0.34** |
| GLOVE | 8.30 | 3.15 | **0.37** |



**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**

$$u_1, ..., u_d \leftarrow \mathrm{PCA}(\{\tilde{v}(w), w \in V\})$$

$$v'(w) \leftarrow \tilde{v} - \sum_{i=1}^{D} \left(u_i^{\mathrm{T}} v(w)\right) u_i$$

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

| | avg. improvement |
|---|---|
| word2vec | 1% |
| GloVe | 2.5% |

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

# Concept Categorization

Group words into different semantic categories.

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

| | avg. improvement |
|---|---|
| word2vec | 7.5% |
| GloVe | 0.6% |

Datasets: ap, ESSLLI, battig

# Sentence-level Evaluation

✓ Sentential Textual Similarity (STS) 2012-2016

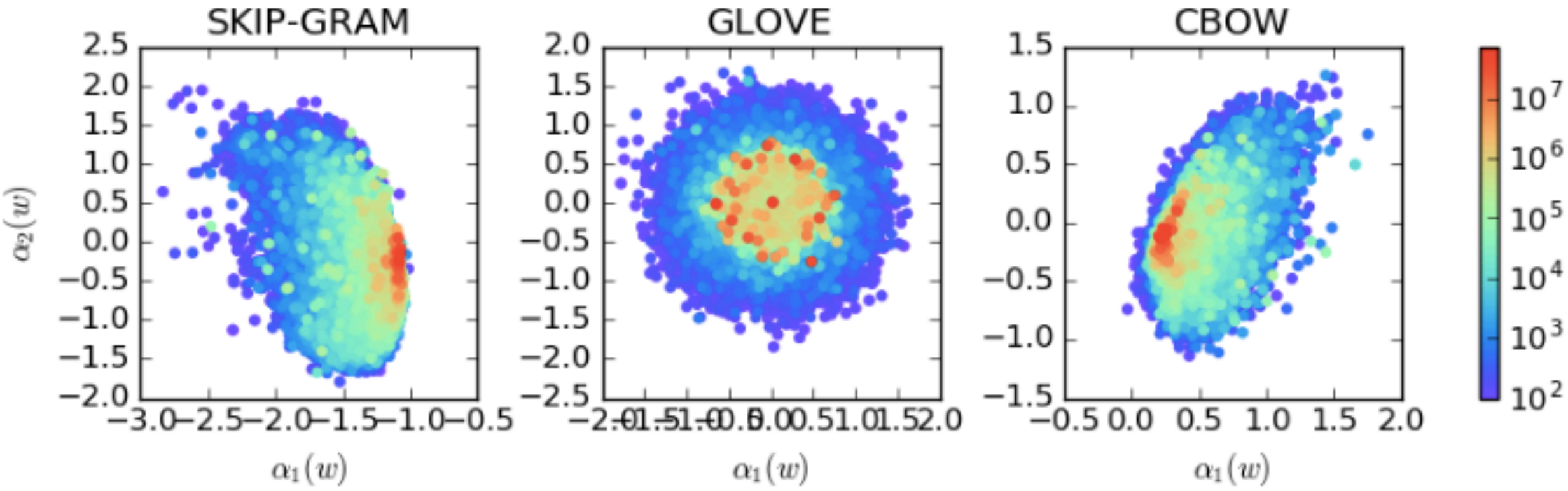
- 21 Different datasets: pairs of sentences
  - algorithm rates  similarity
  - compare to human scores

  - Average improvement of **4%**

# 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

$$\log \frac{p_{W,C}(w,c)}{p_W(w)p_C(c)} \propto v(w)^{\mathrm{T}} v(c)$$

# Post-processing and Isotropy

Measure of isotropy

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

| | before | after |
|---|---|---|
| **word2vec** | 0.7 | **0.95** |
| **GloVe** | 0.065 | **0.6** |

# 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

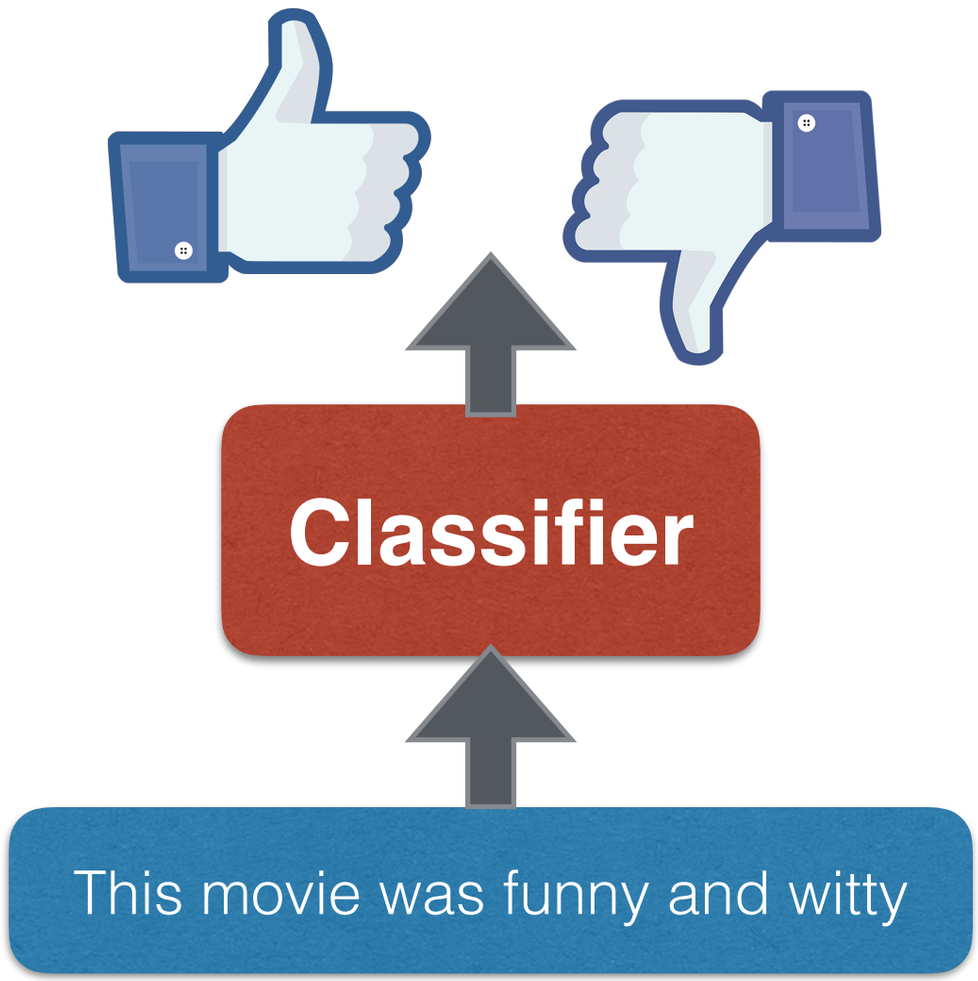
- We improve them universally
  - Angular symmetry

- Other geometries?

# Sentence Representations

# What to preserve?

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

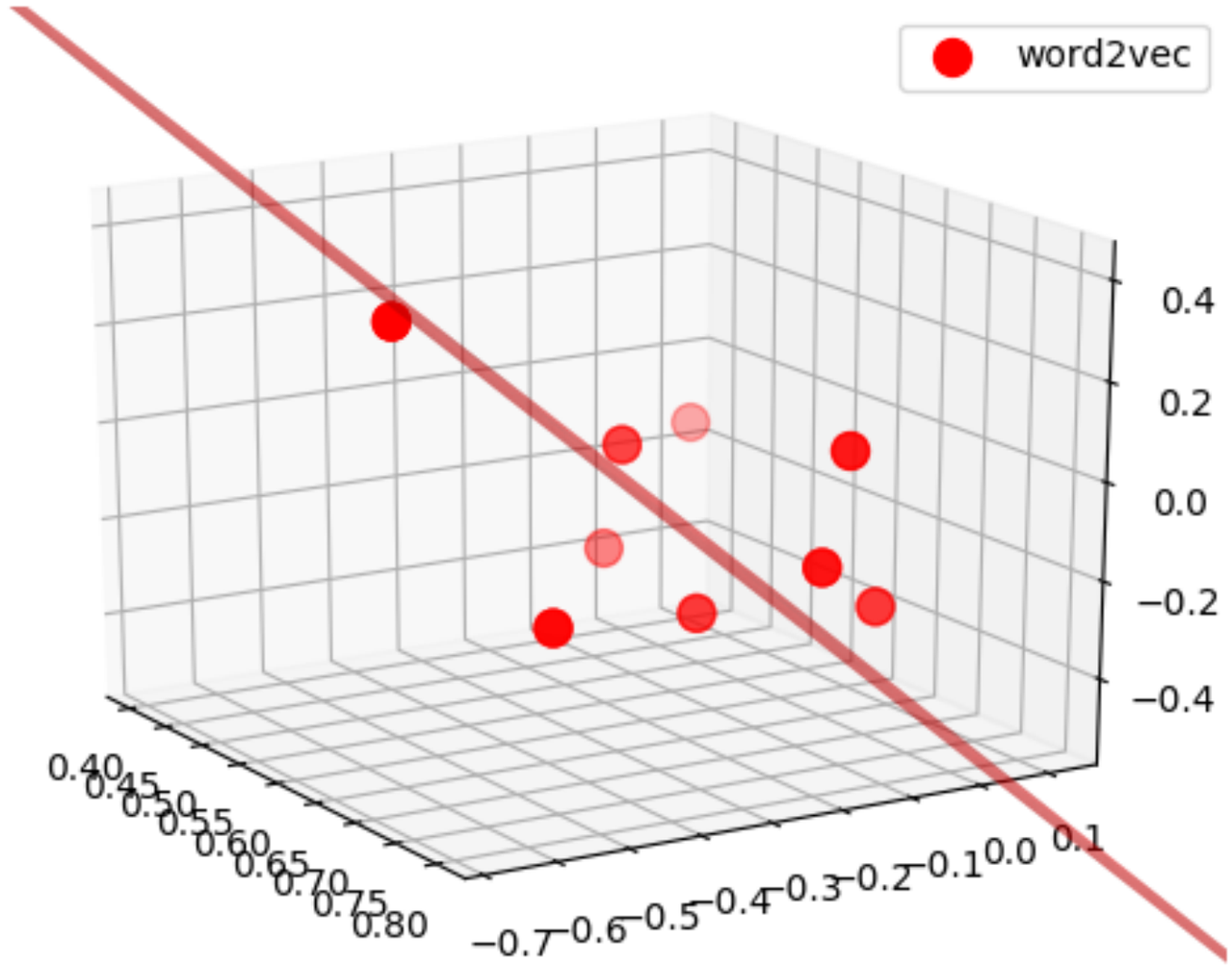


# Representation by Vectors

- Bag-of-words
  - frequency, tf-idf weighted frequency
- Average of word vectors:
  - Wieting et al. 2015, Huang et al. 2012, Adi et al. 2016, Kenter et al. 2016, Arora et al. 2017
- Neural networks:
  - Kim et al. 2014, Kalchbrenner et al. 2014, Sutskever et al. 2014, Le and Mikolov 2014, Kiros et al. 2015, Hill et al. 2016

# 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

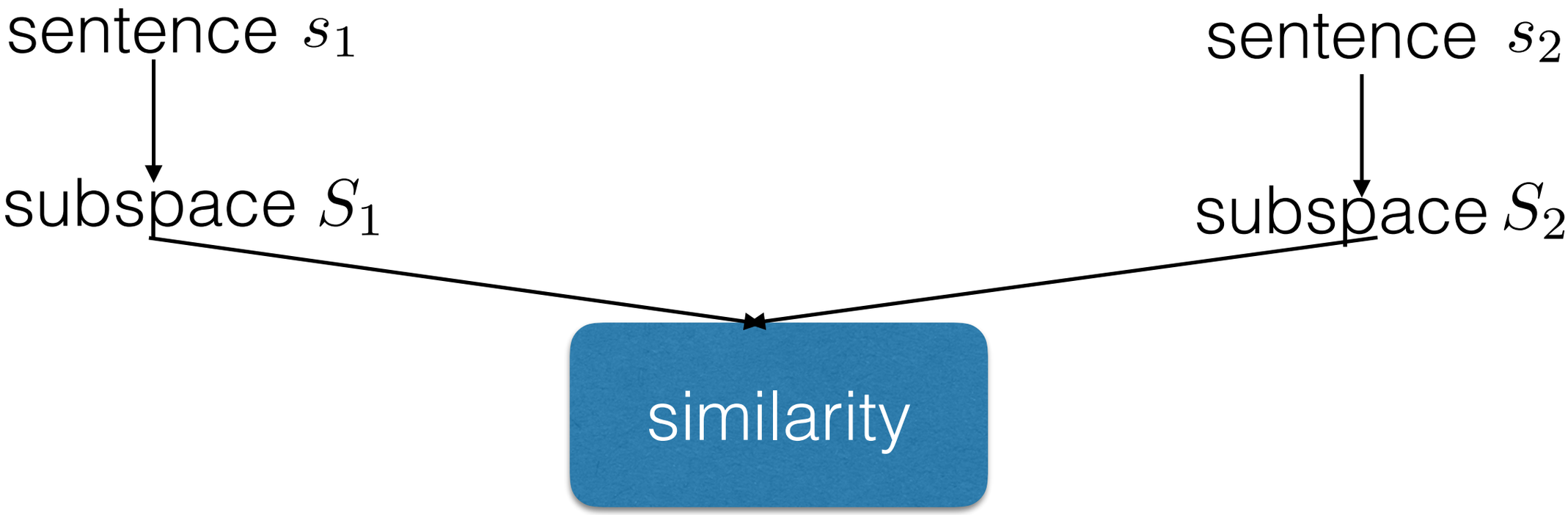
- Input: a sequence of words $\{v(w), w \in s\}$

- Compute the first $N$ principal components

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

- Output: orthonormal basis [Mu, Bhat and **V**, ACL '17]

# Similarity between Sentences

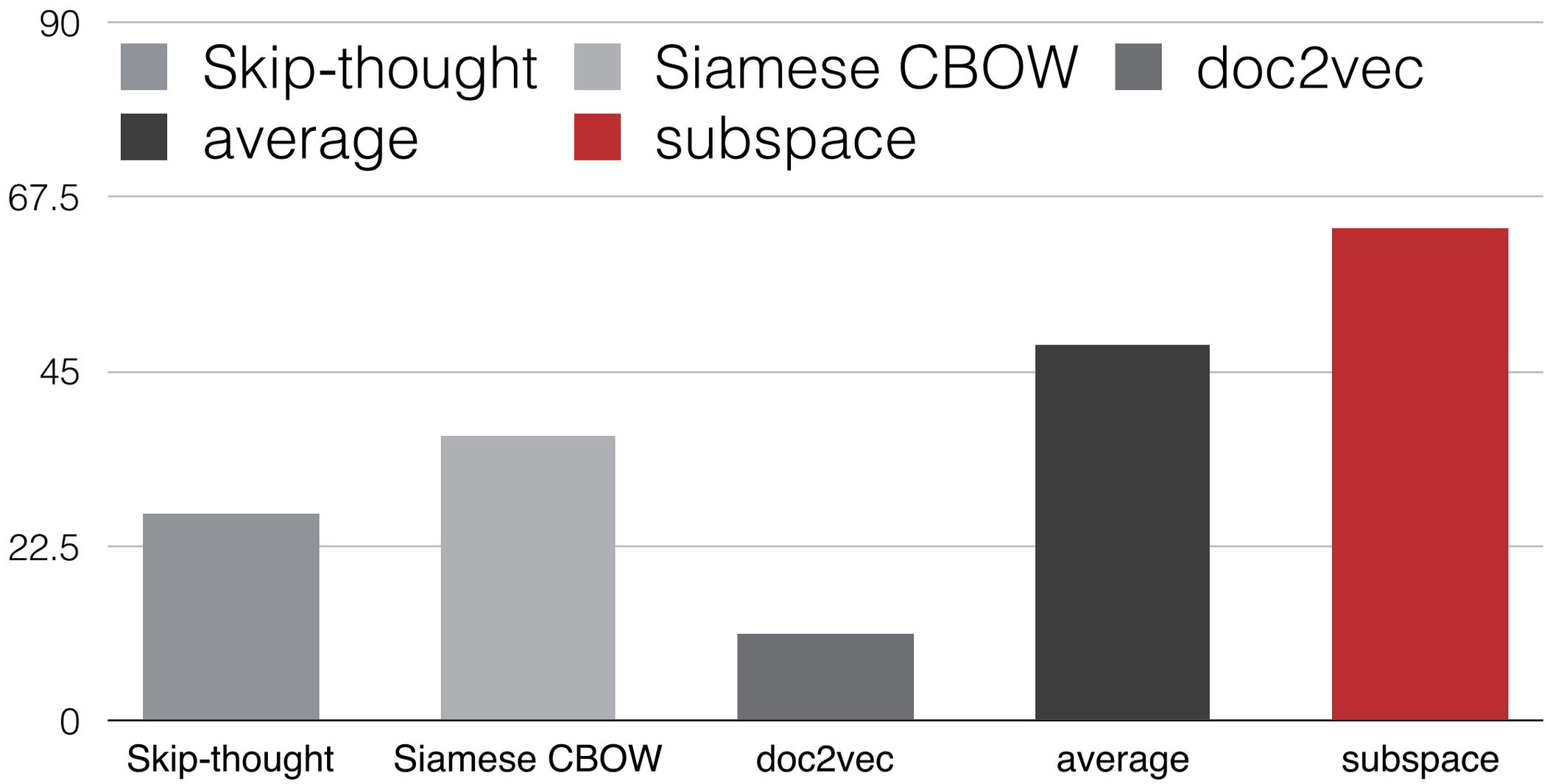



$$\mathrm{CosSim}(s_1, s_2) = \frac{1}{N} d(S_1, S_2)$$
$$\triangleq \frac{1}{N} \sqrt{\mathrm{tr}(S_1 S_1^{\mathrm{T}} S_2 S_2^{\mathrm{T}})}$$

# Examples

| sentence pair | Ground Truth | Predicted Score |
|---|---|---|
| The man is doing exercises. | 0.78 | 0.82 |
| The man is training. | | |
| The man is doing exercises. | 0.28 | 0.38 |
| Two men are hugging. | | |
| The man is doing exercises. | 0.4 | 0.43 |
| Two men are fighting. | | |

# Semantic Textual Similarity Task




**2015.answer-studetns**

# Collaborators

Hongyu Gong

Jiaqi Mu

Suma Bhat