Model-Based Word Embeddings from Decompositions of Count Matrices

Karl Stratos

Joint work with Michael Collins and Daniel Hsu

Columbia University

Motivation: WORD2VEC as Matrix Decomposition

- WORD2VEC (Mikolov et al., 2013) trains word/context embeddings by maximizing some objective:

$$(v_w, v_c) = \arg\max_{u,v} J(u, v)$$

- Recently cast as a *low-rank decomposition* of *transformed co-occurrence counts* (Levy and Goldberg, 2014):

$$v_w^\top v_c = f(\text{count}(w, c))$$

- **Q. Are there other count transformations whose low-rank decompositions yield effective word embeddings?**

This Work

1. Count transformation under **canonical correlation analysis (CCA)** (Hotelling, 1936)
   - Model-based interpretation that permits a variance-stabilizing transformation

2. Unifies a number of existing spectral methods for inducing word embeddings

3. Empirically competitive with other popular methods such as WORD2VEC and GLOVE

Overview

Canonical Correlation Analysis

- Variational Characterization
- Count Transformation for Word Embeddings
- Model-Based Interpretation

Template for Spectral Word Embeddings

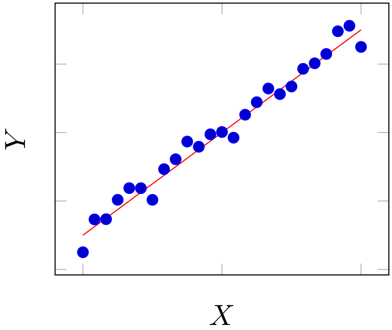
Experiments

Correlation Coefficient

- Correlation coefficient between random variables $X, Y \in \mathbb{R}$:

$$\text{Cor}(X,Y) := \frac{\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}{\sqrt{(\mathbf{E}[X^2] - \mathbf{E}[X]^2) \times (\mathbf{E}[Y^2] - \mathbf{E}[Y]^2)}}$$

Degree of linear relationship $[-1, 1]$



$$\text{Cor}(X,Y) \approx 1$$

Optimization Problem Underlying CCA

Input:

1. $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ // two "views" of an object
2. $m \leq \min(d, d')$ // number of projection vectors

Output: $(a_1, b_1) \ldots (a_m, b_m) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ such that

- (a_1, b_1) is the solution of

$$\arg\max_{a,b} \ \text{Cor}\left(a^\top X,\ b^\top Y\right) \qquad (1)$$

- For $i = 2 \ldots m$: (a_i, b_i) is the solution of (1) subject to:

$$\text{Cor}\left(a^\top X,\ a_j^\top X\right) = 0 \qquad \forall j < i$$

$$\text{Cor}\left(b^\top Y,\ b_j^\top Y\right) = 0 \qquad \forall j < i$$

Exact Solution via Singular Value Decomposition (SVD)

Theorem. (Hotelling, 1936) Define **correlation matrix** $\Omega \in \mathbb{R}^{d \times d'}$:

$$\Omega := \left(\mathbf{E}[XX^\top] - \mathbf{E}[X]\mathbf{E}[X]^\top\right)^{-1/2} \left(\mathbf{E}[XY^\top] - \mathbf{E}[X]\mathbf{E}[Y]^\top\right) \left(\mathbf{E}[YY^\top] - \mathbf{E}[Y]\mathbf{E}[Y]^\top\right)^{-1/2}$$

Let (u_i, v_i) be the left/right singular vectors of Ω corresponding to the i-th largest singular value. Then

$$a_i = \left(\mathbf{E}[XX^\top] - \mathbf{E}[X]\mathbf{E}[X]^\top\right)^{-1/2} u_i$$
$$b_i = \left(\mathbf{E}[YY^\top] - \mathbf{E}[Y]\mathbf{E}[Y]^\top\right)^{-1/2} v_i$$

New Representation under CCA

- Induce new m-dimensional representation $(\underline{X}, \underline{Y})$ of (X, Y):

$$\underline{X}_i = a_i^\top X$$
$$\underline{Y}_i = b_i^\top Y$$

  for $i = 1 \dots m$

- Idea: remove ambient dimensions by projecting to a subspace containing most correlation

Overview

Canonical Correlation Analysis

- Variational Characterization
- Count Transformation for Word Embeddings
- Model-Based Interpretation

Template for Spectral Word Embeddings

Experiments

Two Views of a Word: Word & Context

Extract samples of $(X, Y) := (\textbf{word}, \textbf{context})$ from a corpus:

... `Whatever` **our** **souls** **are** `made of`...

$$\downarrow$$

$$(\mathcal{I}_{\textbf{souls}}, \mathcal{I}_{\textbf{our}}) \quad (\mathcal{I}_{\textbf{souls}}, \mathcal{I}_{\textbf{are}})$$

where $\mathcal{I}_i$ is an indicator vector for i

Need to perform singular value decomposition (SVD) on

$$\hat{\Omega} = \left(\hat{\mathbf{E}}[XX^\top] - \hat{\mathbf{E}}[X]\hat{\mathbf{E}}[X]^\top\right)^{-1/2} \left(\hat{\mathbf{E}}[XY^\top] - \hat{\mathbf{E}}[X]\hat{\mathbf{E}}[Y]^\top\right) \left(\hat{\mathbf{E}}[YY^\top] - \hat{\mathbf{E}}[Y]\hat{\mathbf{E}}[Y]^\top\right)^{-1/2}$$

Simplified Correlation Matrix

When the number of samples is large, the means tend to zero:

$$\hat{\Omega} \approx \hat{\mathbf{E}}\left[XX^\top\right]^{-1/2} \hat{\mathbf{E}}\left[XY^\top\right] \hat{\mathbf{E}}\left[YY^\top\right]^{-1/2}$$

I.e., decompose the following transformed counts!

$$\hat{\Omega}_{w,c} = \frac{\mathsf{count}(w,c)}{\sqrt{\mathsf{count}(w) \times \mathsf{count}(c)}}$$

Previous Work Using CCA for Word Embeddings

- Dhillon et al. (2011, 2012) propose various modifications of CCA, but take the square root of counts,

$$\hat{\Omega}_{w,c} = \frac{\mathsf{count}(w,c)^{1/2}}{\sqrt{\mathsf{count}(w)^{1/2} \times \mathsf{count}(c)^{1/2}}}$$

- The square root was taken for empirical reasons.

- We now provide a **model-based interpretation** that naturally admits this extra transformation.

Overview

Canonical Correlation Analysis

- Variational Characterization
- Count Transformation for Word Embeddings
- Model-Based Interpretation

Template for Spectral Word Embeddings

Experiments

Definition of the "Brown Model" (Brown et al., 1992)

Parameters: same as HMMs

$$\pi(h) = \text{probability of state } h \text{ starting a sequence}$$
$$t(h'|h) = \text{probability of transitioning from state } h \text{ to state } h'$$
$$o(w|h) = \text{probability of word } w \text{ under state } h$$

Assumption: every word w has a *single* possible state h

- Define emission matrix O where $O_{w,h} = o(w|h)$
- Rows of O can be seen as *state-revealing* word embeddings

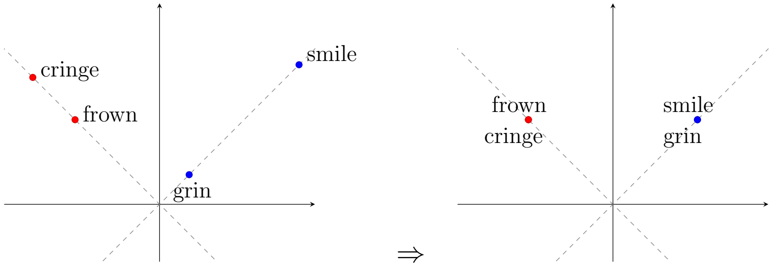
$$O_{\text{smile}} = [0.3 \quad 0.0] \qquad O_{\text{frown}} = [0.0 \quad 0.25]$$
$$O_{\text{grin}} = [0.7 \quad 0.0] \qquad O_{\text{cringe}} = [0.0 \quad 0.75]$$

Using the Scaled, Rotated Rows of O as Word Embeddings

Suppose we had $\overline{O} := \text{diag}(s_1)O^{\langle a \rangle}\text{diag}(s_2)Q^\top$ where

- s_1 and s_2 are any positive vectors
- $O^{\langle a \rangle}$ is an element-wise power of O with any $a \neq 0$
- Q is any orthogonal matrix

Normalized rows of $\overline{O}$ have the same representational power as normalized rows of O!



CCA for Estimating O up to Scaling and Rotation

Theorem. Pick any $a \neq 0$. Let $\hat{U}$ be the top m left singular vectors of $\hat{\Omega}^{\langle a \rangle}$ where

$$\hat{\Omega}^{\langle a \rangle}_{w,c} = \frac{\text{count}(w,c)^a}{\sqrt{\text{count}(w)^a \times \text{count}(c)^a}}$$

Then as the sample size grows:

$$\hat{U} \to O^{\langle a/2 \rangle} \text{diag}(s) Q^\top$$

for some $s > 0$ and orthogonal Q

Proof. Extension of Stratos et al. (2014)

Choosing the Value of a

So we can choose any $a \neq 0$, what should it be?

One answer: $a = 1/2$

Why?

- Assume word counts drawn from a multinomial distribution
- Equivalent to drawing from independent Poisson distributions (conditioned on the length of the corpus)
- Square-root is a **variance-stabilizing** transformation for Poisson random variables (Bartlett, 1936):

$$X \sim \text{Poisson}(np)$$
$$\text{Var}(X^{1/2}) \to \mathbf{1/4} \qquad \text{as } n \to \infty$$

Why does Variance Stabilization Help?

SVD minimizes unweighted squared-error loss:

$$\min_{u_w, v_c} \sum_{w,c} \left(\Omega_{w,c}^{\langle a \rangle} - u_w^\top v_c \right)^2$$

But minimizing *variance-weighted* squared-error loss is more statistically efficient (Aitken, 1936):

$$\min_{u_w, v_c} \sum_{w,c} \frac{1}{\mathsf{Var}\left(\Omega_{w,c}^{\langle a \rangle} \right)} \left(\Omega_{w,c}^{\langle a \rangle} - u_w^\top v_c \right)^2$$

Generally intractable (Srebro et al., 2003)

Using $a = 1/2$ makes $\mathsf{Var}\left(\Omega_{w,c}^{\langle a \rangle} \right)$ approximately **constant** and removes the need for explicit variance weighting!

Overview

Template

Input: count(w, c), dimension m, transform t, scaling s

- count$(w) := \sum_c$ count(w, c)
- count$(c) := \sum_w$ count(w, c)

Output: embedding $v(w) \in \mathbb{R}^m$ for each word w

---

1. Transform counts

2. Scale counts to construct matrix $\hat{\Omega}$

3. Do **rank**-m **SVD** on $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ and let $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Template: No Scaling (Pennington et al., 2014)

Input: count(w, c), dimension m, $t = \log$, $s = $ —

- count$(w) := \sum_c$ count(w, c)
- count$(c) := \sum_w$ count(w, c)

Output: embedding $v(w) \in \mathbb{R}^m$ for each word w

---

1. Transform counts

   count$(w, c) \leftarrow \log(1 + \text{count}(w, c))$

2. Scale counts to construct matrix $\hat{\Omega}$

$$\hat{\Omega}_{w,c} = \text{count}(w, c)$$

3. Do **rank**-m **SVD** on $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ and let $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Template: PPMI (Levy and Goldberg, 2014)

Input: count(w, c), dimension m, $t =$ —, $s =$ ppmi

- count$(w) := \sum_c$ count(w, c)
- count$(c) := \sum_w$ count(w, c)

Output: embedding $v(w) \in \mathbb{R}^m$ for each word w

---

1. Transform counts

$$\text{count}(w, c) \leftarrow \text{count}(w, c) \qquad \text{count}(w) \leftarrow \text{count}(w)$$
$$\text{count}(c) \leftarrow \text{count}(c)$$

2. Scale counts to construct matrix $\hat{\Omega}$

$$\hat{\Omega}_{w,c} = \max\left(0, \log \frac{\text{count}(w, c) \times \sum_{w,c} \text{count}(w, c)}{\text{count}(w) \times \text{count}(c)}\right)$$

3. Do **rank-m SVD** on $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ and let $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Template: CCA (Stratos et al., 2014)

Input: count(w, c), dimension m, $t = $ —, $s = $ cca

- count$(w) := \sum_c$ count(w, c)
- count$(c) := \sum_w$ count(w, c)

Output: embedding $v(w) \in \mathbb{R}^m$ for each word w

---

1. Transform counts

$$\text{count}(w,c) \leftarrow \text{count}(w,c) \qquad \text{count}(w) \leftarrow \text{count}(w)$$
$$\text{count}(c) \leftarrow \text{count}(c)$$

2. Scale counts to construct matrix $\hat{\Omega}$

$$\hat{\Omega}_{w,c} = \frac{\text{count}(w,c)}{\sqrt{\text{count}(w) \times \text{count}(c)}}$$

3. Do **rank-m SVD** on $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ and let $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Template: CCA with Square-Root (this work)

Input: count(w, c), dimension m, $t =$ sqrt, $s =$ cca

- count$(w) := \sum_c$ count(w, c)
- count$(c) := \sum_w$ count(w, c)

Output: embedding $v(w) \in \mathbb{R}^m$ for each word w

---

1. Transform counts

$$\text{count}(w,c) \leftarrow \sqrt{\text{count}(w,c)} \qquad \text{count}(w) \leftarrow \sqrt{\text{count}(w)}$$

$$\text{count}(c) \leftarrow \sqrt{\text{count}(c)}$$

2. Scale counts to construct matrix $\hat{\Omega}$

$$\hat{\Omega}_{w,c} = \frac{\text{count}(w,c)}{\sqrt{\text{count}(w) \times \text{count}(c)}}$$

3. Do **rank-m SVD** on $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ and let $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Overview

Canonical Correlation Analysis

- Variational Characterization
- Count Transformation for Word Embeddings
- Model-Based Interpretation

Template for Spectral Word Embeddings

Experiments

Setting

Corpus: pre-processed English Wikipedia (1.4 billion words)

Evalution

- Word similarity: correlation with human judgment on ranking similar words, averaged across 3 datasets (AVG-SIM)
- Word analogy: answering analogy questions of form

  $$\texttt{Beijing} : \texttt{China} \sim \texttt{Kampala} : \ ?$$

  Syntactic (SYN), syntactic+semantic (MIXED)
- Semi-supervised learning: improving performance of supervised learner

Comparison with GLOVE (Pennington et al., 2014) and CBOW, SGNS (implemented in WORD2VEC) (Mikolov et al., 2013)

- Default hyperparameter configuration

Effect of Power Transformation in CCA

Different values of a in

$$\hat{\Omega}^{\langle a \rangle}_{w,c} = \frac{\text{count}(w,c)^a}{\sqrt{\text{count}(w)^a \times \text{count}(c)^a}}$$

1000 dimensions

a	AVG-SIM	SYN	MIXED
1	0.572	39.68	57.64
2/3	0.650	60.52	74.00
1/2	**0.690**	**65.14**	**77.70**

Word Similarity and Analogy

- LOG: log transform, no scaling
- PPMI: no transform, PPMI scaling
- CCA: square-root transform, CCA scaling

500 dimensions

Method		AVG-SIM	SYN	MIXED
Spectral	LOG	0.652	59.52	67.27
	PPMI	0.628	43.81	58.38
	CCA	**0.655**	68.38	74.17
Others	GLOVE	0.576	68.30	78.08
	CBOW	0.597	75.79	73.60
	SGNS	0.642	**81.08**	**78.73**

Semi-Supervised Learning

- Features in named-entity recognition (CoNLL 2003)
- RCV1 corpus (205 million words)

30 dimensions

Features	Dev	Test
—	90.04	84.40
BROWN	92.49	88.75
LOG	92.27	88.87
PPMI	92.25	89.27
CCA	**92.88**	**89.28**
GLOVE	91.49	87.16
CBOW	92.44	88.34
SGNS	92.63	88.78

(BROWN: 1000 Brown clusters (Brown et al., 1992))

Summary

We developed a new statistical understanding of word embeddings based on transformed counts

- CCA transformations: recovery of Brown model

Unified many spectral word embedding methods

Future work includes:

- Applying square-root in other SVD applications
- Relaxing Brown model assumption