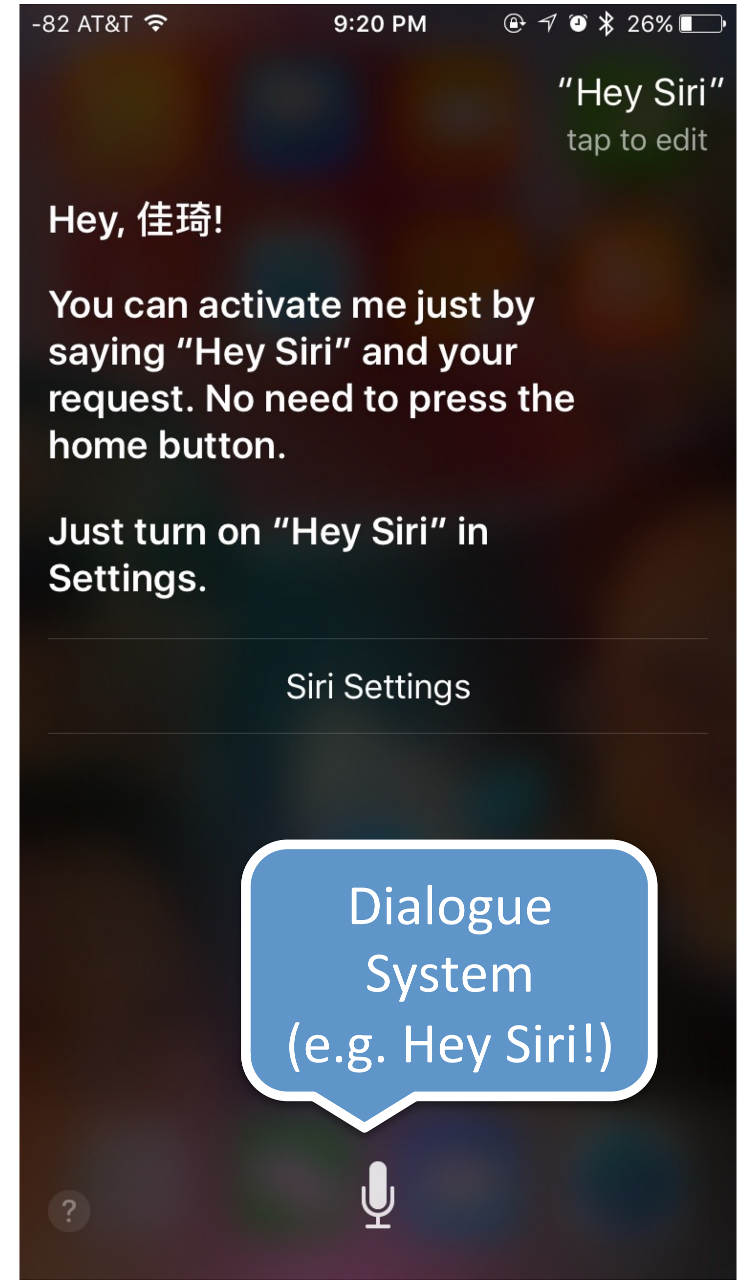
Geometries of Word Embeddings

Pramod Viswanath

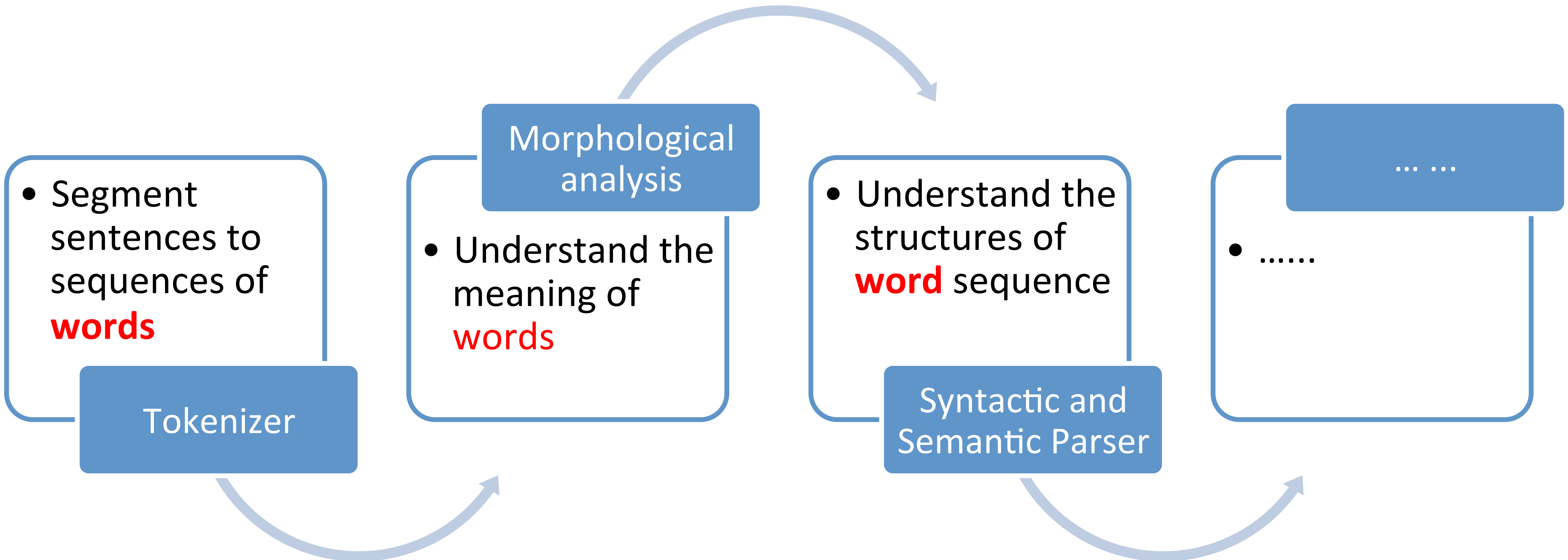
University of Illinois

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

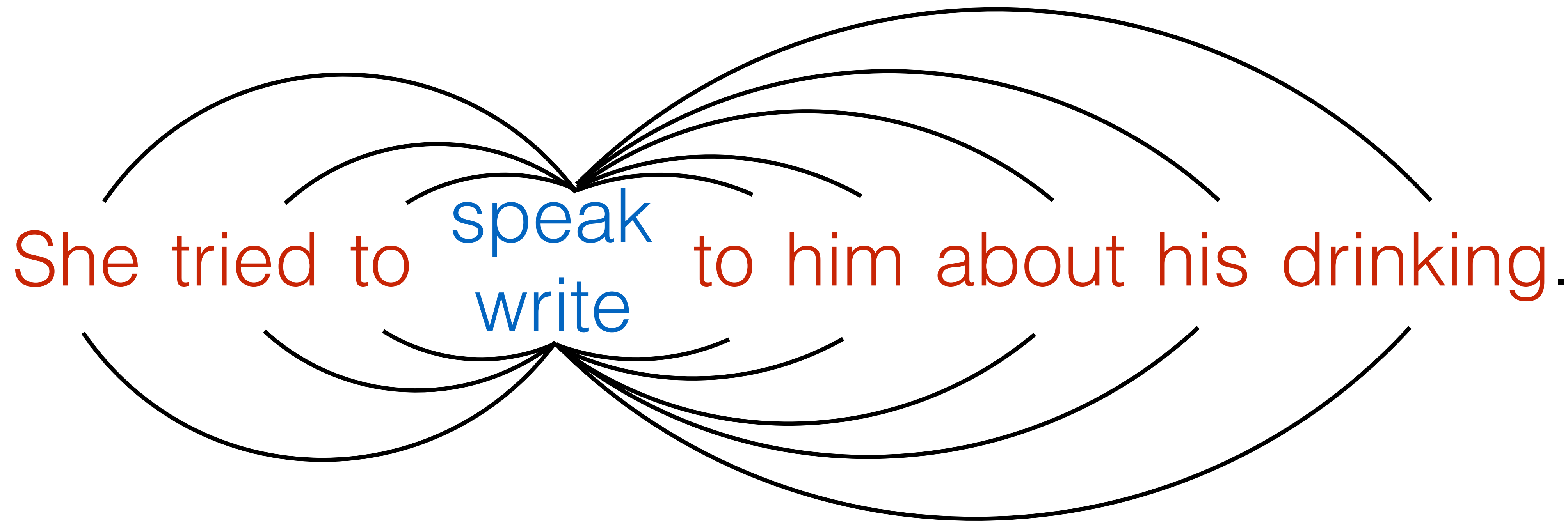
- **Word2Vec** (2013)
  - Google
  - Publicly available

- **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	…
…	…	…	…

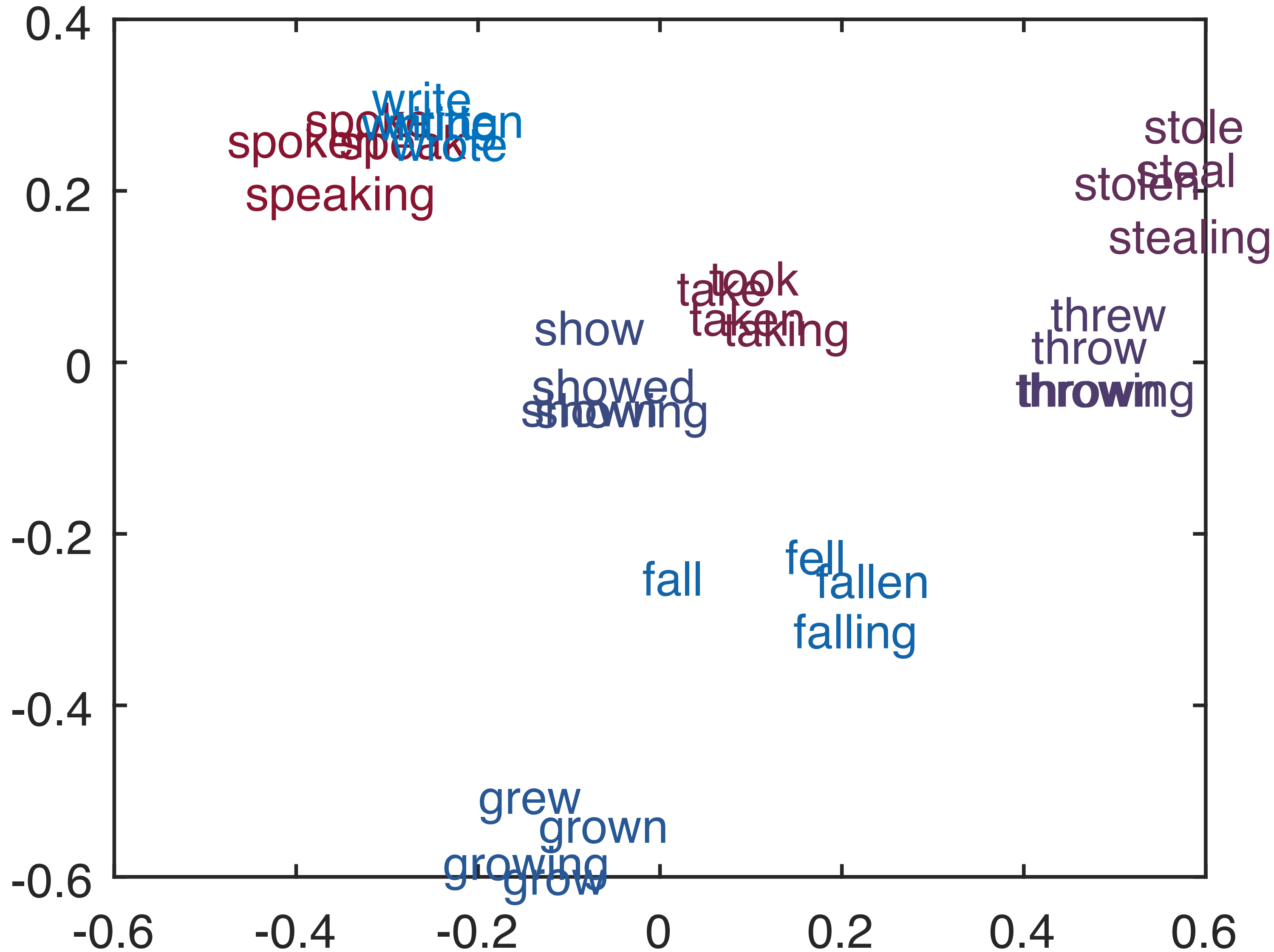
PMI matrix is low rank

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



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

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

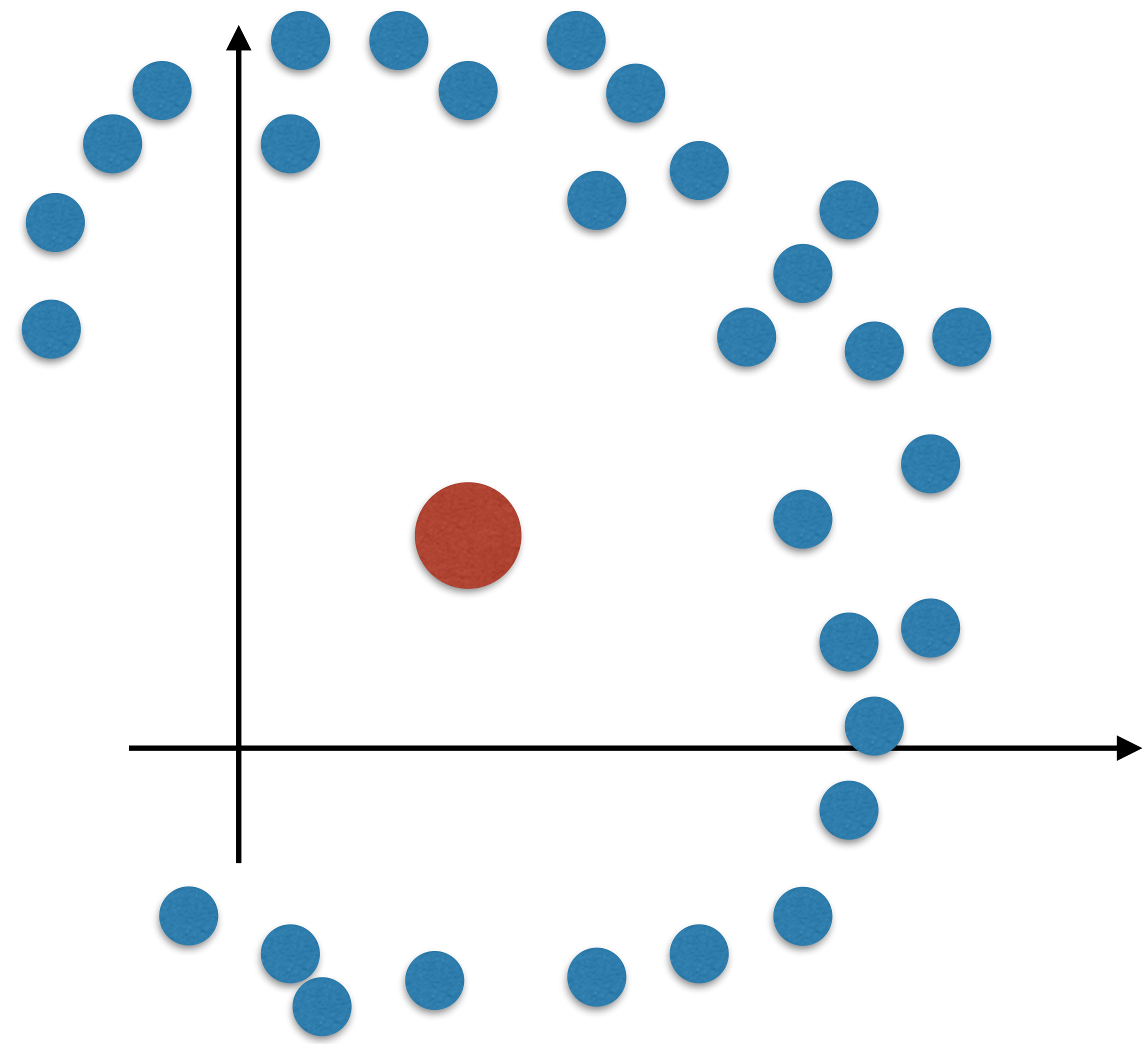
Isotropy and Word Vectors

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

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

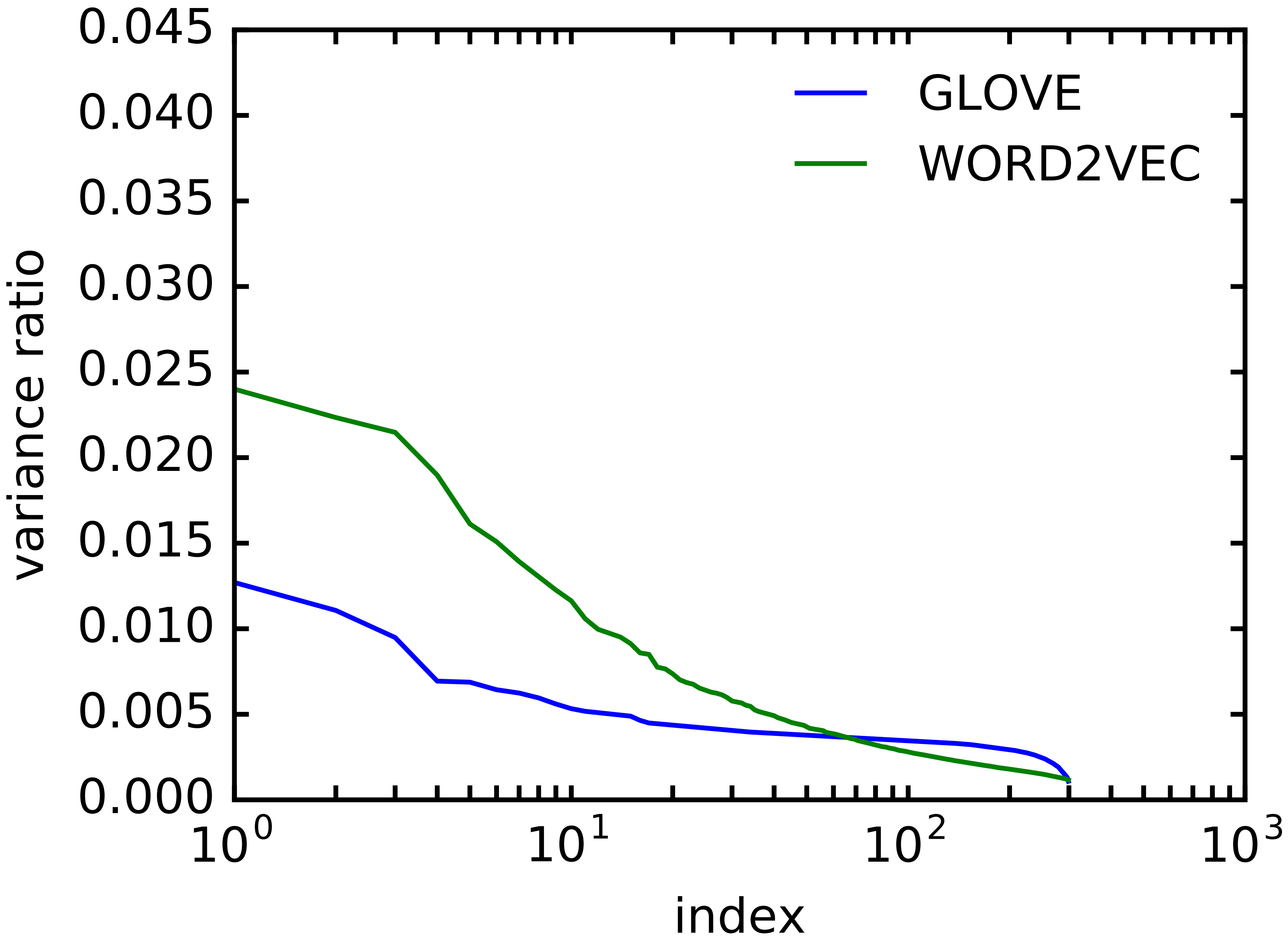
Geometry of word vectors

	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

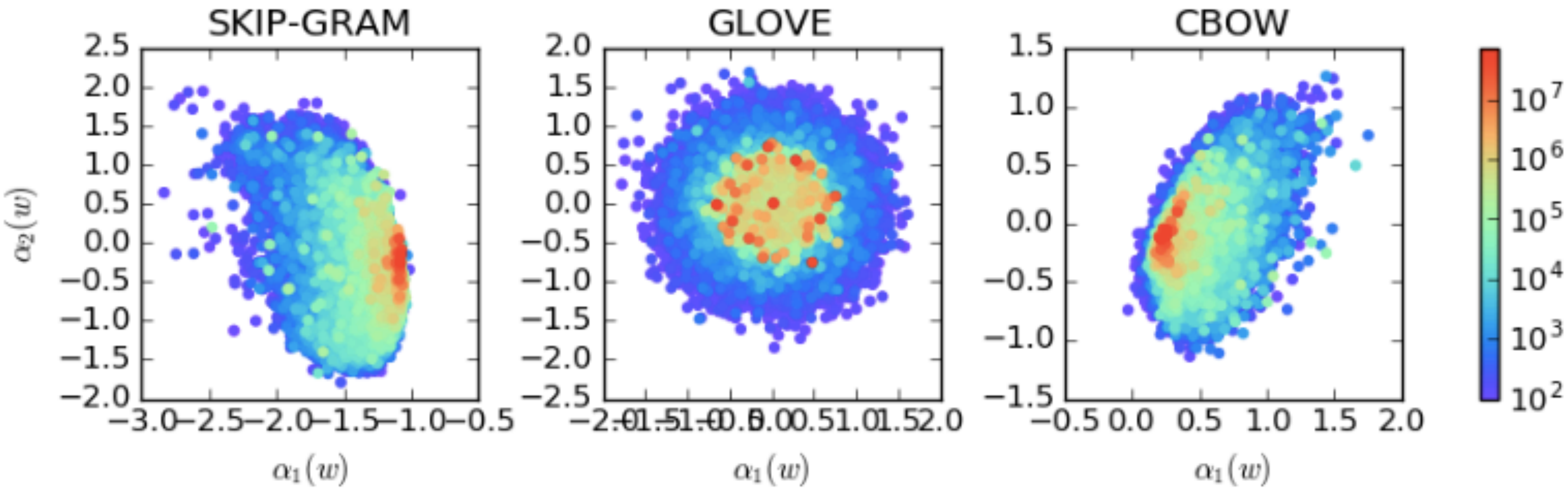
- 21 Different datasets: pairs of sentences
  - algorithm rates similarity
  - compare to human scores

  - 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

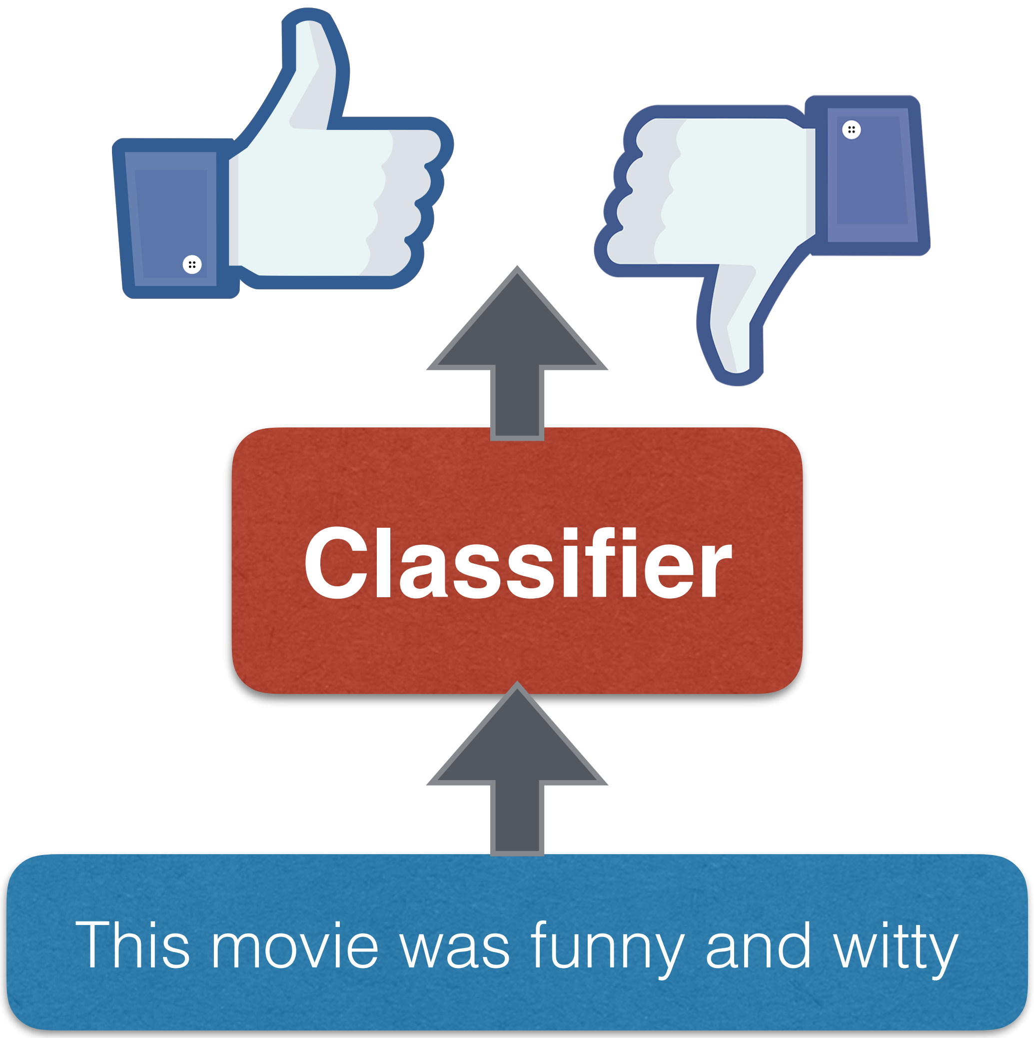
Summary

- 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

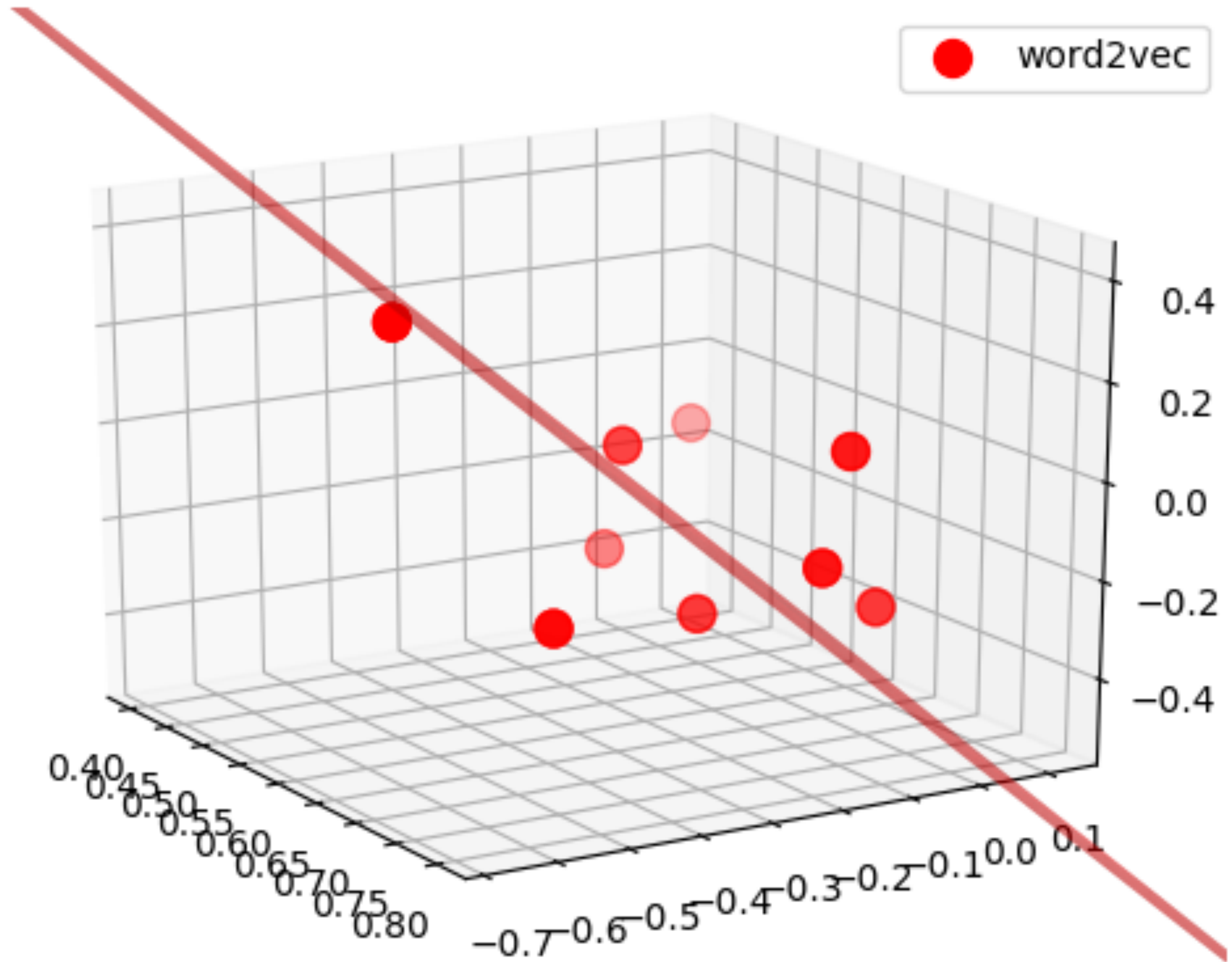


Representation by Vectors

- 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

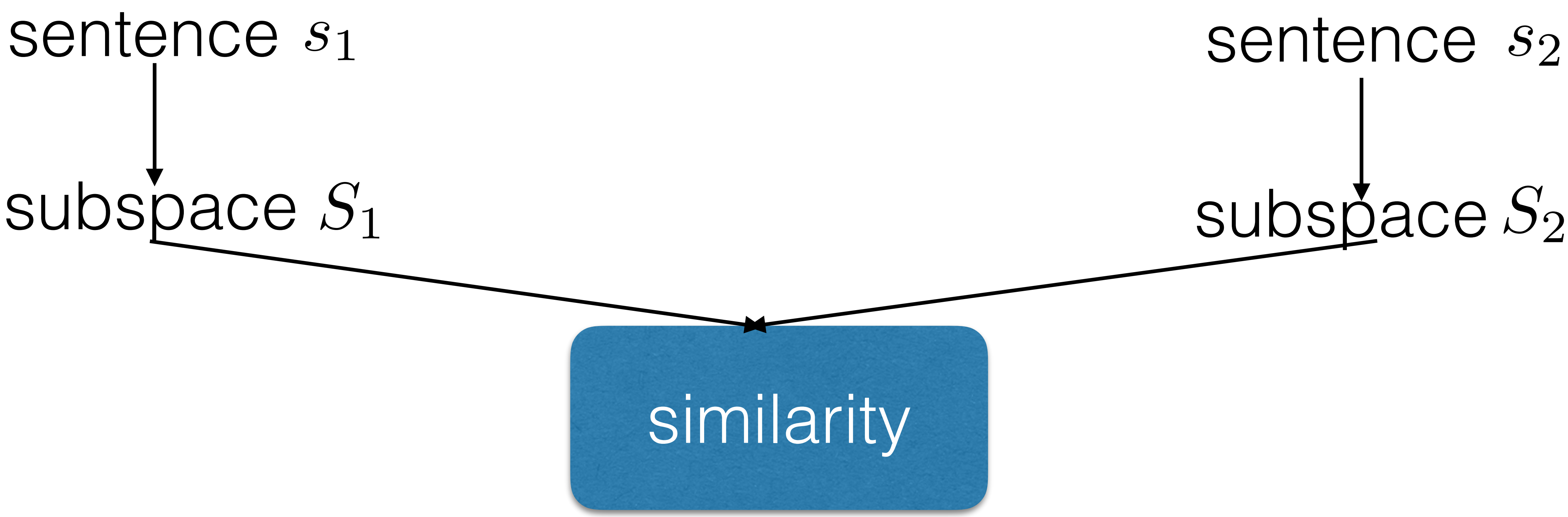
- Input: a sequence of words $\{v(w), w \in s\}$

- 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

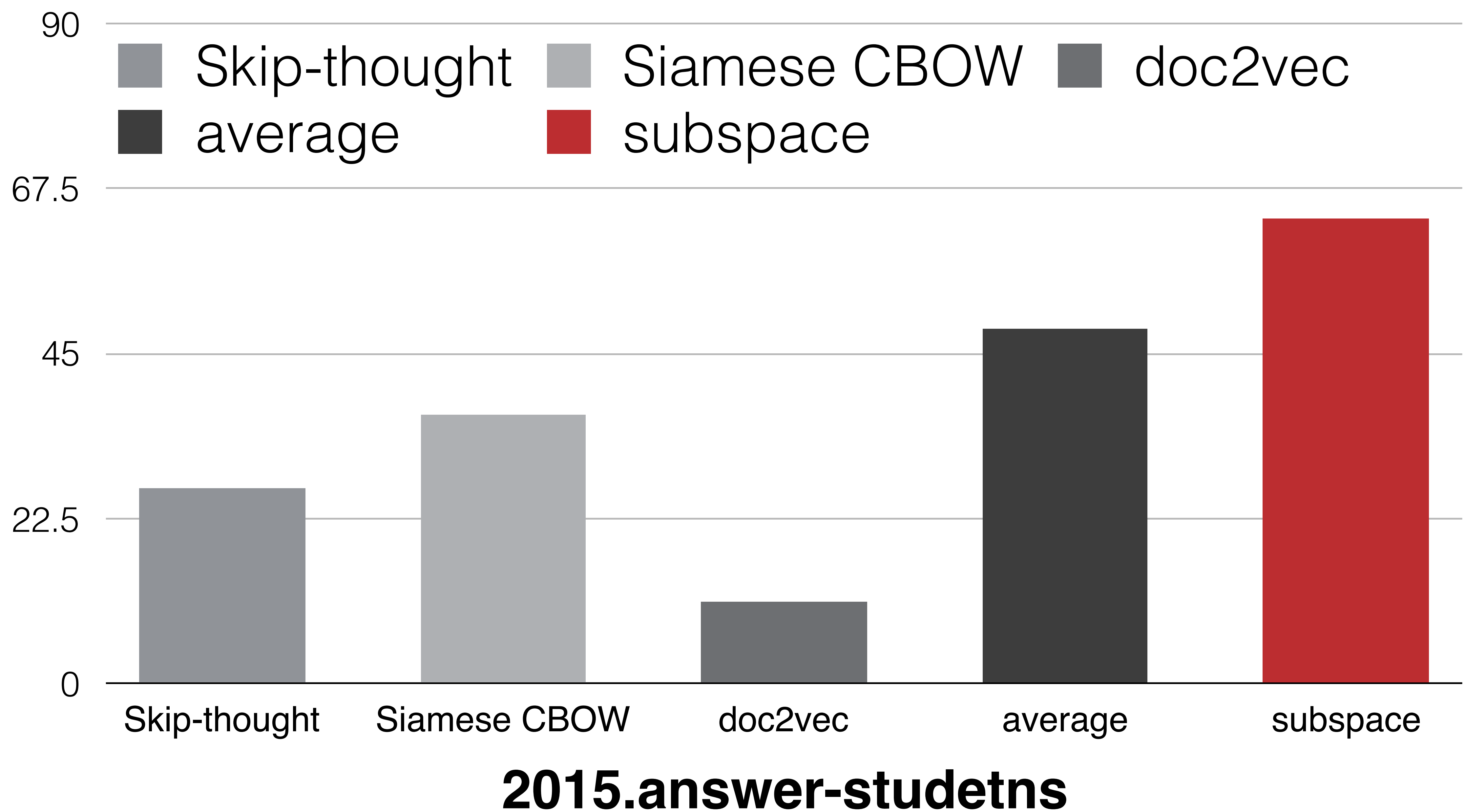


$$\begin{aligned}\mathrm{CosSim}(s_1, s_2) &= \frac{1}{N} d(S_1, S_2) \\ &\triangleq \frac{1}{N}\sqrt{\mathrm{tr}(S_1 S_1^{\mathrm{T}} S_2 S_2^{\mathrm{T}})}\end{aligned}$$

Examples

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

Semantic Textual Similarity Task



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

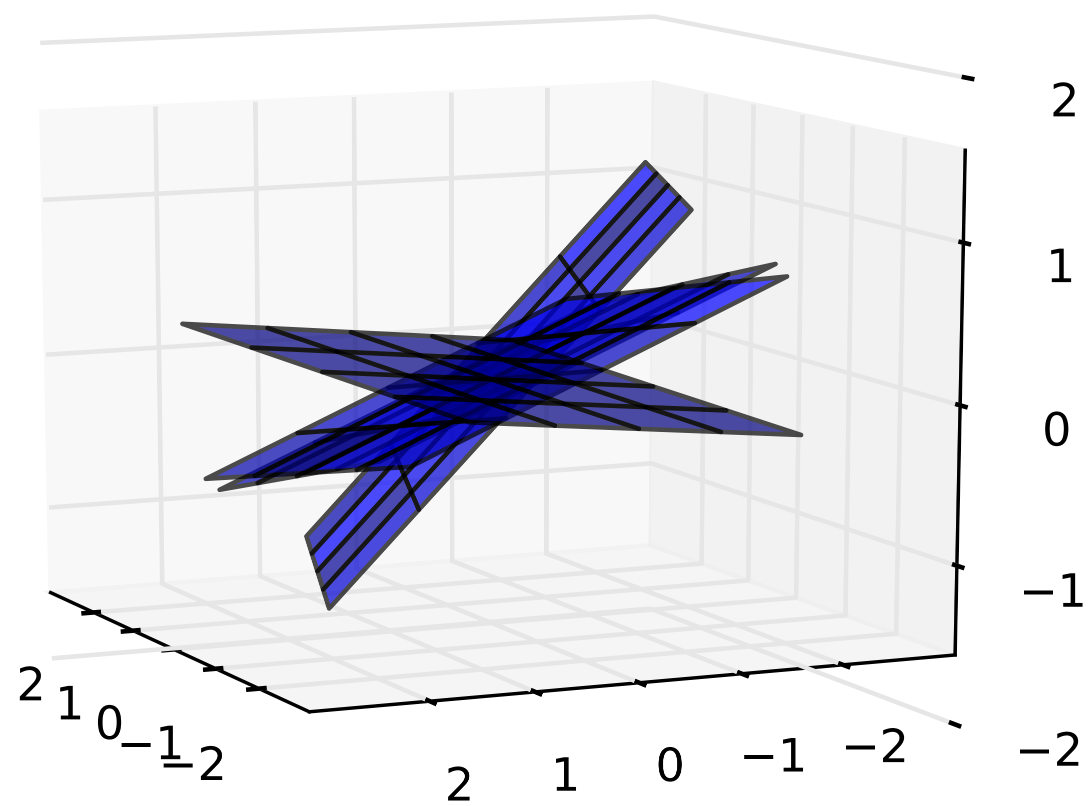
Disambiguation via Context

- (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

“typhoon”



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

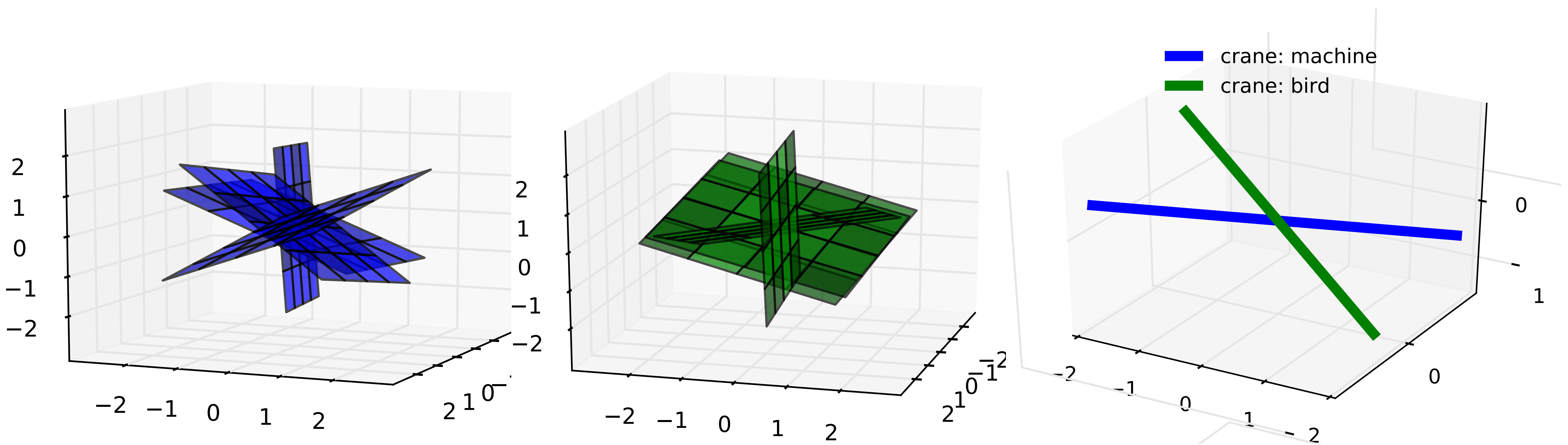
$$\hat{u}(w) = \arg\min_{\|u\|=1} \sum_{w \in c} d(u, S(c \setminus w))^2$$

$$= \arg\min_{\|u\|=1} \sum_{w \in c} \sum_{n=1}^{N} \left(u^{\mathrm{T}} u_n(c \setminus w)\right)^2$$

rank-1 PCA of $\{u_n(c \setminus w)\}_{c,n=1,\ldots,N}$

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 $c_1, ..., c_M$
    number indicating the number of senses K.

- Output: partition the M contexts into K sets $S_1, ..., S_K$

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

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

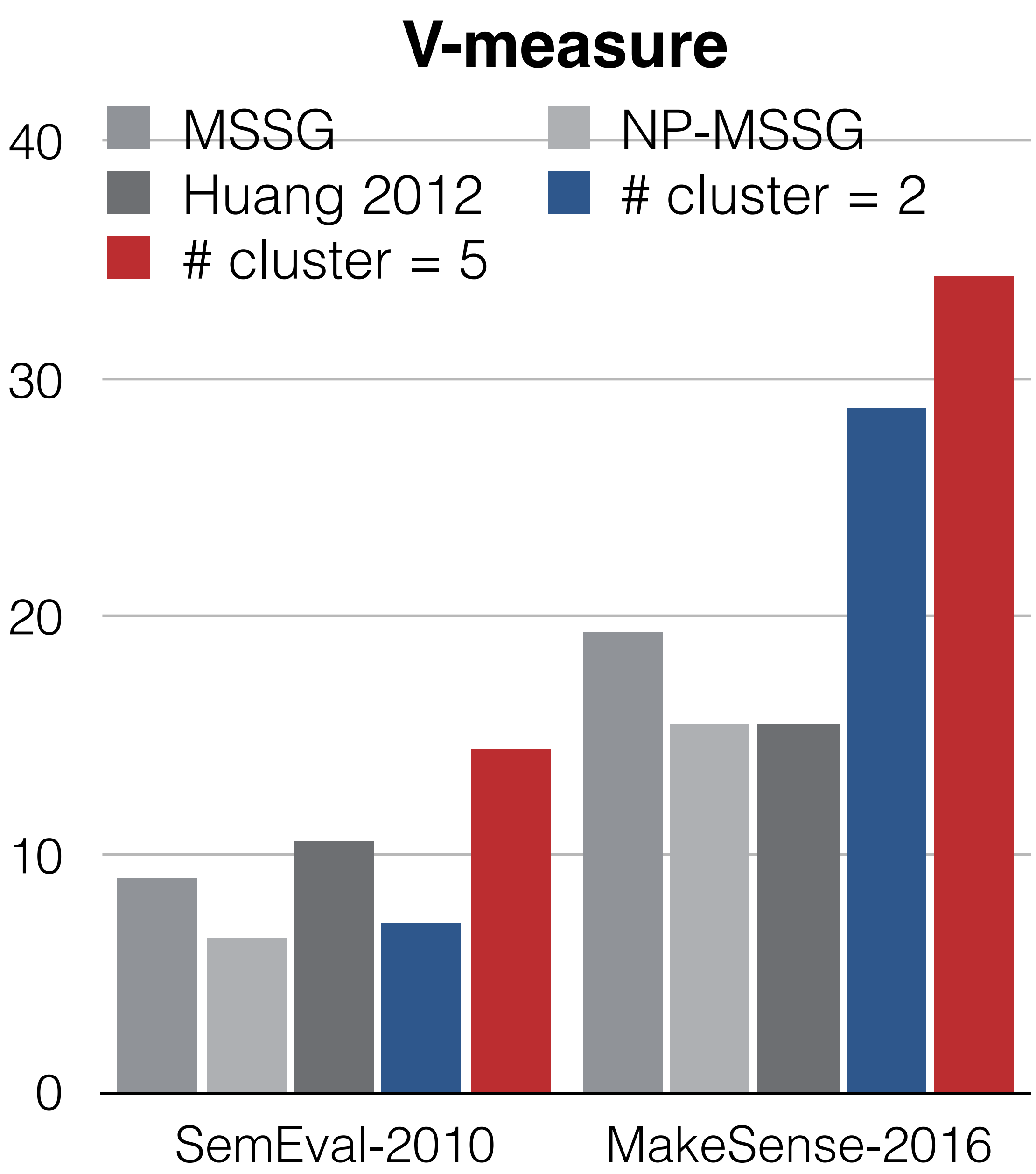
- Soft Decoding: output a probability distribution

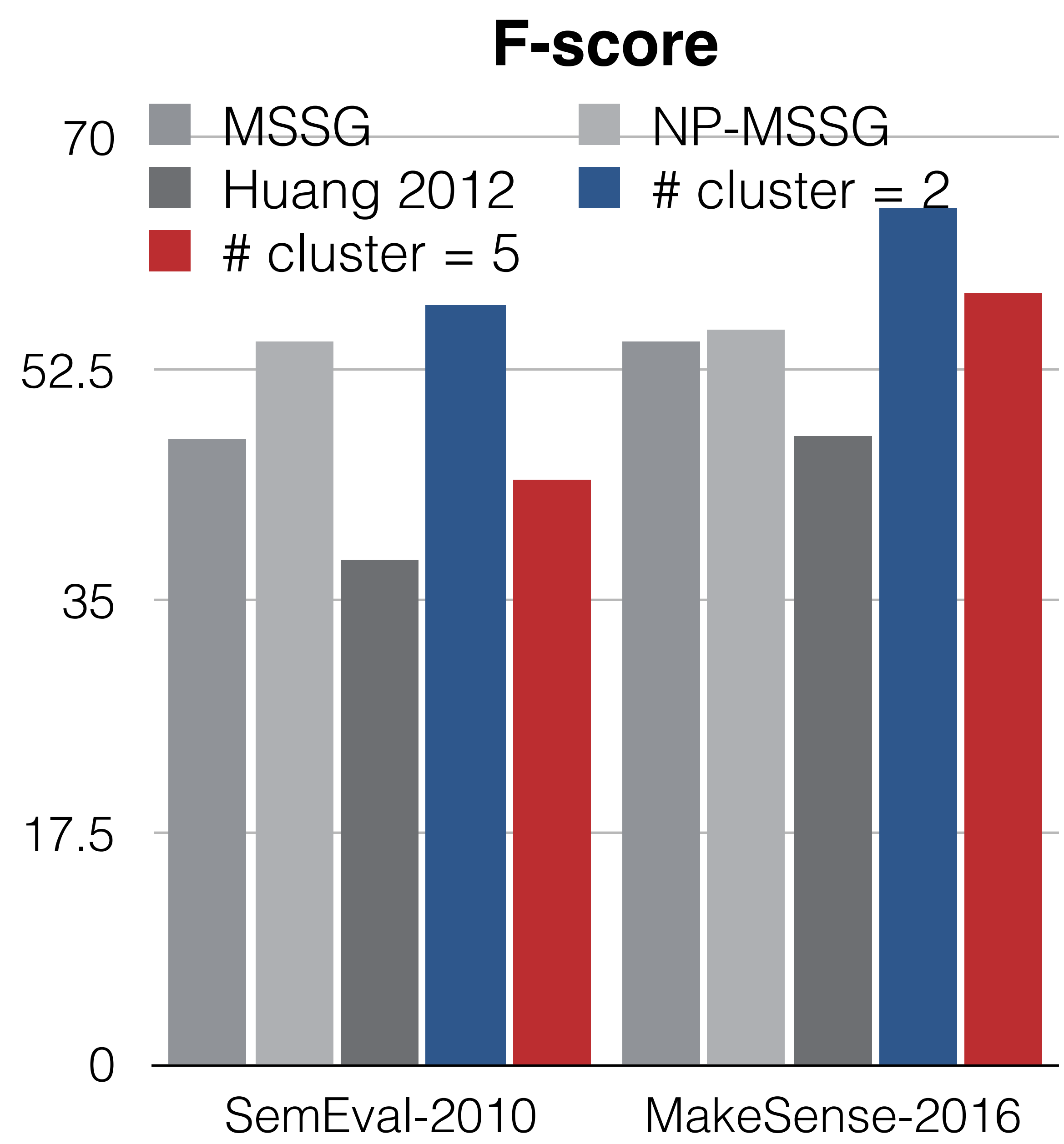
$$P(w, c, k) = \frac{\exp(-d(u_k(w), S(c \setminus w)))}{\sum_{k'} \exp(-d(u_{k'}(w), S(c \setminus w)))}$$

- Hard Decoding: output a deterministic classification

$$k^* = \arg\min_k d(u_k(w), S(c \setminus w))$$

SemEval Share Tasks



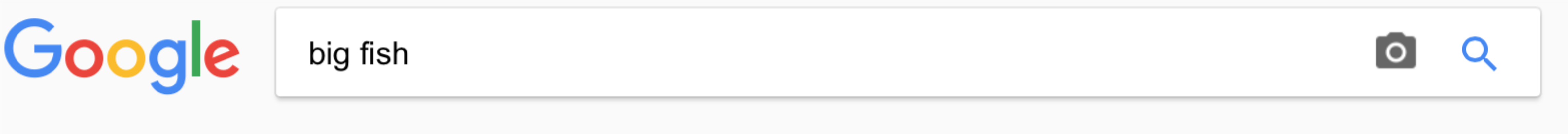


[Mu, Bhat and **V**, ICLR '17]

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 idiomaticity

- Challenge: context dependent

Previous Works

- Linguistic resources
  - Wikitionary: list definitions
  - WordNet: lexical supersenses
  - Psycholinguistic database: infer feelings conveyed

- Our contribution: integrate with polysemy

View idiomaticity as a rare sense

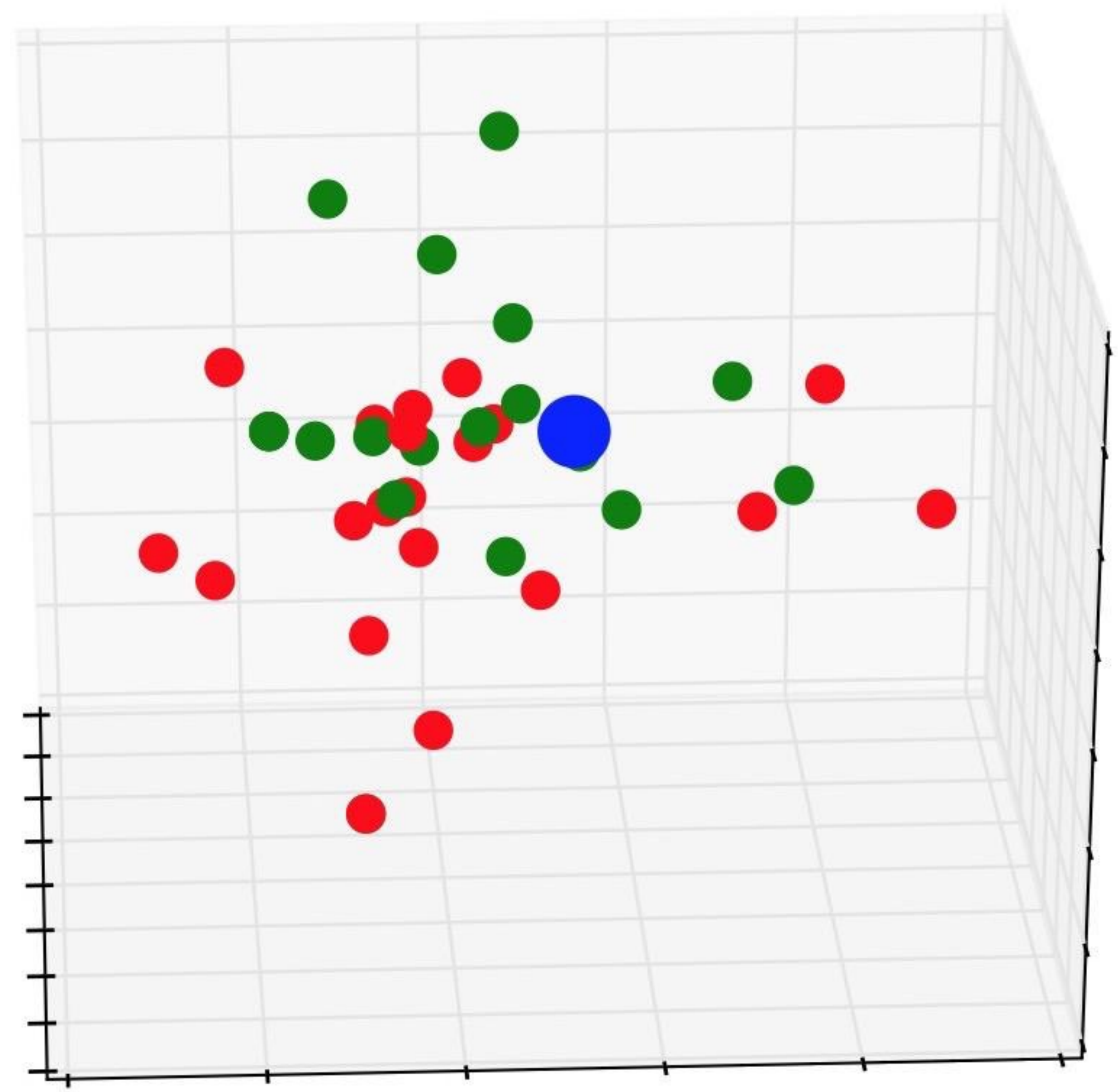
Compositional or Not

- (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

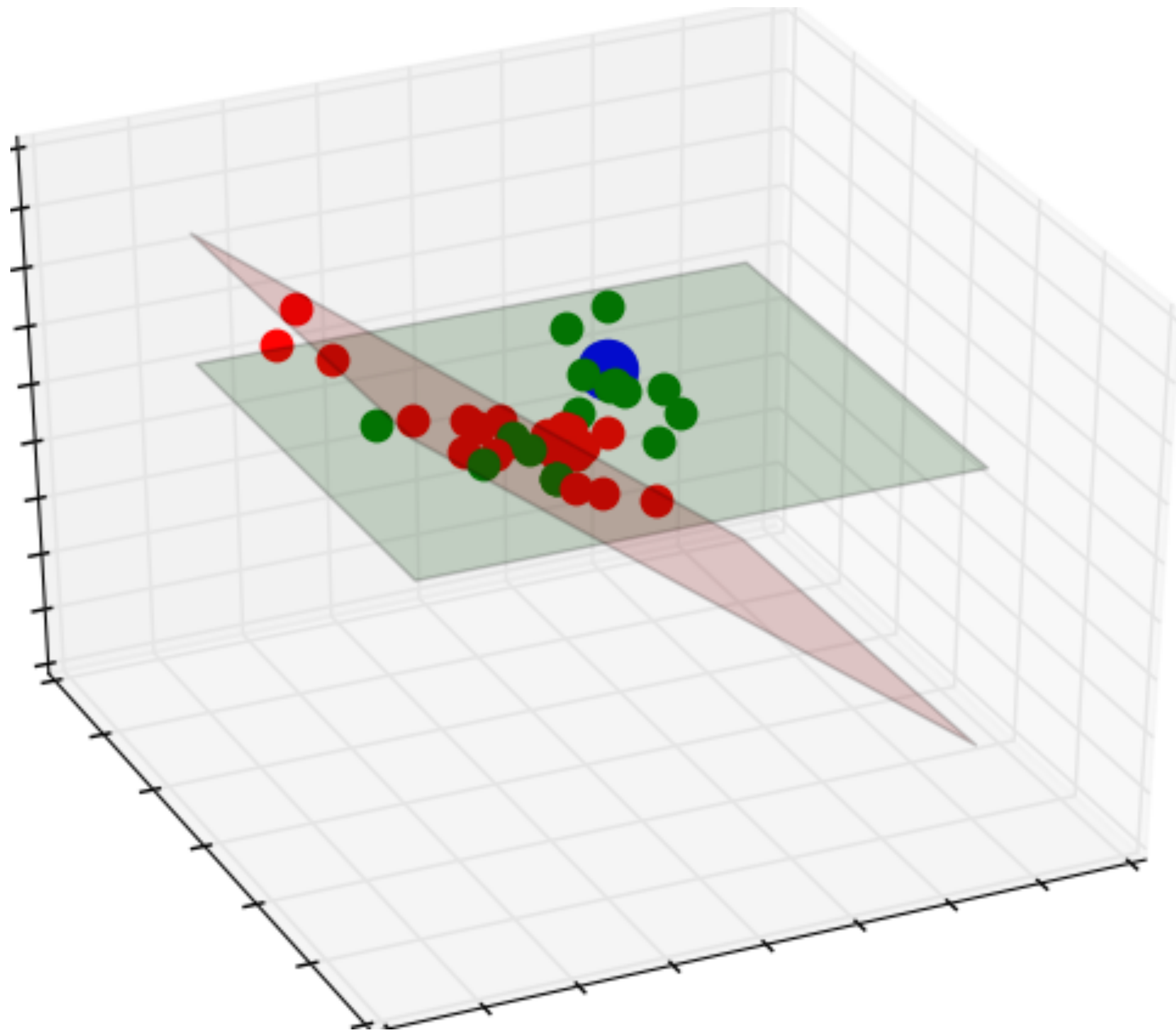
Geometry of Context Words

- "cutting edge"
- all words
  -- compositional
- all words
  -- idiomatic



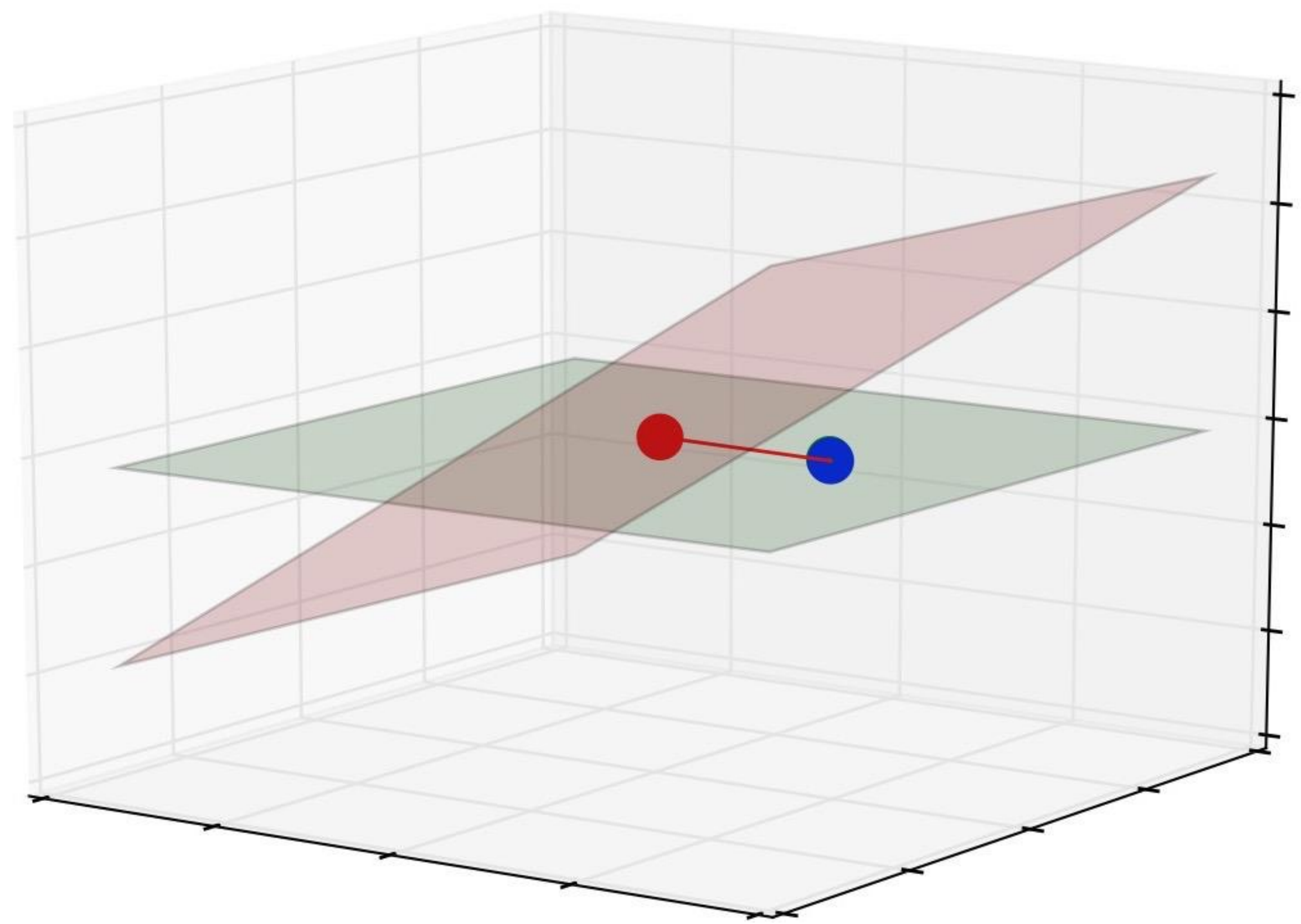
Geometry of Context Subspace

- "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

Irony

- **Ironic**

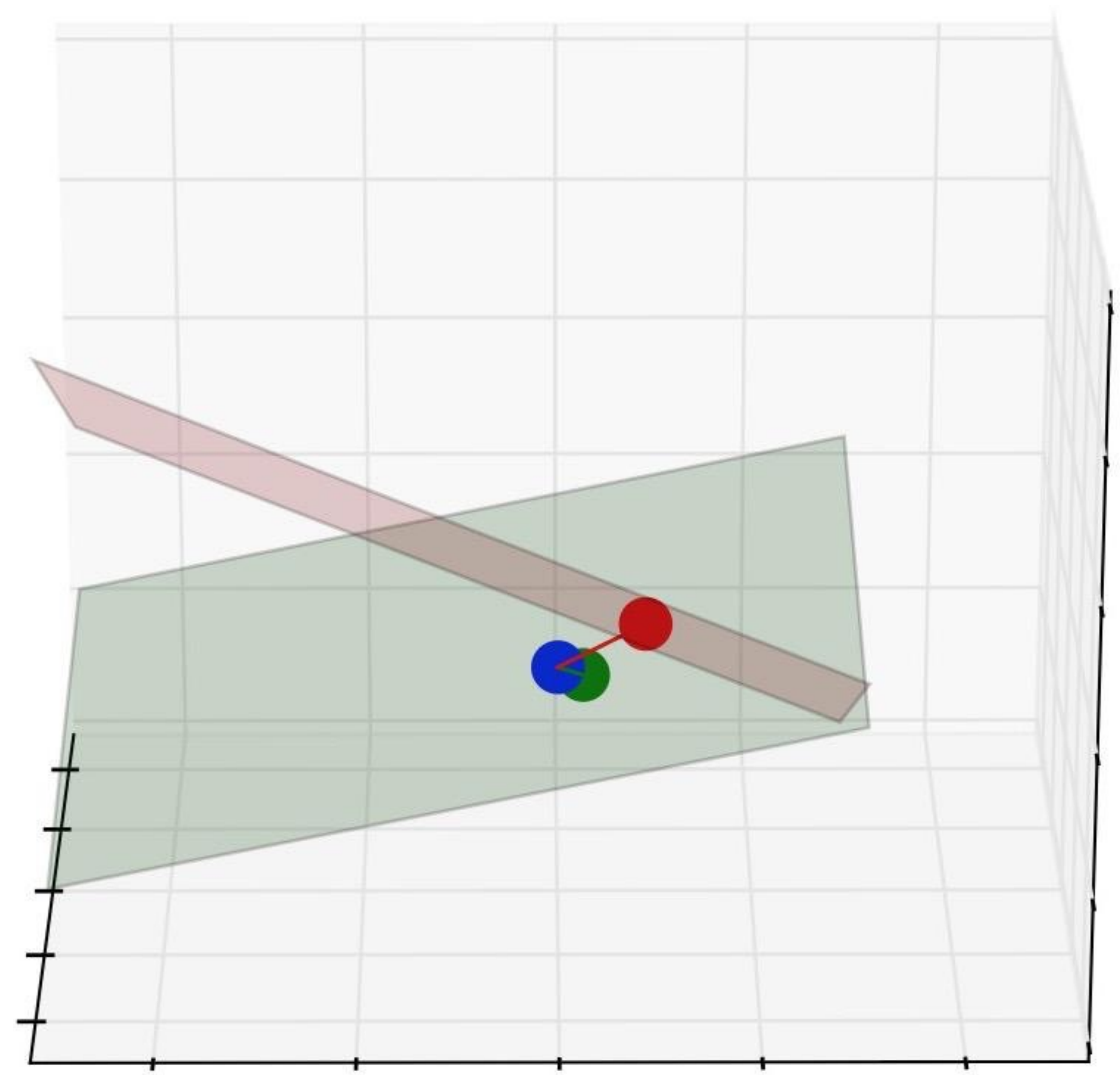
I **Love** going to the dentist! Looking forward to it all week.

- **Non-ironic**

Love to hear that youthcamp was so awesome!

Subspace-based Algorithm

- "glad"
- sentence subspace -- non-irony
- sentence subspace -- irony



- Irony detection: distance from target phrase to context space

Metaphor

- Figurative speech that refers to one thing by mentioning another

  - **Metaphor**

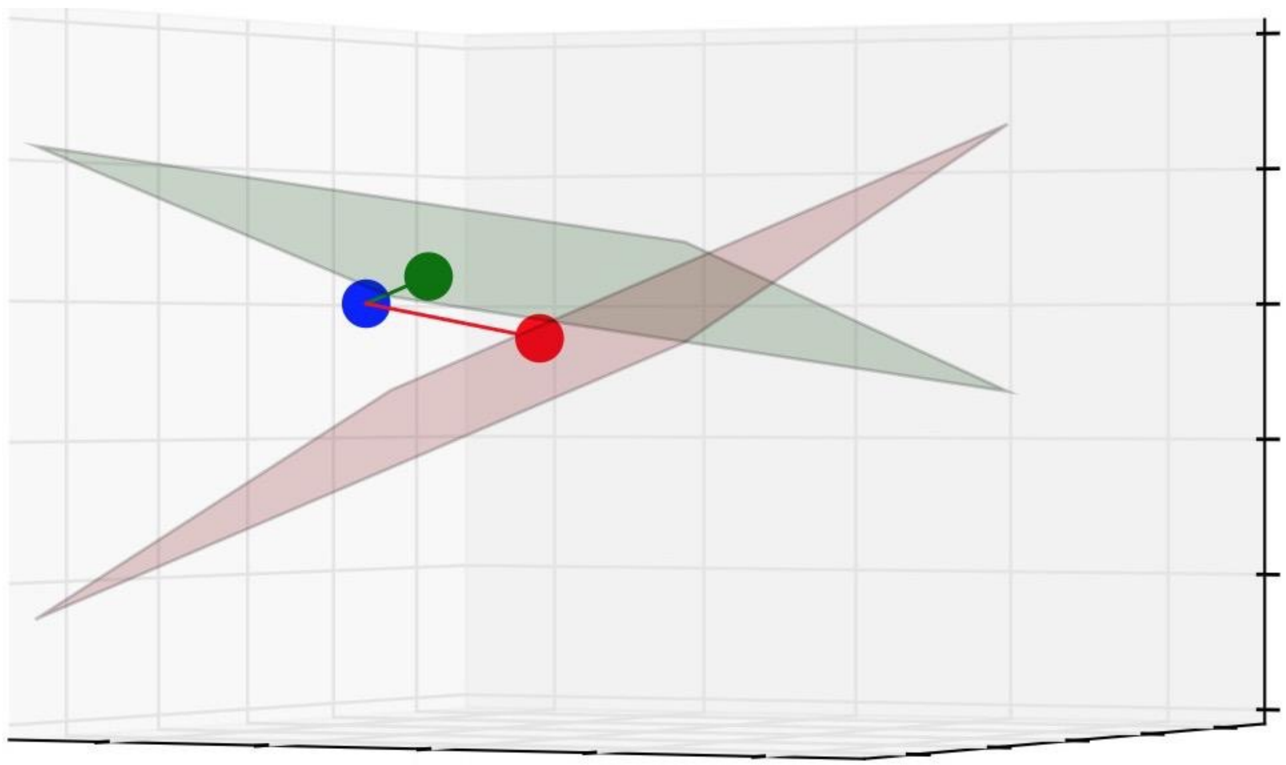
  They often **wear** an attitude that says – 'I can get away with anything’

  - **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

Dataset	Method	F1 score (%)
ENC Dataset	***State-of-art***	75.5
	This talk	**84.2**
EVPC Dataset	***State-of-art***	39.8
	This talk	**46.2**
GNC Dataset	***PMI***	61.1
	This talk	**62.4**

Dataset	Method	Accuracy (%)
Chinese Dataset	***Baseline***	78.1
	This talk	**88.3**

Prepositions: Polysemous Nature

“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*
- ….

Context Implying True Sense

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

[1,2] Kenneth C Litkowski and Orin Hargraves. 2005. The Preposition Project.
[3] Samuel Ritter, et al. 2015. Leveraging preposition ambiguity to assess compositional distributional models of semantics.

Intrinsic Evaluation: SemEval

System	Resources	Accuracy
Our system	English corpus	**0.80**
Litkowski, 2013	Lemmatizer, dependency parser	0.86
Srikumar and Roth, 2013	dependency parser, WordNet	0.85
Gonen and Goldberg, 2016	multilingual corpus, aligner, dependency parser	0.81
Ye and Baldwin, 2007	chunker, WordNet dependency parser	0.69

Intrinsic Evaluation: OEC

System	Resources	Accuracy
Our system	English corpus	**0.40**
Litkowski, 2013	Lemmatizer, dependency parser, WordNet	0.32

Intrinsic Evaluation: Spatial Relation

Preposition	Spatial Relation	Example
in	Full Containment	apple in the bag
	Partial Containment	finger in the ring
on	Adhesion to Vertical Surface	sign on the building
	Support by Horizontal Surface	leaf on the ground
	Support from Above	bat on the branch

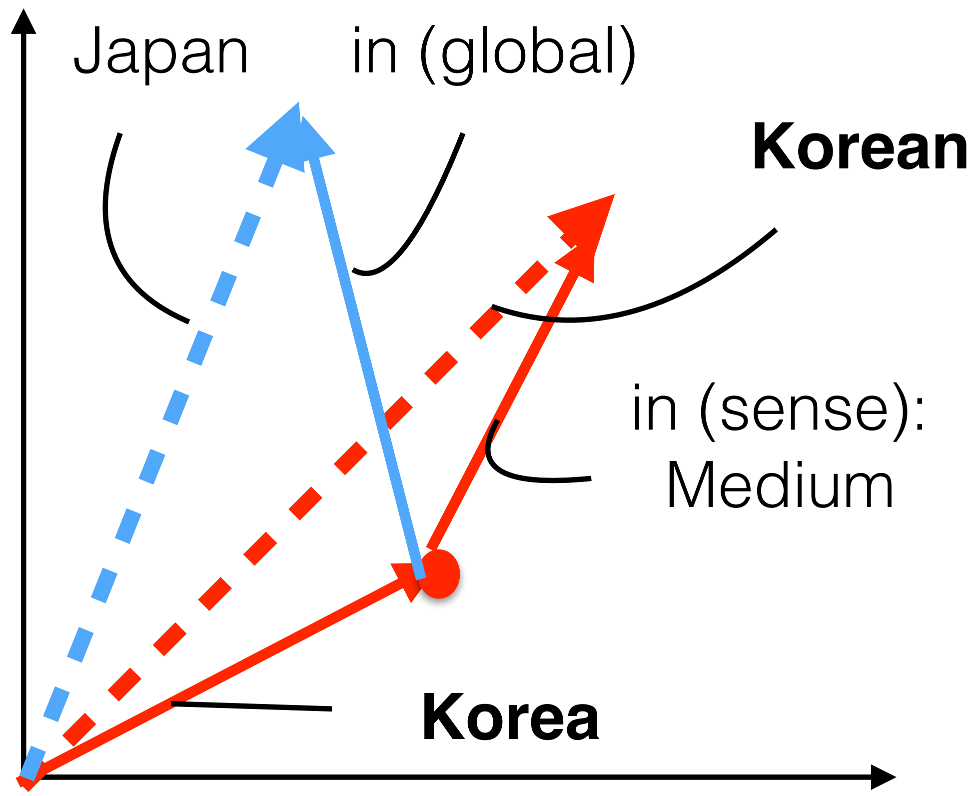
Our system achieves an accuracy of 77.5%, compared with 71% achieved by the state-of-art

Extrinsic Evaluation

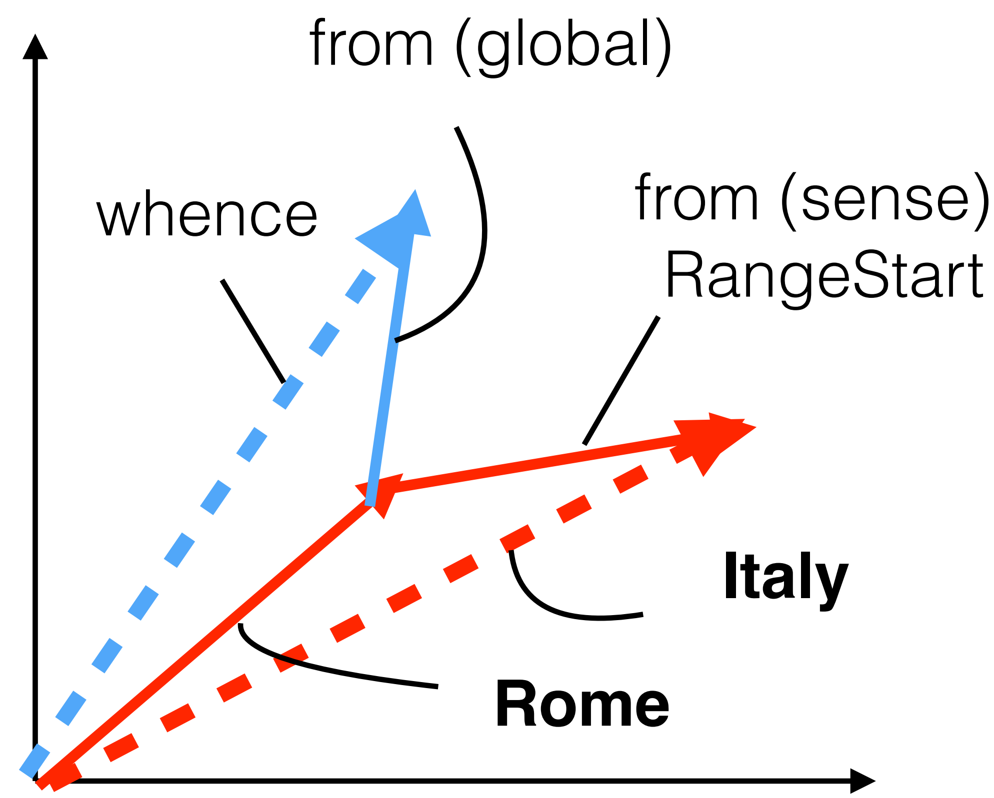
- Light-weight disambiguation system
  - no reliance on external linguistic resources
- Efficient scaling to enrich large corpus
  - train sense representations
- Extrinsic evaluation
  - semantic relation
  - paraphrasing of phrasal verbs

Extrinsic Evaluation: Semantic Relation

- Sense representations encode relations

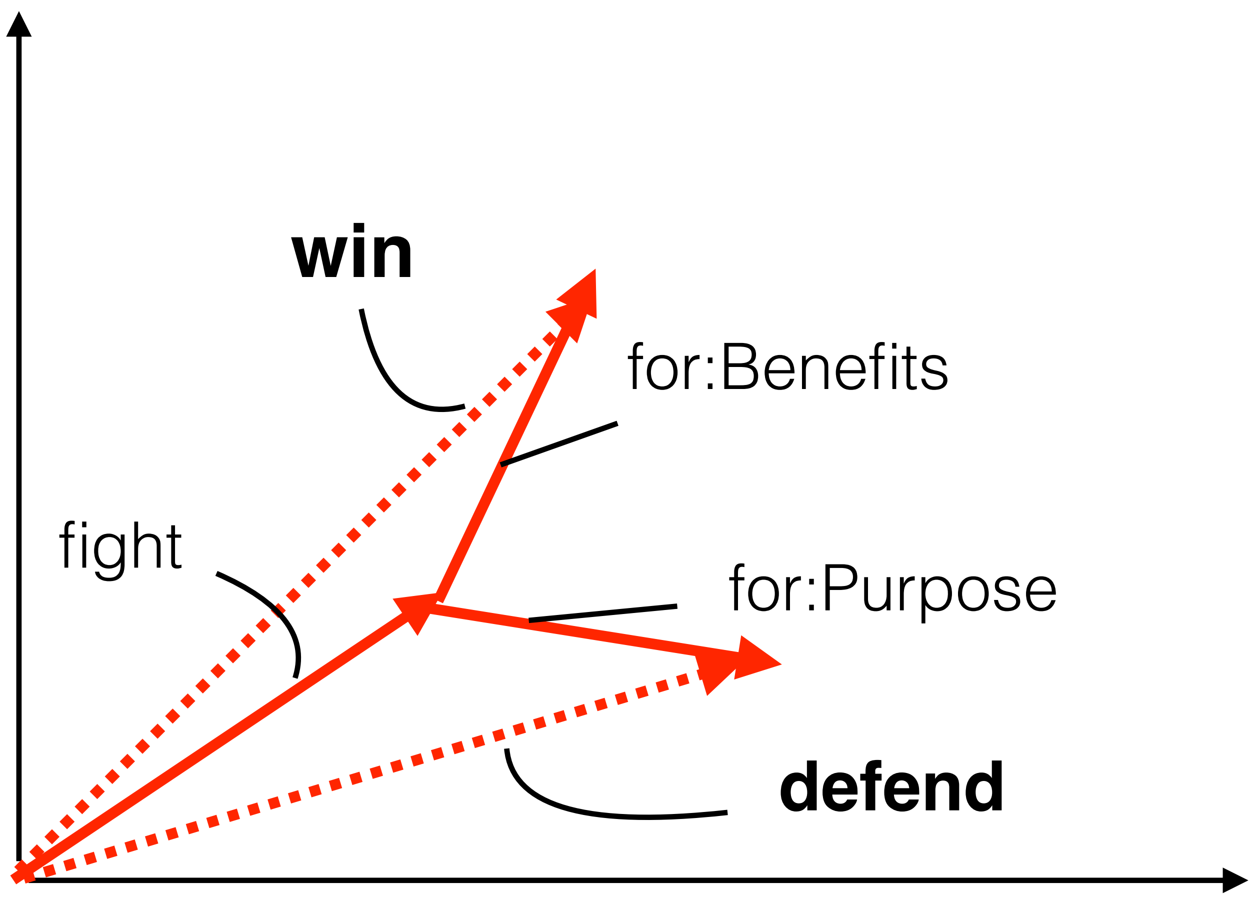


in (Location) + Korea ~ Korean



from (RangeStart) + Rome ~ Italy

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