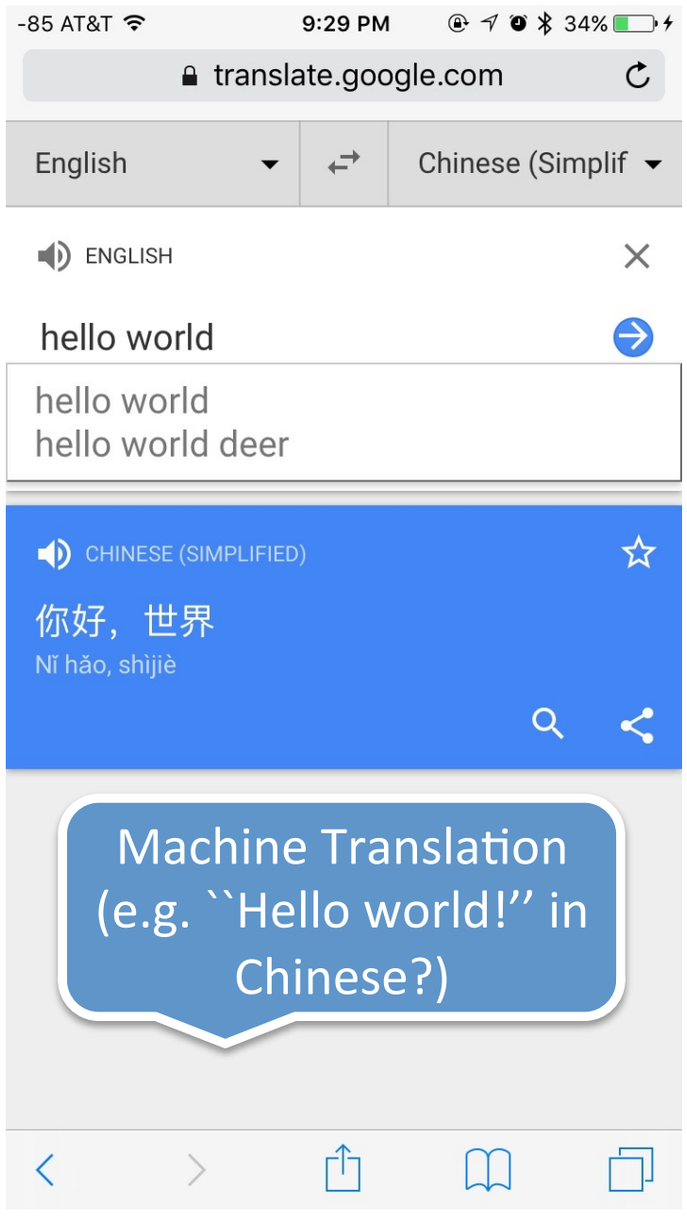
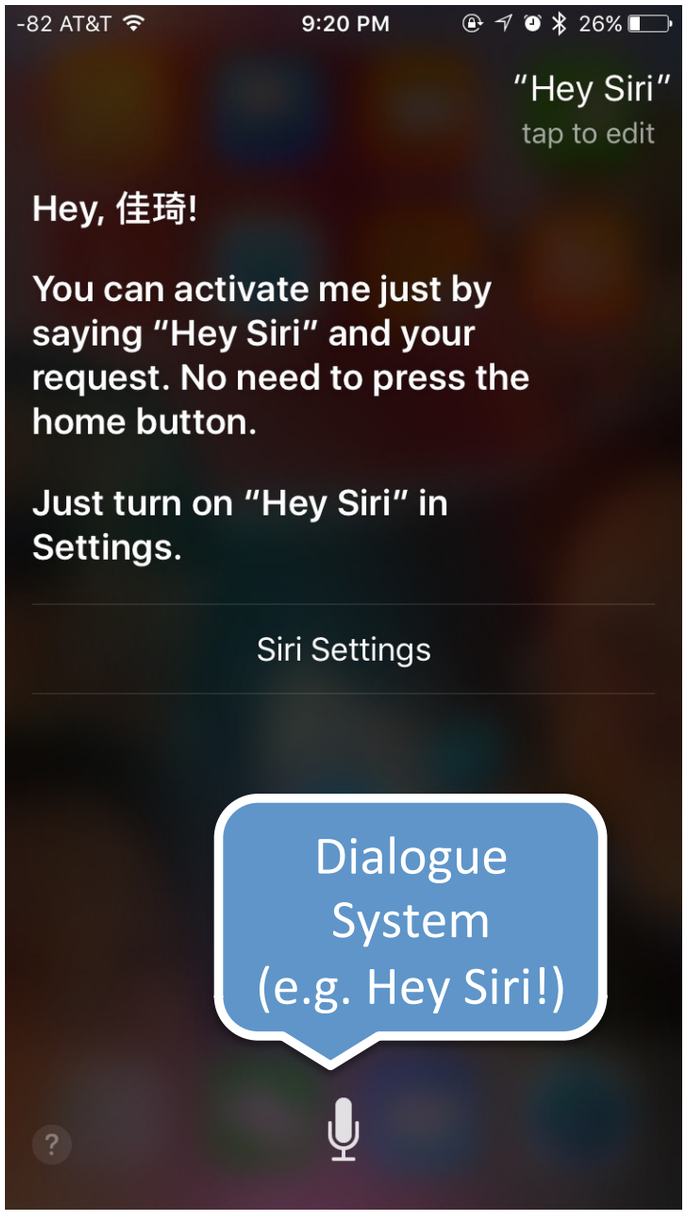
Geometry of Word Embeddings and Its Applications

Jiaqi Mu (jiaqimu2@illinois.edu)
joint work with Hongyu Gong, Pramod Viswanath and Suma Bhat

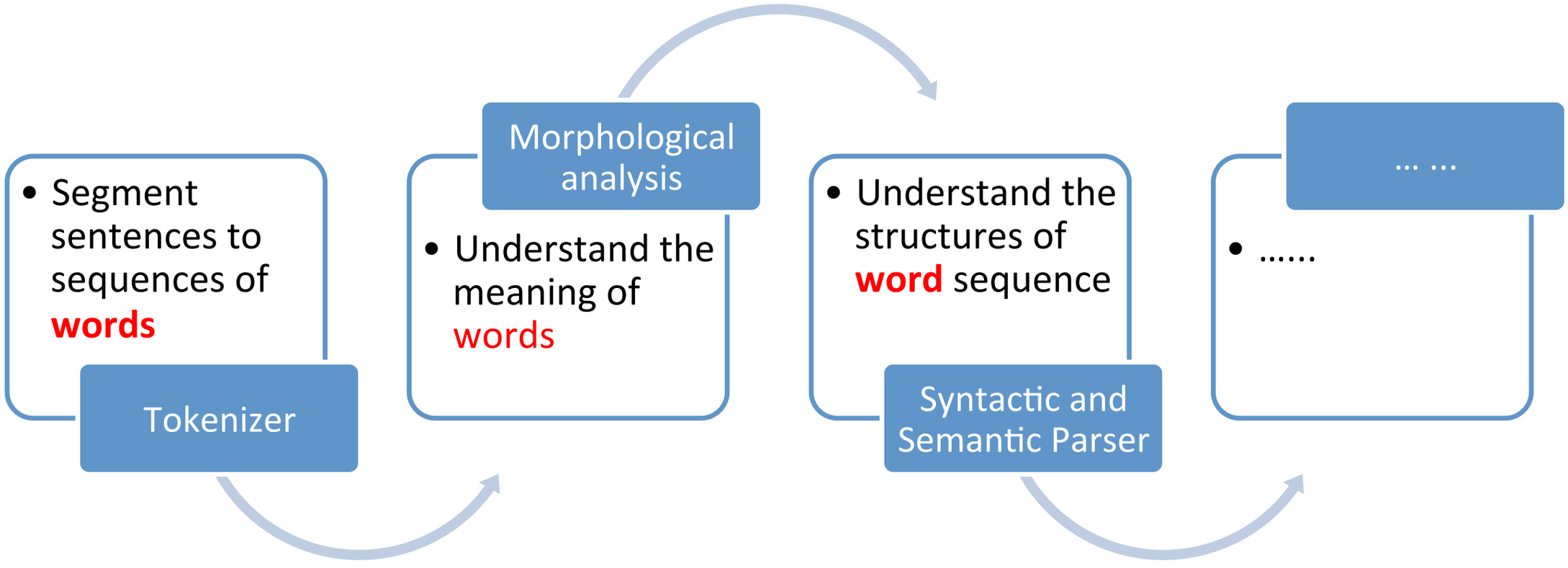


natural language processing



Natural language processing is widely used in daily life.

Natural Language Processing Pipeline



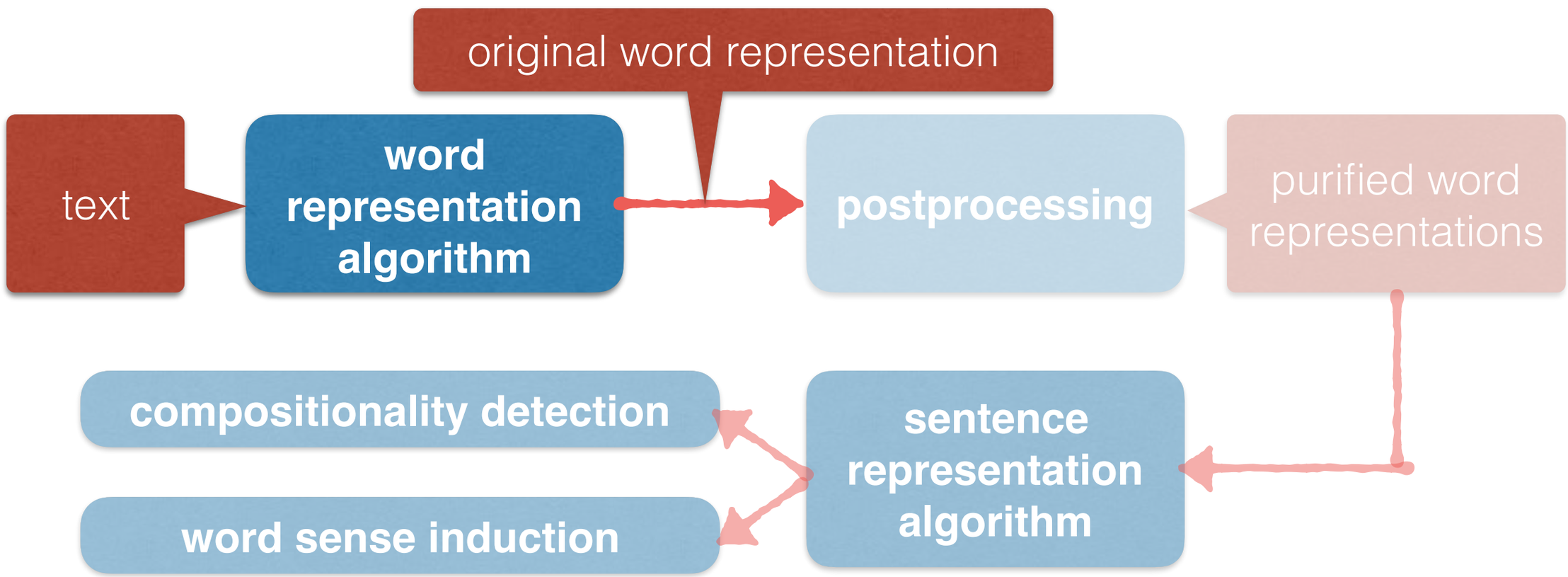
Word is the basic unit of natural language.

How to Represent Words?

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

Words could be represented by vectors.

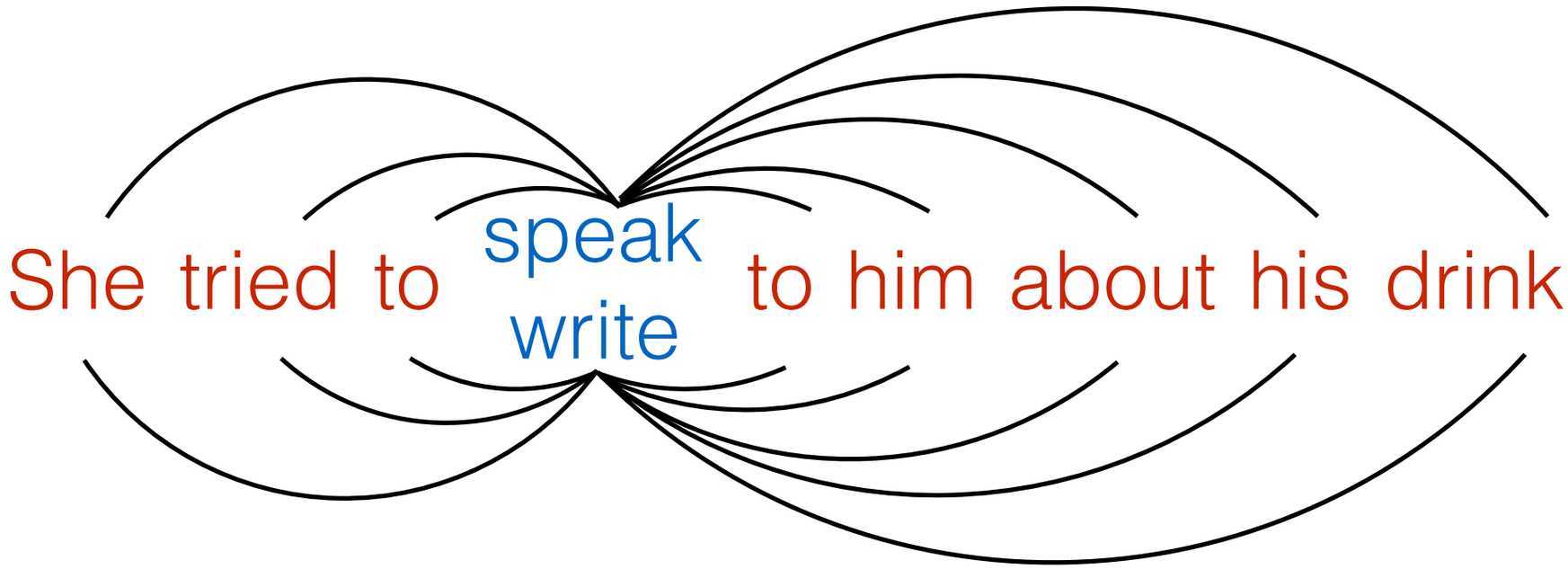
Word2vec: What and Why



What to Preserve?

"A word is characterized by the company it keeps."

— Firth 1957



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	…
…	…	…	…

Factorization of PMI Matrix

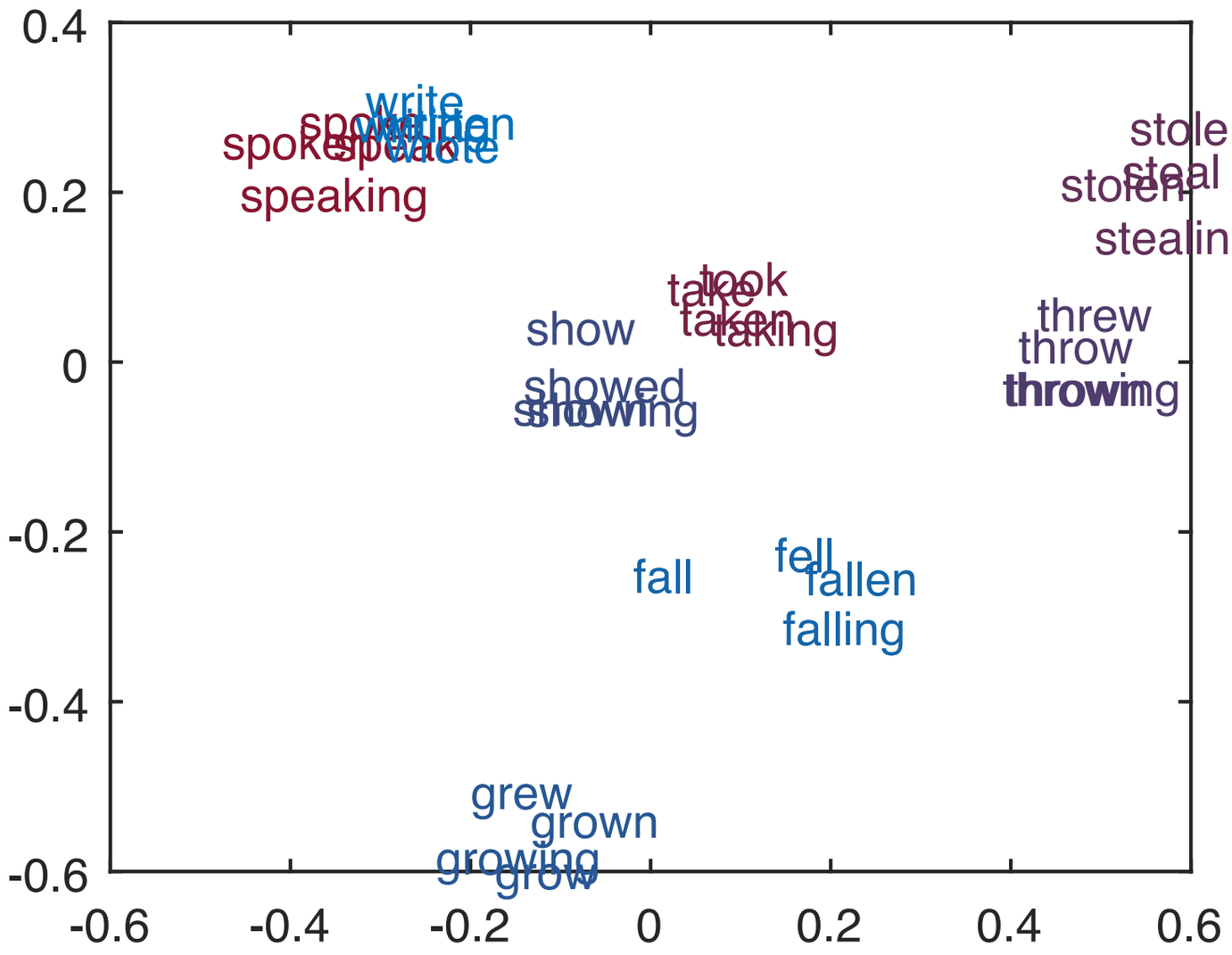
word2vec (Mikolov 2013) and **GloVe** (Pennington 2014)

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

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

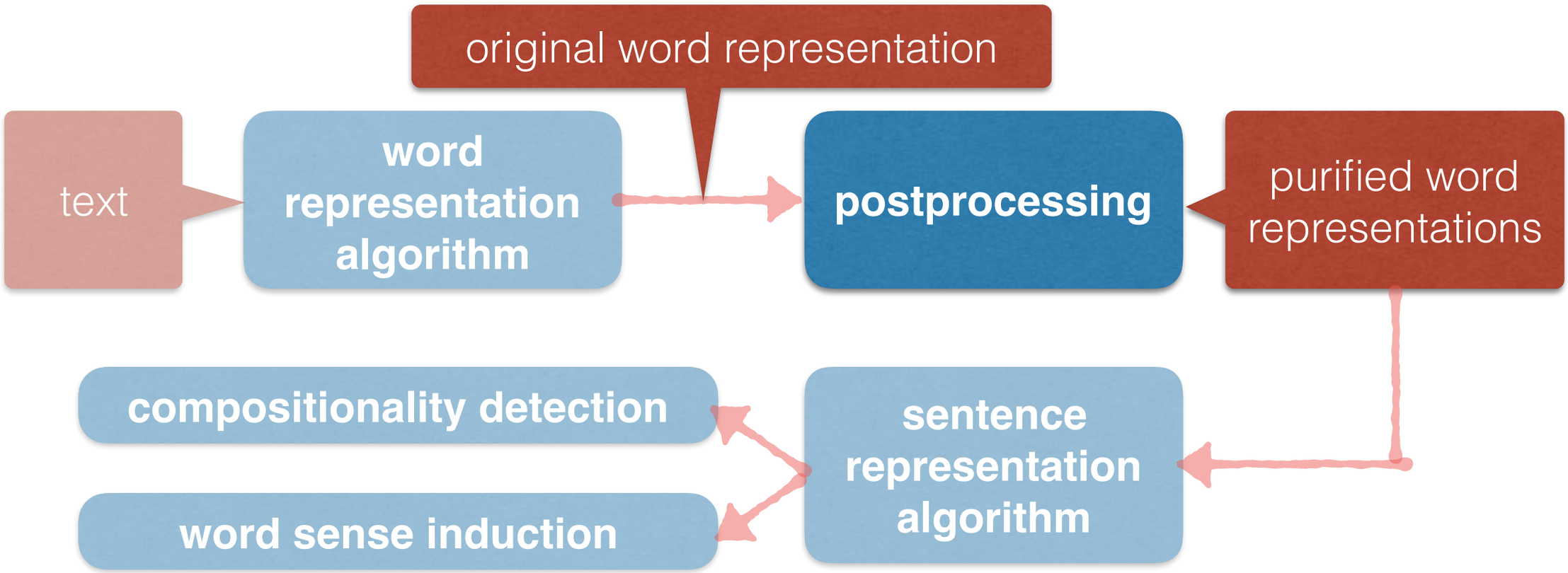
Ref: https://arxiv.org/abs/1502.03520

Word Similarity



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

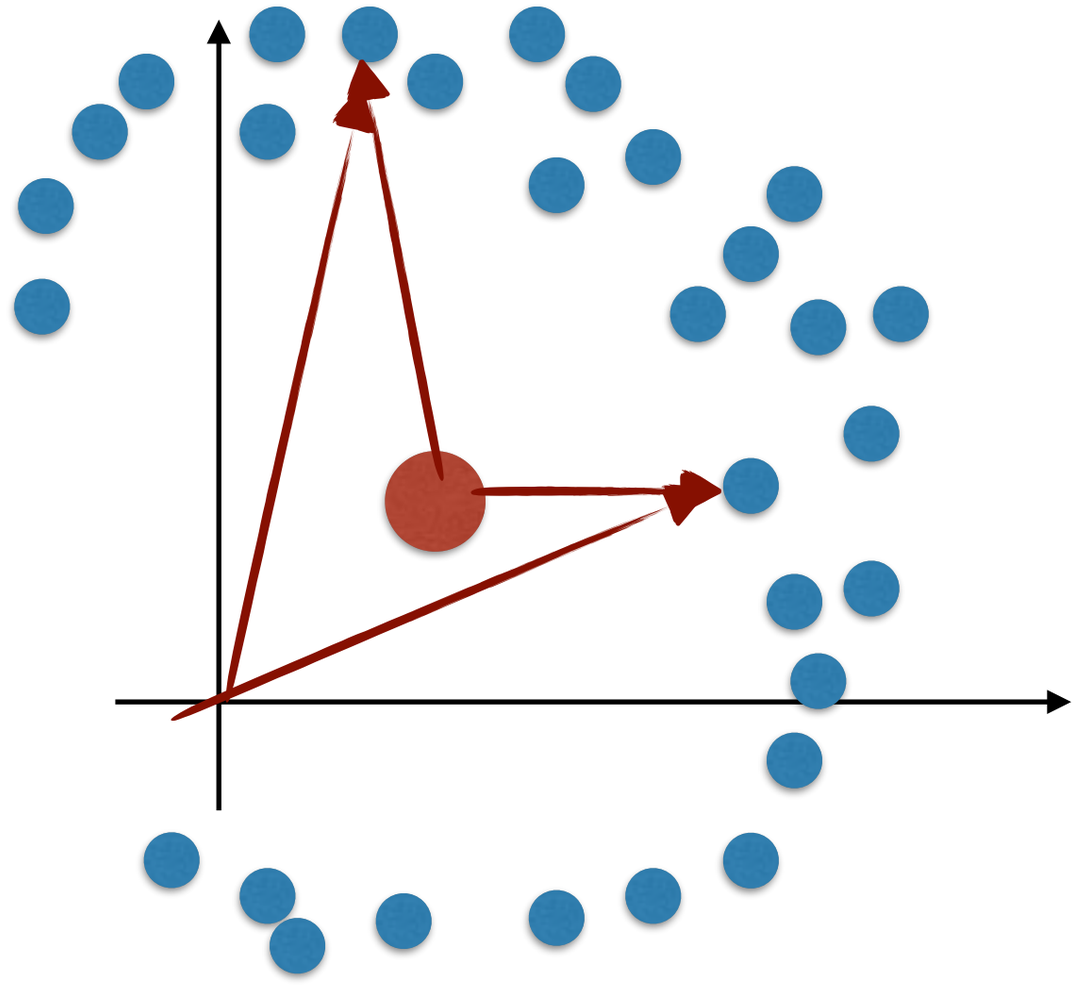
All-but-the-top: a Simple but Effective Post-processing Algorithm



Ref: https://arxiv.org/abs/1702.01417

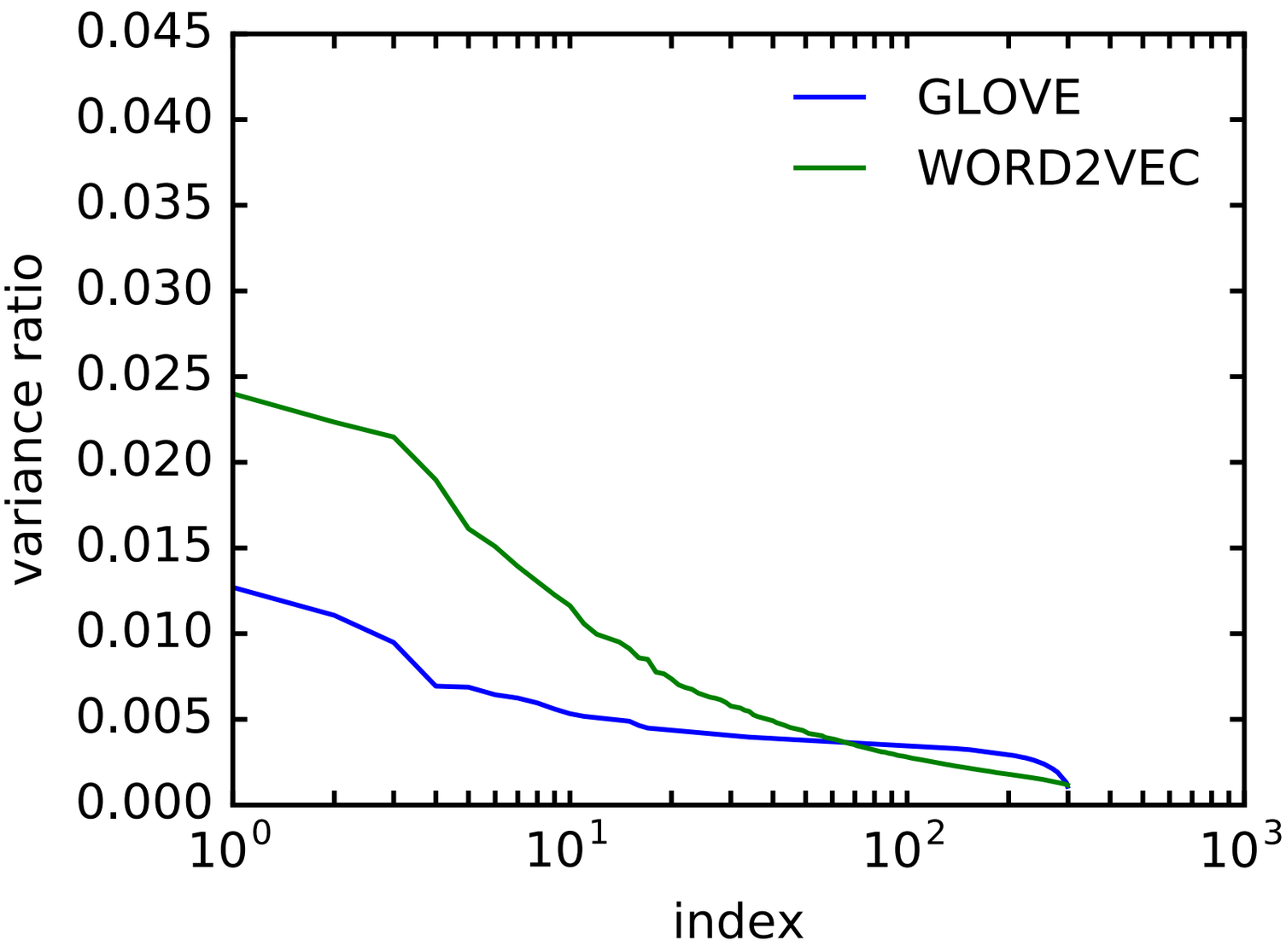
Statistics of Word Vectors

	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

Statistics of Word Vectors



Postprocessing

- Remove the non-zero mean of $\{v(w), w \in V\}$

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

- Remove the dominating D components

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

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

A simple post processing 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	0.95
GloVe	2.43

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

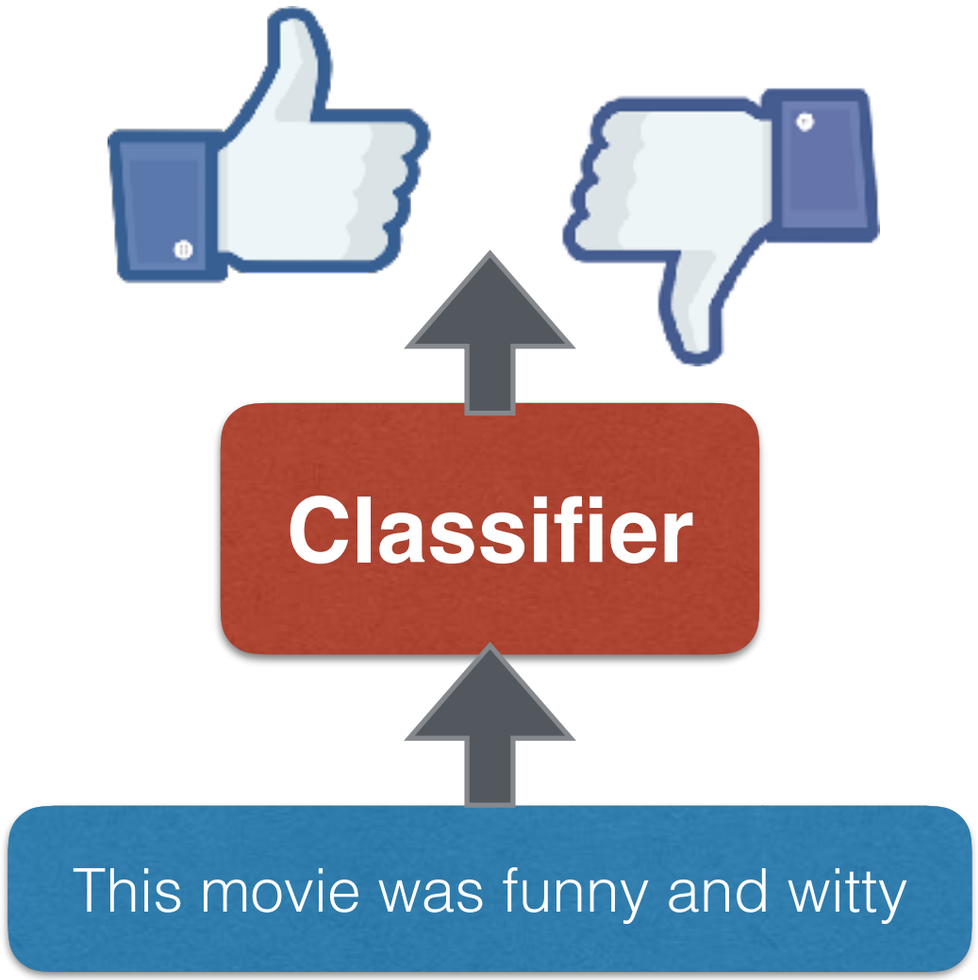
	avg. improvement
word2vec	7.37
GloVe	0.62

Datasets: ap, ESSLLI, battig

Sentence-level Evaluation

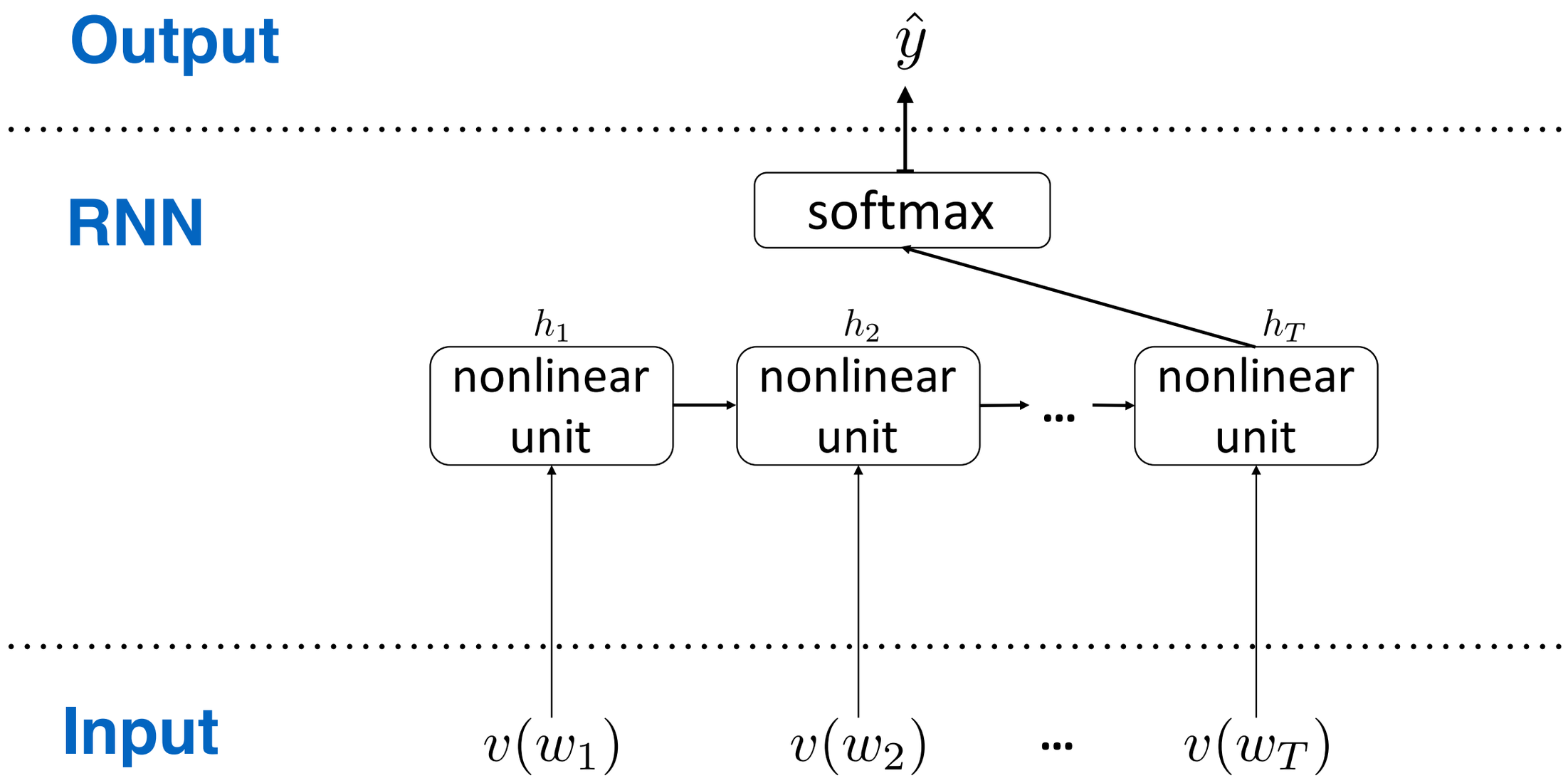
- ✓ Sentiment Analysis

Sentiment Analysis

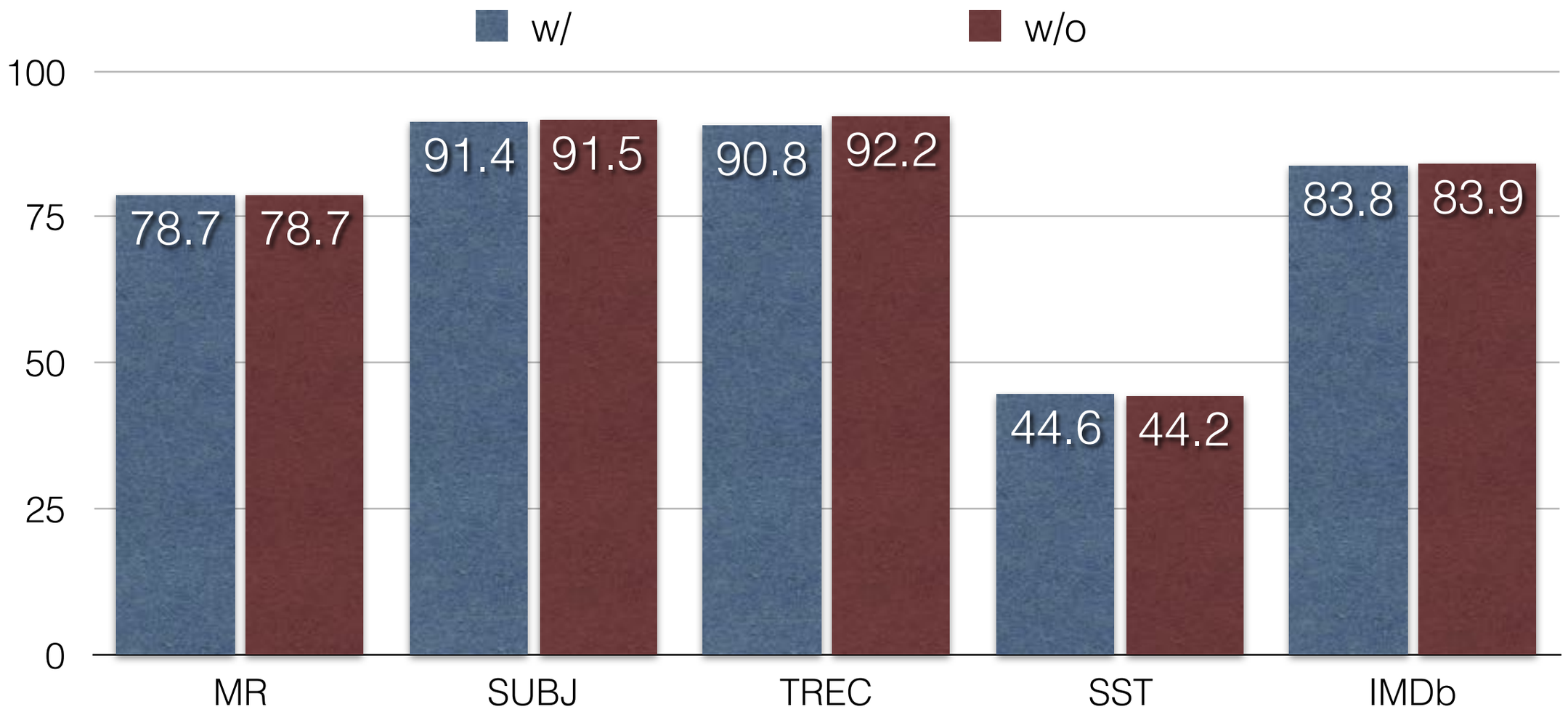


Postprocessing in downstream applications

Recurrent Neural Networks

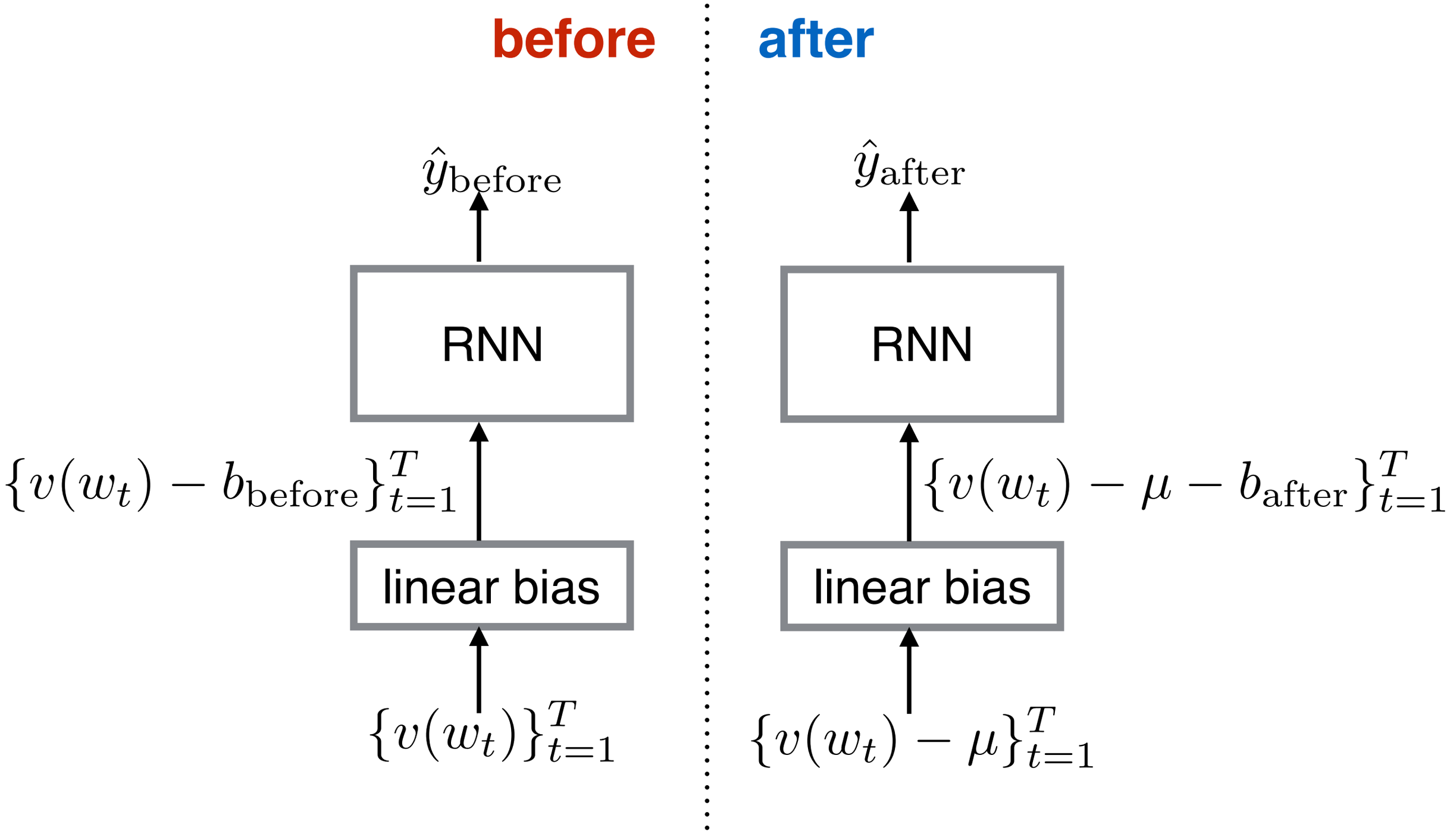


Experiment Results



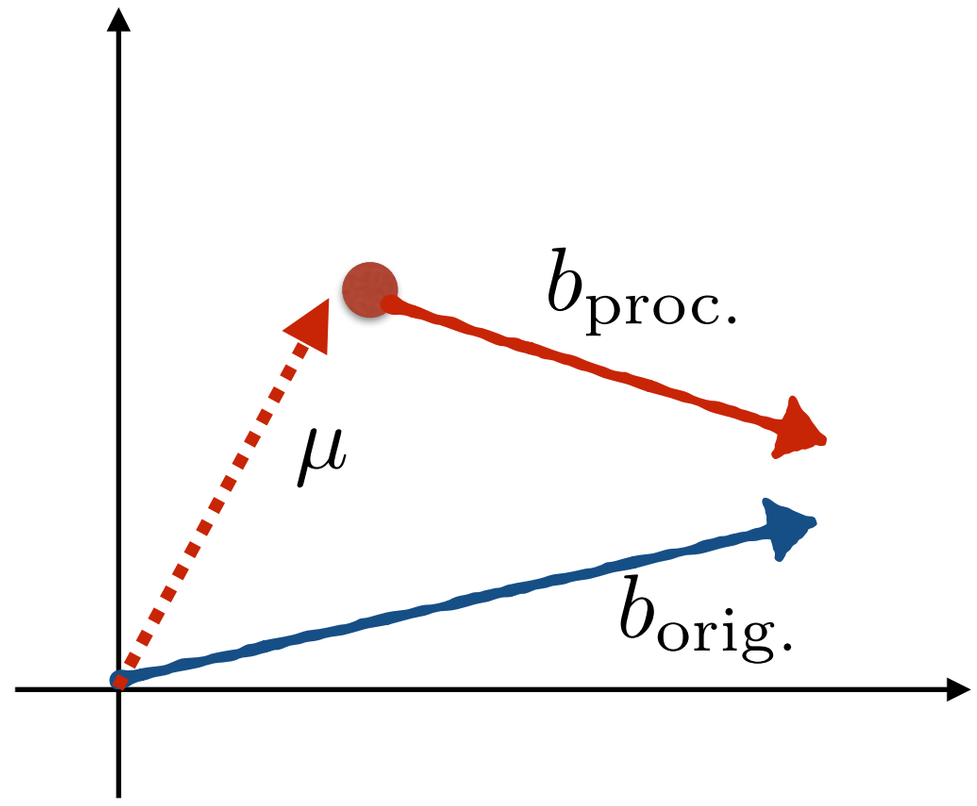
The performances are similar for vectors w/ and w/o postprocessing.

RNN with An Appended Layer



Linear bias <-> Mean Vector

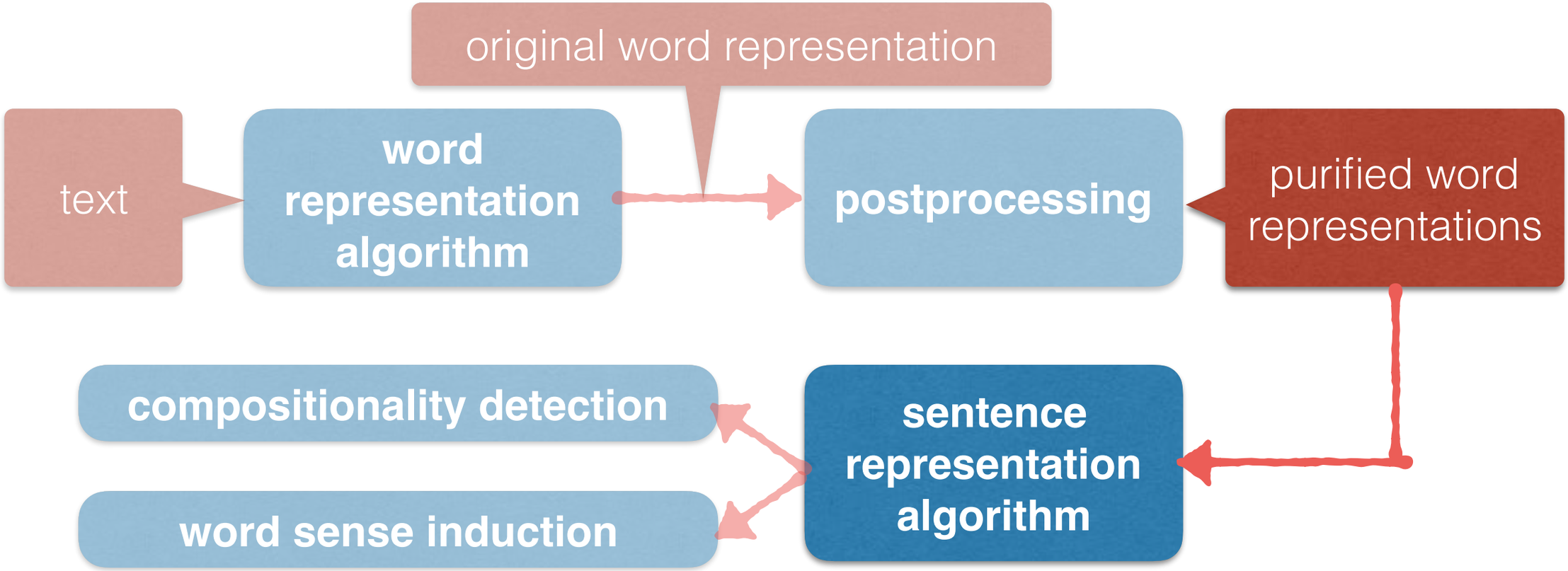
Jointly trained b_{after} and $b_{\text{before}} + \mu$ are highly correlated



avg. cosine similarity 0.69

The Neural network architectures ``automatically learn" the postprocessing operation.

Geometry of Sentences:
Compositionality and Polysemy

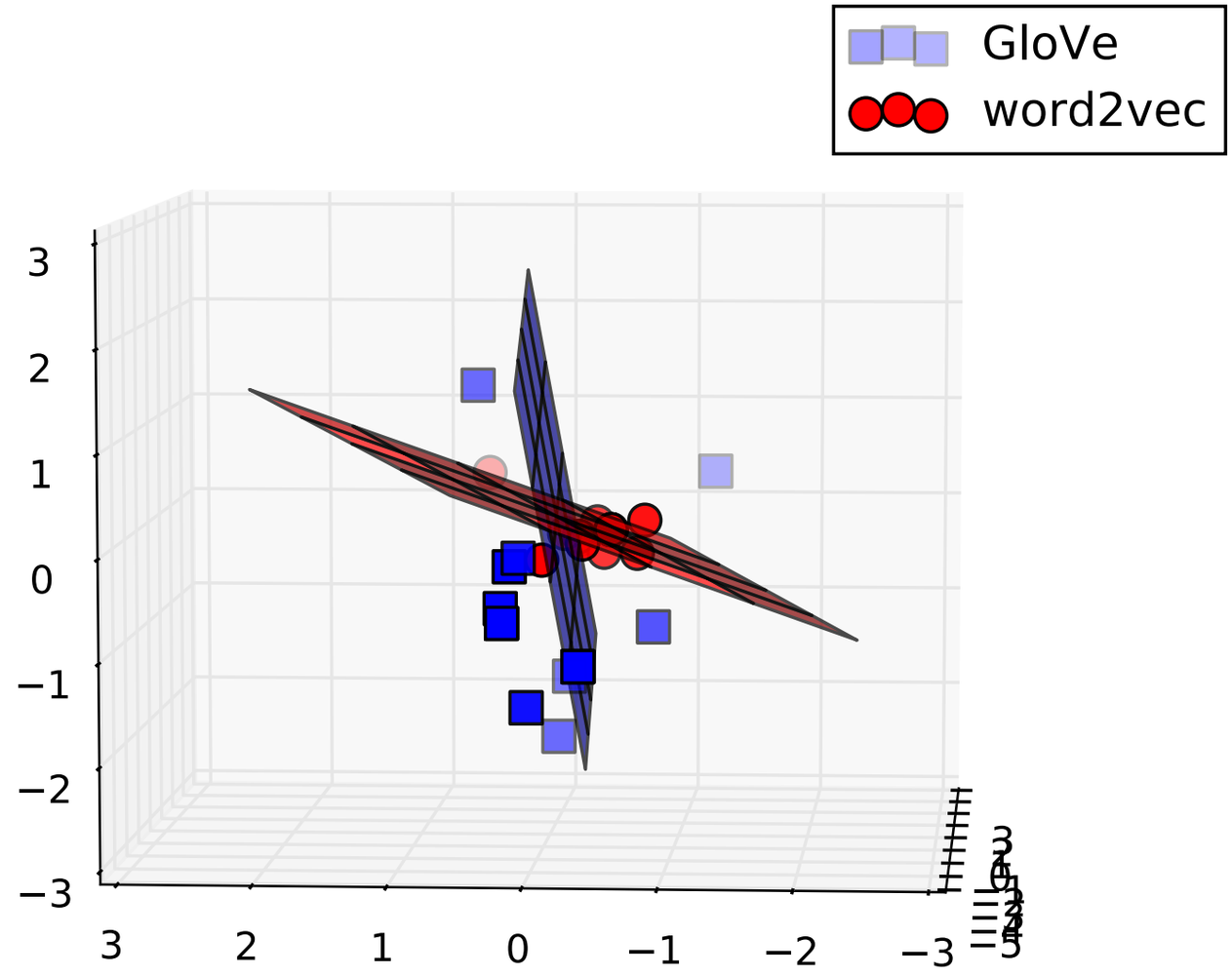


How to Represent Sentences?

- Neural networks:
  - Kim 2014, Kalchbrenner 2014, Sutskever 2014, Le and Mikolov 2014, Kiros 2015, Hill 2016
- Average of word vectors:
  - Wieting 2015, Huang 2012, Adi 2016, Kenter 2016

Sentences are usually represented by vectors

Grassmannian Manifold



"They would not tell me if there was any pension left here, and would only tell me if there was (and how much there was) if they saw I was entitled to it."

Sentences can also be represented by subspaces

Subspace Representation

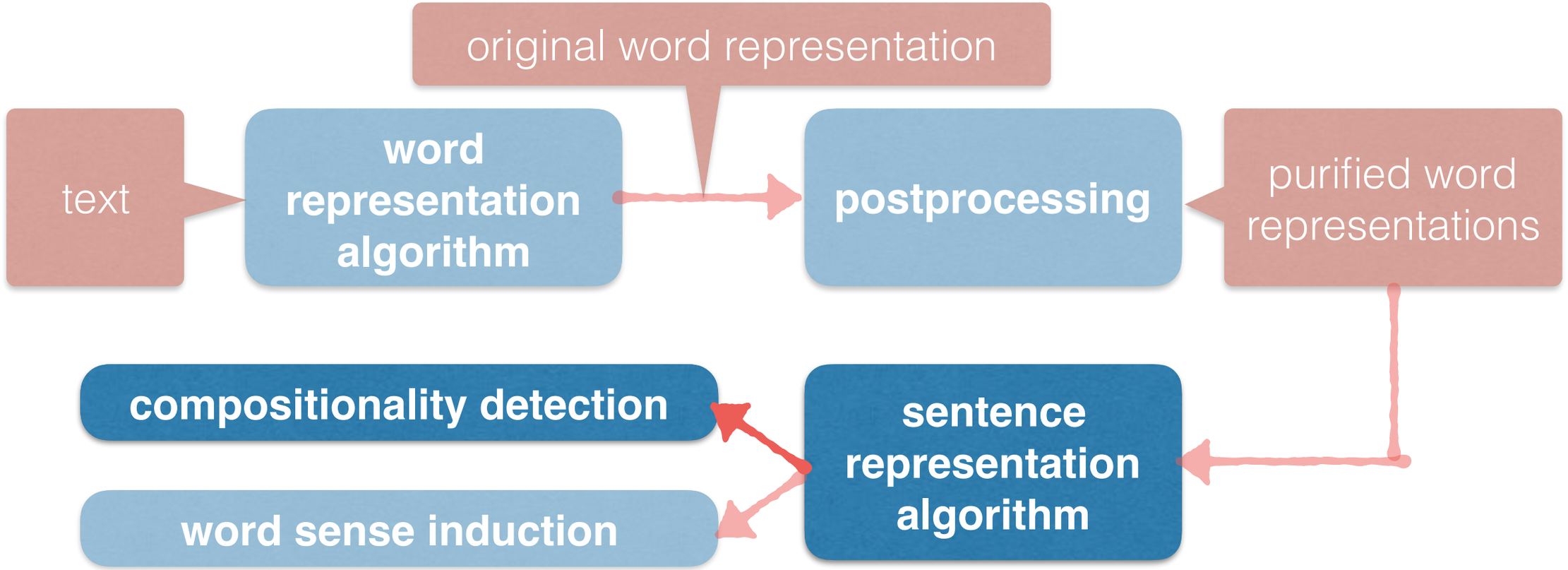
- Input: a sequence of words $s = \{w_1, ..., w_t\}$
- Compute the first N principal components of $\{v(w'), w' \in s\}$

$$u_1, ..., u_N \leftarrow \text{PCA}(v(w'), w' \in s),$$

$$S \leftarrow \left\{ \sum_{n=1}^{N} : \alpha_n u_n, \alpha_n \in R \right\}$$

- Output: N orthonormal basis $u_1, ..., u_N$ and a subspace S

Compositionality Detection



Ref: https://arxiv.org/abs/1611.09799

Non-compositional Phrases

- (English) he enjoys being a big fish, playing with the politicians who make a difference
- (Chinese) 在 當時 人 們看 來， 有 文化， 有 墨 水 的 人， 就 是 知 識 分子 。

Previous Work

- Wiktionary: list definitions, tag phrases as literal or idiomatic
- WordNet: Lexical supersenses
- Psycholinguistic database: infer feelings conveyed

Previous approaches rely heavily on linguistic resources

Our Approach

- Compositionality can be inferred from the context.

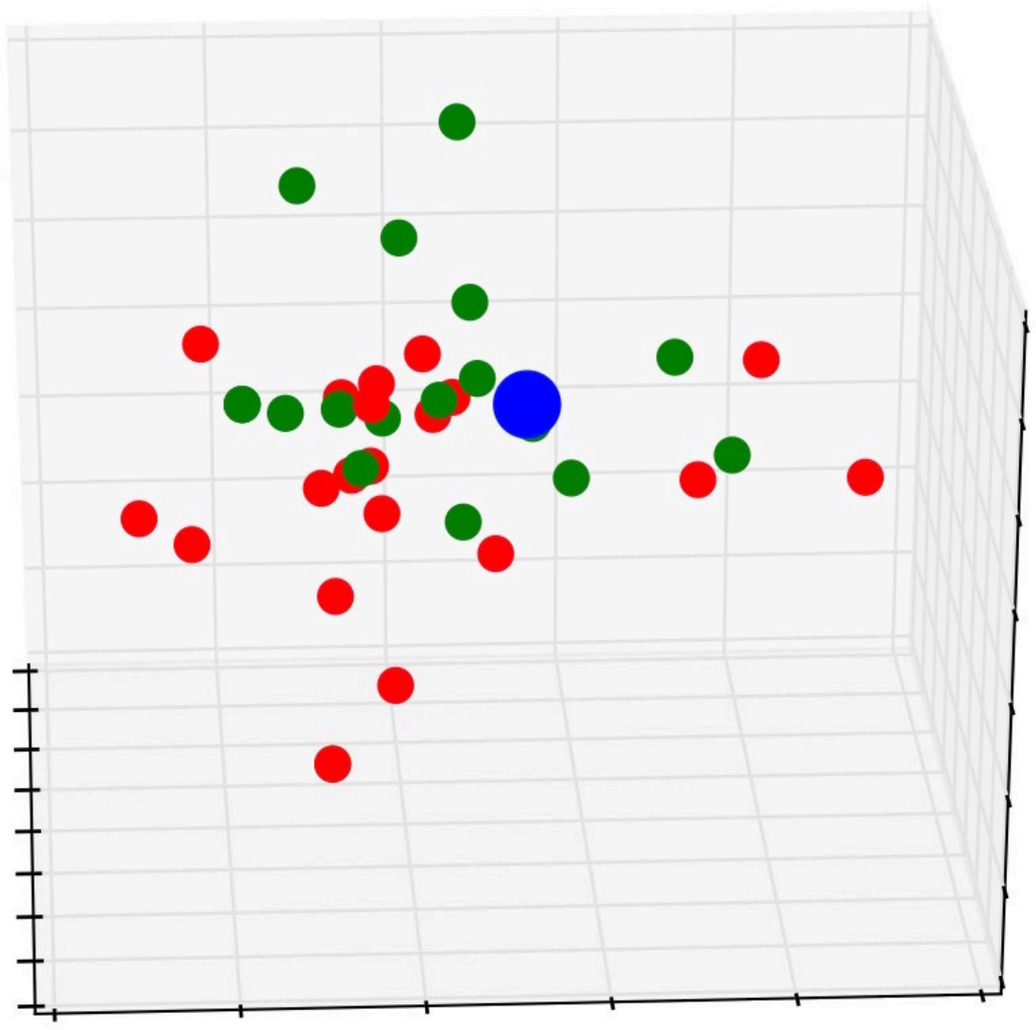
Compositional: Knife has a cutting edge, a sharp side formed by the intersection of two surfaces of an object

Idiomatic: Utilize his vast industry contacts and knowledge while creating a cutting edge artworks collection

Word embeddings facilitate geometry

Geometry of Context Words

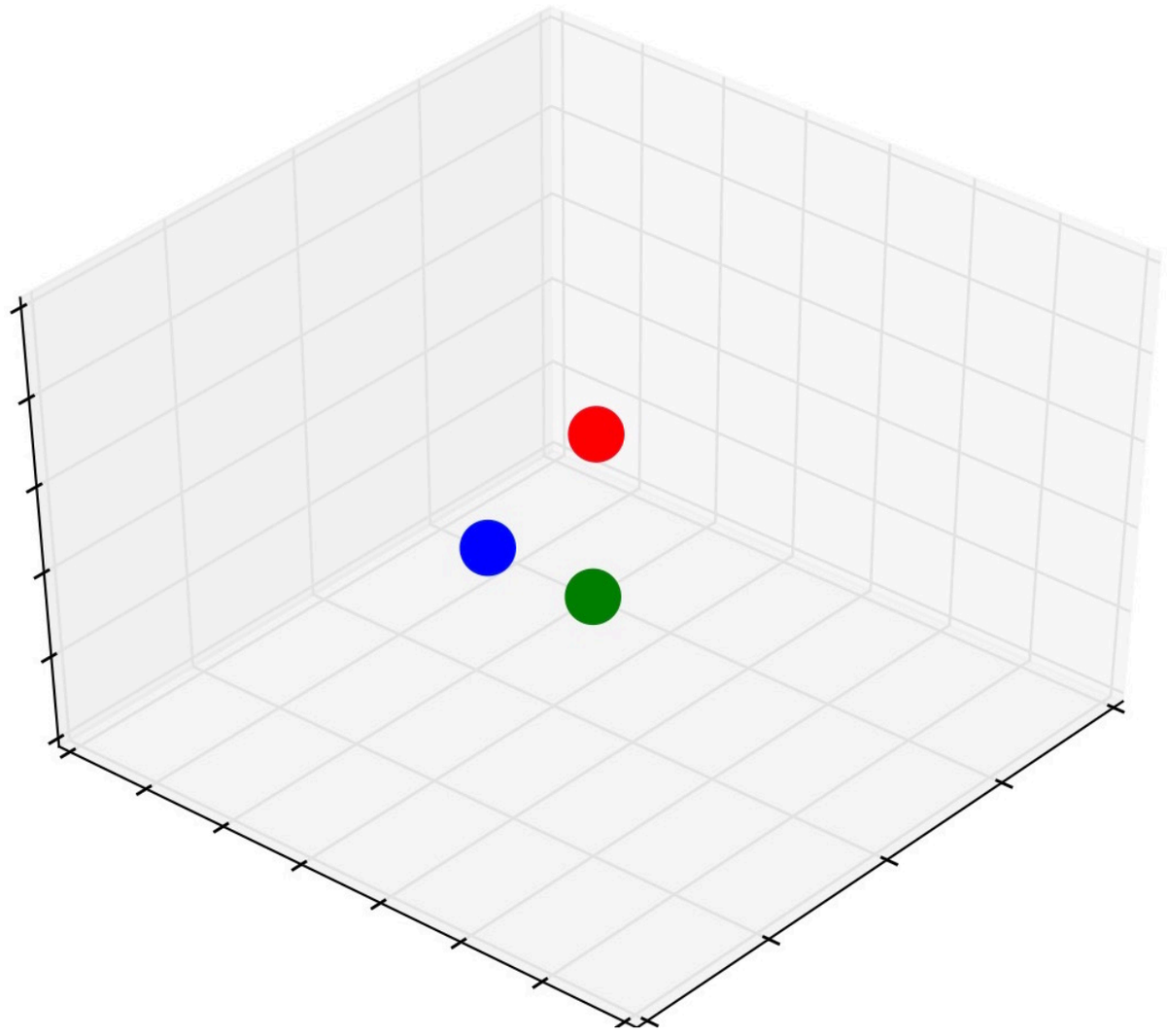
- “cutting edge”



- context words in compositional sentences
- context words in non-compositional sentences

Average Representation

- "cutting edge"
- average of the compositional context
- average of the non-compositional context

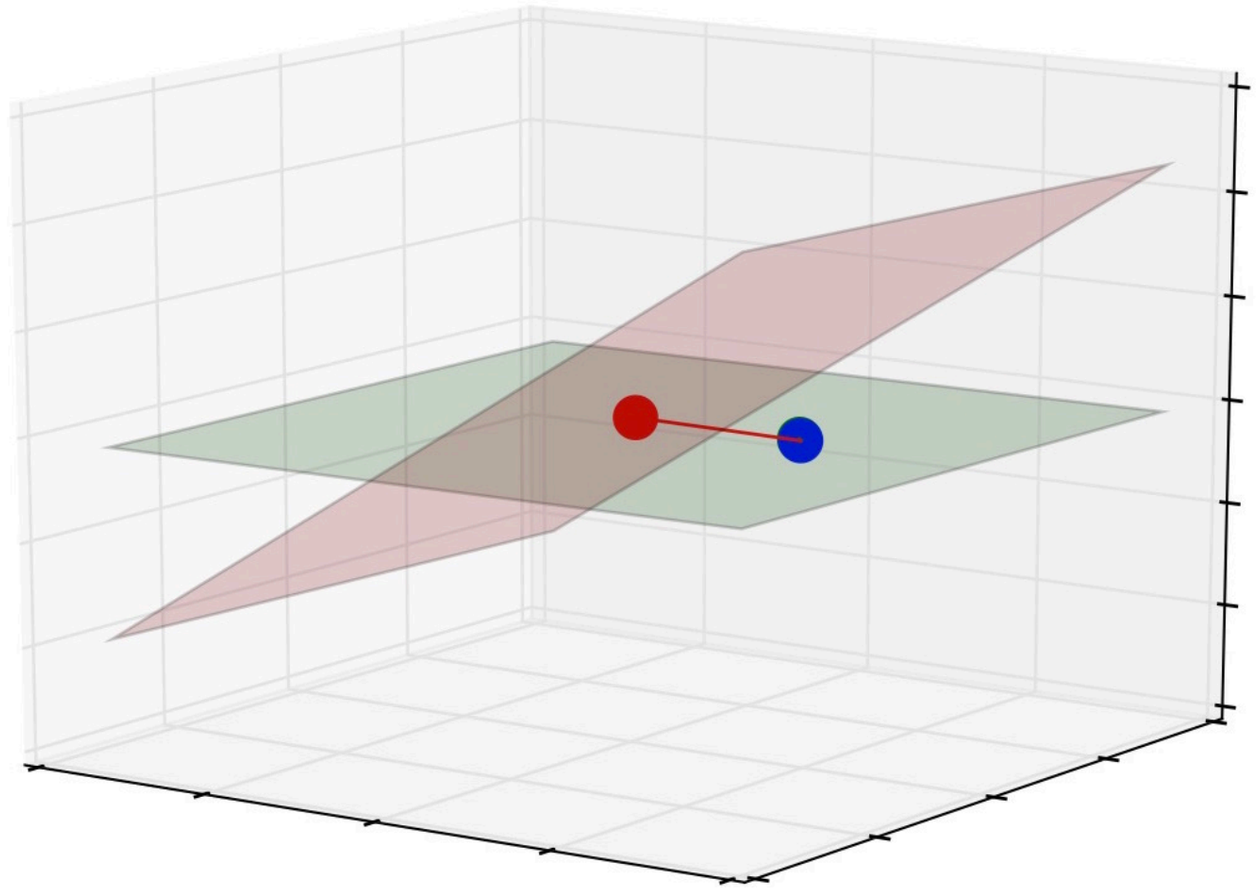


Subspace Representation

- "cutting edge"

- subspace of the compositional context

- subspace of the non-compositional context



Detection Algorithm

- Extract principal components from stacked context words

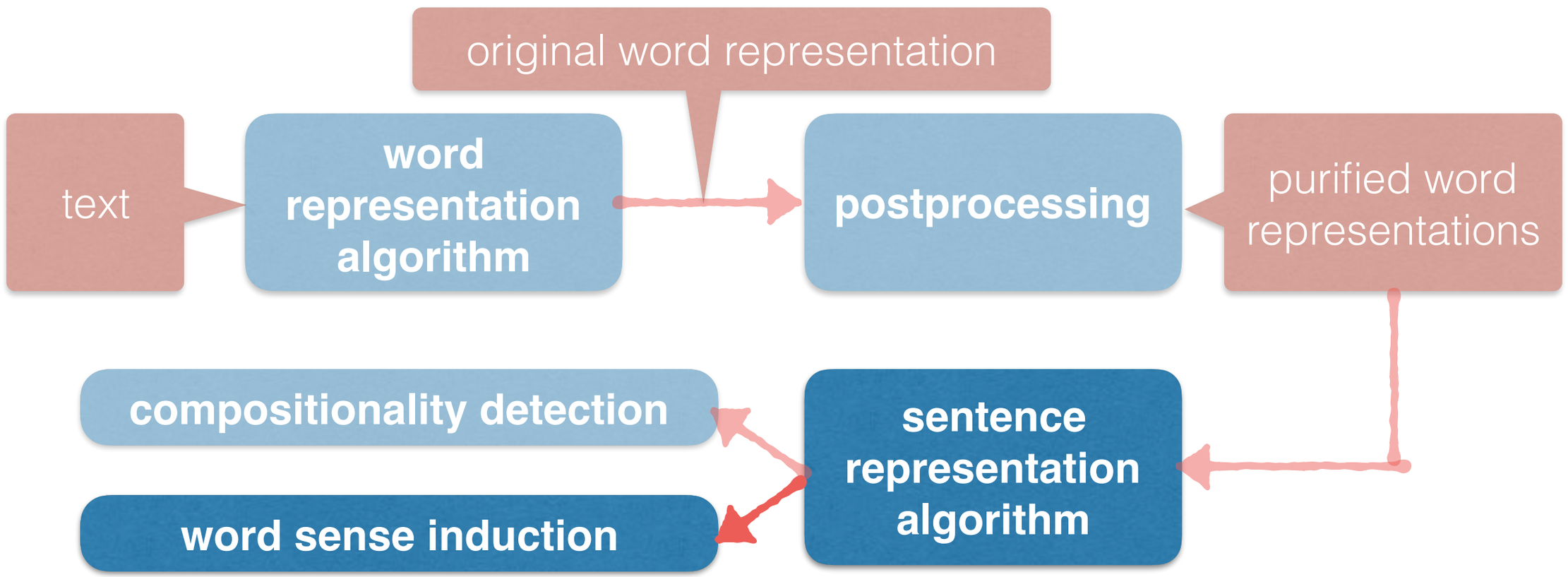
- Construct the subspace of principal components

- Compositionality with context is measured by the distance between target word/phrase and the subspace

Detection Algorithm

- Linguistic resource-independent
- Multilingual applicability: English, German and Chinese
- Context sensitivity
- Accurate detection in extensive experiments
  - compositionality, irony, metaphor

Word Sense Induction, Disambiguation and Representation



Ref: https://arxiv.org/abs/1610.07569

Polysemous Nature of Words

"crane"

Sense Representation

- supervised: aided by hand-crafted lexical resources (for 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.

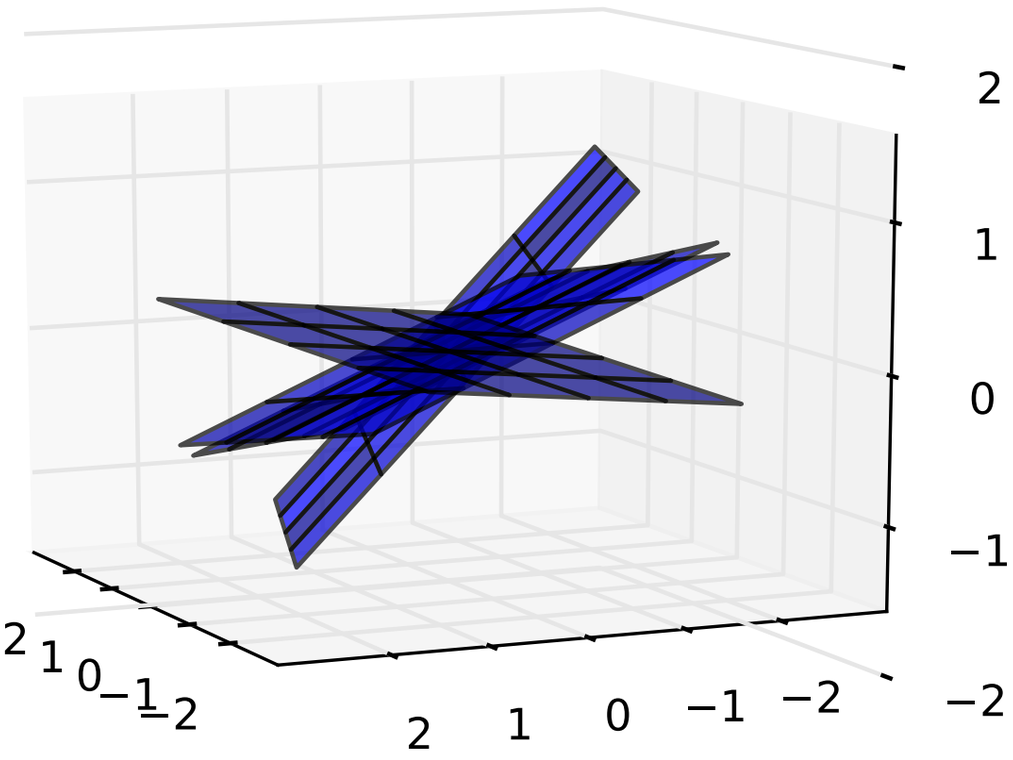
Context Representation

average

subspace

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

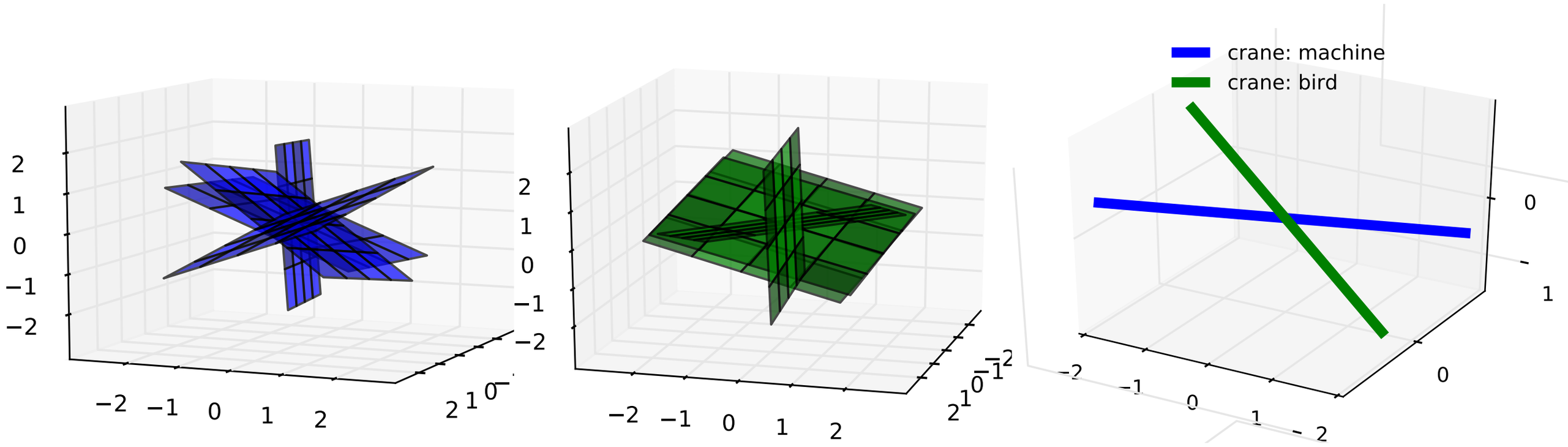
$$\hat{u}(w) = \arg\min_{\|u\|=1} \sum_{w \in c} d(u, S(c \setminus w))^2$$

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

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 , M contexts $c_1, ..., c_M$ containing w and a number K indicating the number of senses w has.
- 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

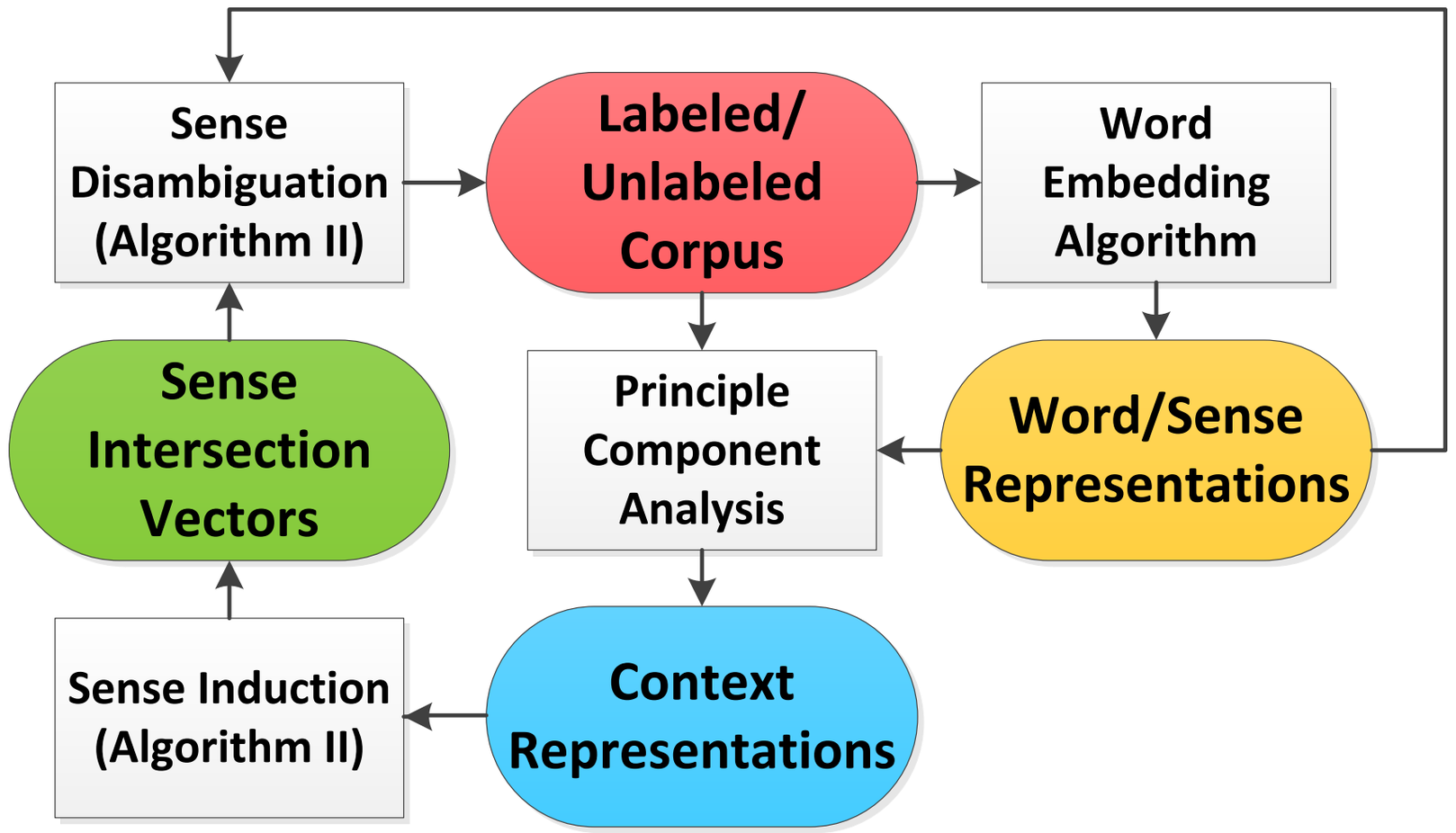
$$P(w, c, k) = \frac{\exp(-d(u_k(w), S(c \setminus w)))}{\sum_{k'} \exp(-d(u_{k'}(w), S(c \setminus w)))}$$

- Hard Decoding: output a deterministic classification

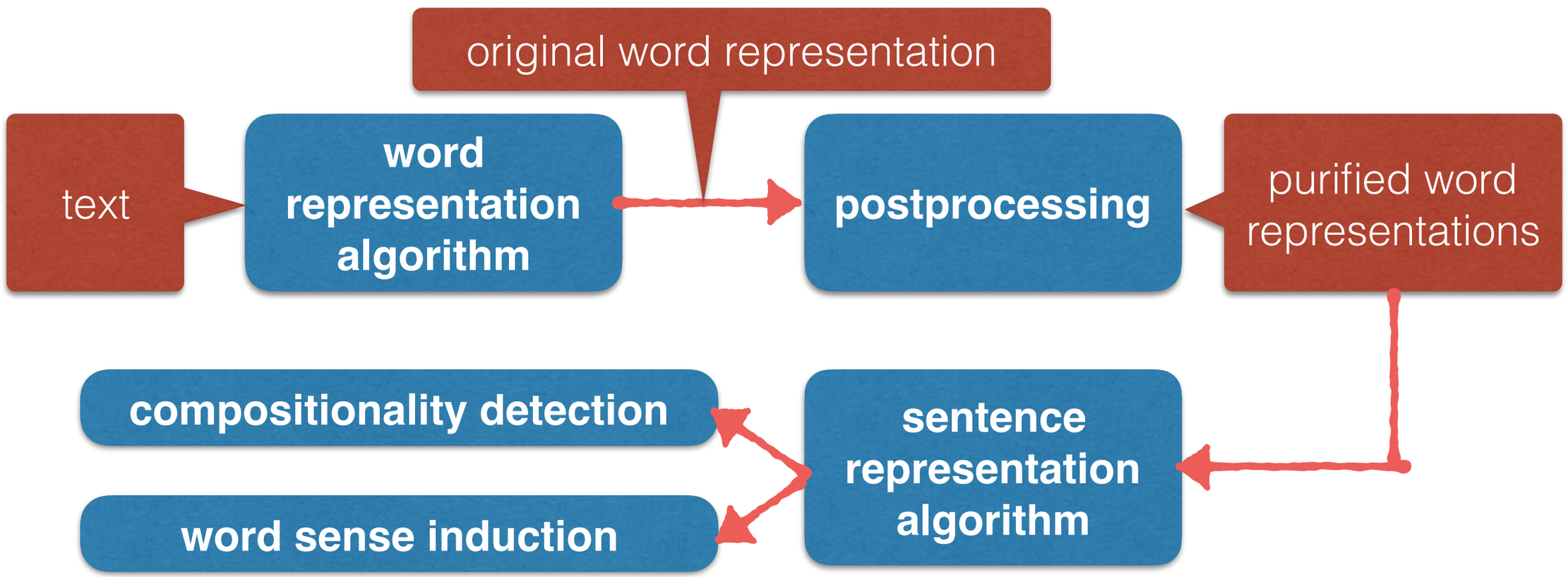
$$k^* = \arg\min_k d(u_k(w), S(c \setminus w))$$

Sense Representation

Similar Senses should have similar representations.



Geometry of Representations



Thank you!!