Model-Based Word Embeddings from Decompositions of Count Matrices

Karl Stratos, Michael Collins, Daniel Hsu

担当：村岡雅康(M2)
乾・岡崎研究室
東北大学大学院 情報科学研究科

モチベーション

- [Levy&Goldberg14]: skip-gram[Mikolov+13]は **postive PMI(PPMI)に変換**された共起行列の分解と等価

- **Question: 他の優れた変換って何かないの？**

本研究の概要

- CCA(Canonical Correlation Analysis)[Hotelling, 1936]で次元圧縮するときに行う変換を拡張した手法を提案

- similarity, analogy, NERのタスクでskip-gramとcomparableな結果

Outline

- **CCAによる次元圧縮**
- Brownモデルによる拡張
- テンプレートの導入
- 実験

CCA(Canonical Correlation Analysis)とは

- 2つのベクトル(X, Y)を次元圧縮する方法
  - 今回の入力は単語&文脈ベクトル
- 特徴：2つのベクトル(X, Y)間の相関ができるだけ高くなるような空間に射影

↓

- 冗長な次元が削減された空間
- 射影するための行列(A, B)を学習する必要あり

SVDによる解法

- SVD(Singular Value Decompositon)を使った解法 [Hottelling,1936]で厳密解が求まる

相関行列： $\Omega \in \mathbb{R}^{d \times d'}$

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

SVDによる分解

$$\Omega \approx U\Sigma V^\top$$

$\xrightarrow{\text{SVD}}$

$$A = \left(\mathbf{E}[XX^\top] - \mathbf{E}[X]\mathbf{E}[X]^\top\right)^{-1/2} U$$

$$B = \left(\mathbf{E}[YY^\top] - \mathbf{E}[Y]\mathbf{E}[Y]^\top\right)^{-1/2} V$$

XとYの表現を工夫することで簡略化できる

単語/文脈のone-hot表現

- … Whatever **our souls are** made of …

$\downarrow$

- $(x^{(i)}, y^{(i)}) = (\mathcal{I}_{\textbf{souls}}, \mathcal{I}_{\textbf{our}}), \quad (\mathcal{I}_{\textbf{souls}}, \mathcal{I}_{\textbf{are}})$
- サンプル数→大のとき，平均→0

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

対角行列

対角行列

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

既存研究

- [Dhillon+11;12]

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

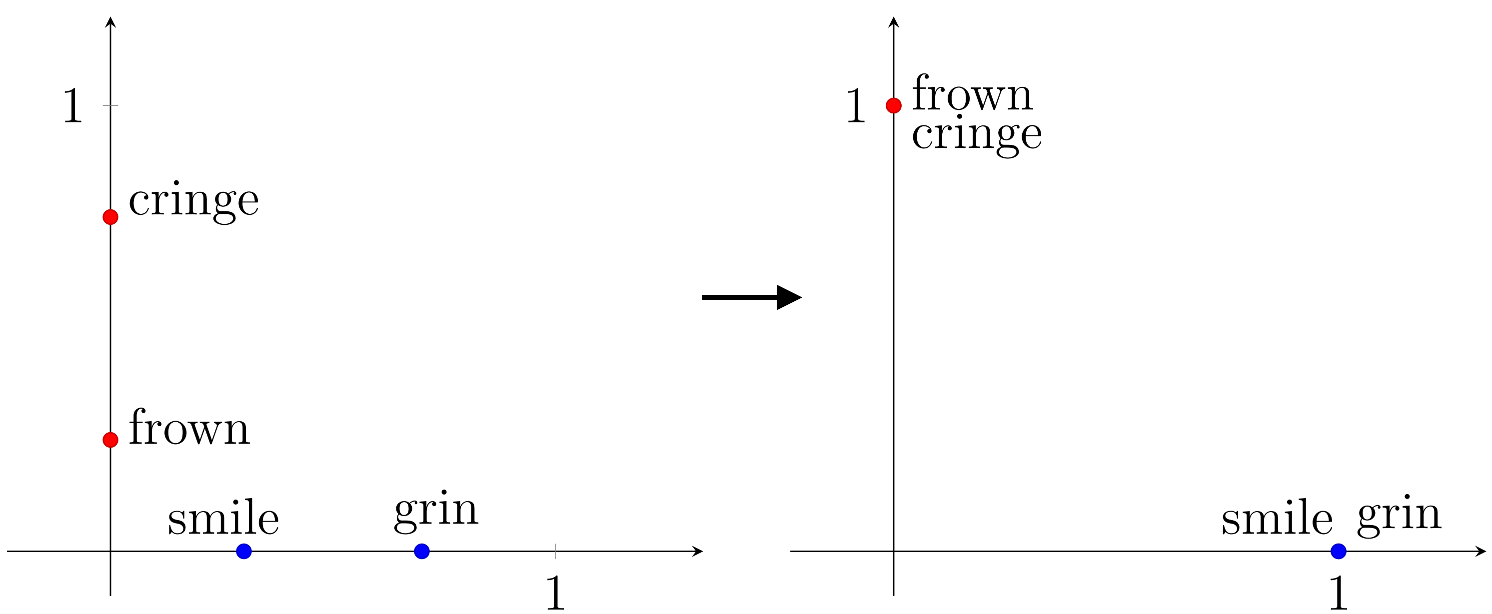
- 1/2は経験的(empirical)な理由
- 本研究ではブラウンモデル[Brown+1992]を使って理論的な裏付けを行う

Outline

- CCAによる次元圧縮
- **Brownモデルによる拡張**
- テンプレートの導入
- 実験

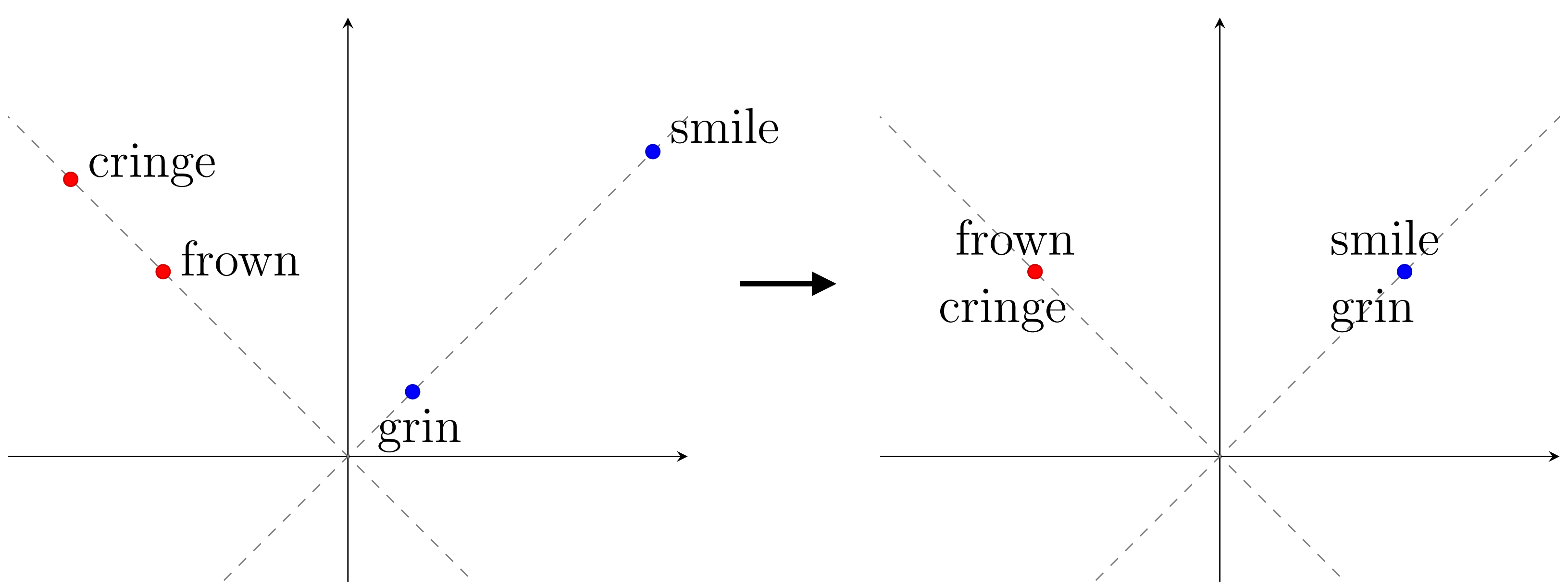
ブラウンモデル[Brown+1992]

- 隠れ状態に制約のあるHMM
- 制約(ブラウン仮定)：各単語には高々1個の隠れ状態
- ⇒ 出力行列Oの各行は1要素以外全てゼロ
  ただし, $O_{w,h} = o(w|h)$
- 正規化することで隠れ状態数だけクラスタができる



回転&スケールしても表現力は同じ

- $\overline{O} := \mathrm{diag}(s_1) O^{\langle a \rangle} \mathrm{diag}(s_2) Q^\top$
  - $s_1, s_2 > 0$ を満たすベクトル
  - $O^{\langle a \rangle}$：要素ごとに指数乗した出力行列
  - Q：任意の直交(変換)行列



定理

- $a \neq 0$. $\hat{U}$: $\hat{\Omega}^{\langle a \rangle}_{w,c}$ の左特異値ベクトル

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

- サンプル数が大きいとき

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

- (証明) Appendix A および[Stratos+14]を参照

- 主張：(回転&スケールされた) O を用いれば任意の a を選ぶことができる！

$a = 1/2$ が最適

- 根拠：
  - 単語の出現は多項分布に従うと仮定
  - これはそれぞれ独立なポアソン分布と等価
  - ポアソン分布の2乗根は分散安定な変換[Bartlett,1936]

$$X \sim \text{Poisson}(np)$$

$$\text{Var}(X^{1/2}) \to \mathbf{1/4} \quad (n \to \infty)$$

分散安定のうれしさ

- SVDの目的関数：重みなし二乗誤差
  - 分散不均一データに関してはsuboptimal

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

- 分散で重み付けられた二乗誤差[Aitken,1936]😀

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

- これだと一般的にintractable[Srebro+03]😕
- でも今は定数で近似できる！😀

Outline

- CCAによる次元圧縮
- Brownモデルによる拡張
- **テンプレートの導入**
- 実験

SVDモデル(テンプレート)

- 入力：共起頻度count(w, c), 次元m, 変換t, スケールs
  - $\text{count}(w) := \sum_c \text{count}(w, c)$
  - $\text{count}(c) := \sum_w \text{count}(w, c)$
- 出力：m次元の単語ベクトルv(w)

1.頻度の変換

2.スケール

3.SVD: $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top \longrightarrow v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

SVDモデル(提案手法)

- 入力：共起頻度count(w, c), 次元m, 変換**sqrt**, スケール**cca**
  - $\text{count}(w) := \sum_c \text{count}(w, c)$
  - $\text{count}(c) := \sum_w \text{count}(w, c)$
- 出力：m次元の単語ベクトルv(w)

1.頻度の変換 $\text{count}(w, c) \leftarrow \sqrt{\text{count}(w, c)}$ $\text{count}(w) \leftarrow \sqrt{\text{count}(w)}$

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

2.スケール
$$\hat{\Omega}_{w,c} = \frac{\text{count}(w, c)}{\sqrt{\text{count}(w) \times \text{count}(c)}}$$

3.SVD: $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top$ $\longrightarrow$ $v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

SVDモデル[Levy&Goldberg14]

- 入力：共起頻度count(w, c), 次元m, 変換**なし**, スケール**ppmi**
  - $\mathsf{count}(w) := \sum_c \mathsf{count}(w, c)$
  - $\mathsf{count}(c) := \sum_w \mathsf{count}(w, c)$
- 出力：m次元の単語ベクトルv(w)

1.頻度の変換 $\mathsf{count}(w, c) \leftarrow \mathsf{count}(w, c)$ $\mathsf{count}(w) \leftarrow \mathsf{count}(w)$

$\mathsf{count}(c) \leftarrow \mathsf{count}(c)$

2.スケール

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

3.SVD: $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top \longrightarrow v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

SVDモデル[Peninngton+14]

- 入力：共起頻度count(w, c), 次元m, 変換**log**, スケール**なし**
  - $\mathsf{count}(w) := \sum_c \mathsf{count}(w, c)$
  - $\mathsf{count}(c) := \sum_w \mathsf{count}(w, c)$
- 出力：m次元の単語ベクトルv(w)

1.頻度の変換

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

2.スケール

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

3.SVD: $\hat{\Omega} \approx \hat{U}\hat{\Sigma}\hat{V}^\top \longrightarrow v(w) = \hat{U}_w / \left|\left|\hat{U}_w\right|\right|$

Outline

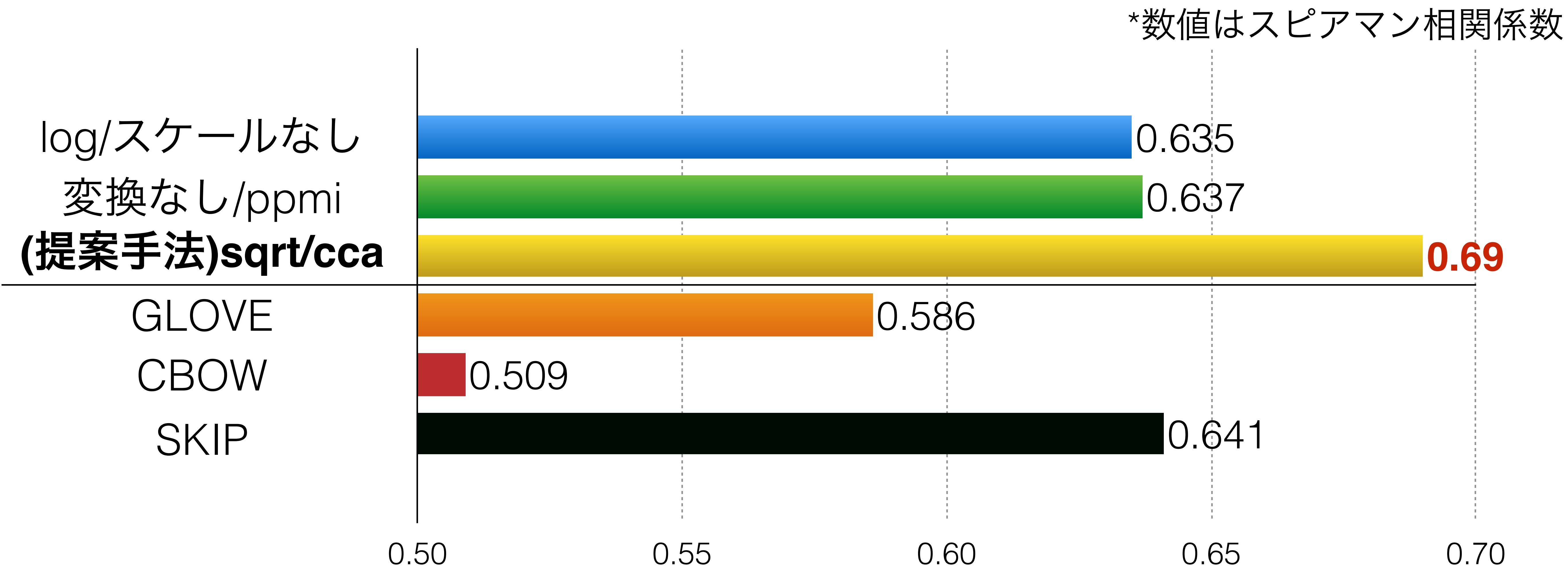
- CCAによる次元圧縮
- Brownモデルによる拡張
- テンプレートの導入
- **実験**

実験

- コーパス：English Wikipedia (1.4 billion words)
- 評価タスク
  - **Word similarity**: 人手のスコアとのスピアマン相関
    - (money, cash) → 9.08, (king, cabbage) → 0.23
  - **Word analogy**: Beijing : China ~ Tokyo : ?
  - **NER (CoNLL 2003)**: embeddingを素性として使用
- 外部モデル
  - GLOVE [Pennington+14]
  - CBOW, SKIP [Mikolov+13]
    - ハイパーパラメータはデフォルト

結果：Word similarity

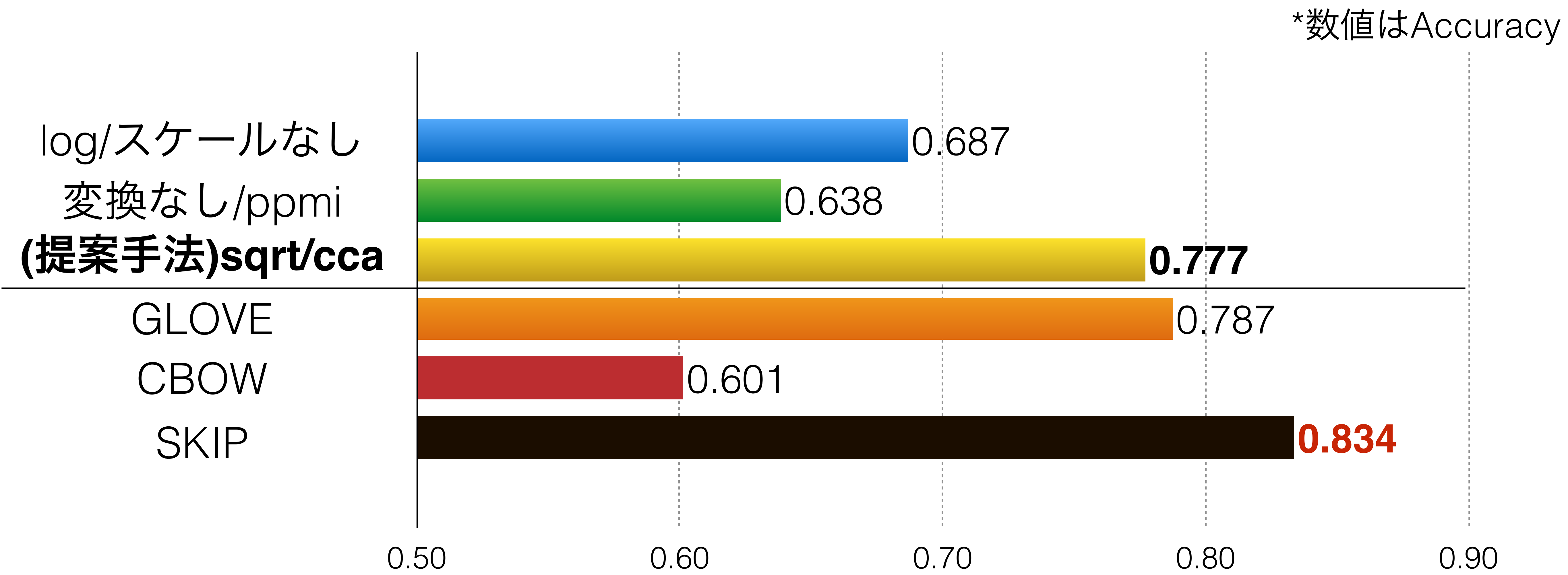
- 1000次元



提案手法(sqrt/cca)が圧勝

結果：Word analogy

- 1000次元

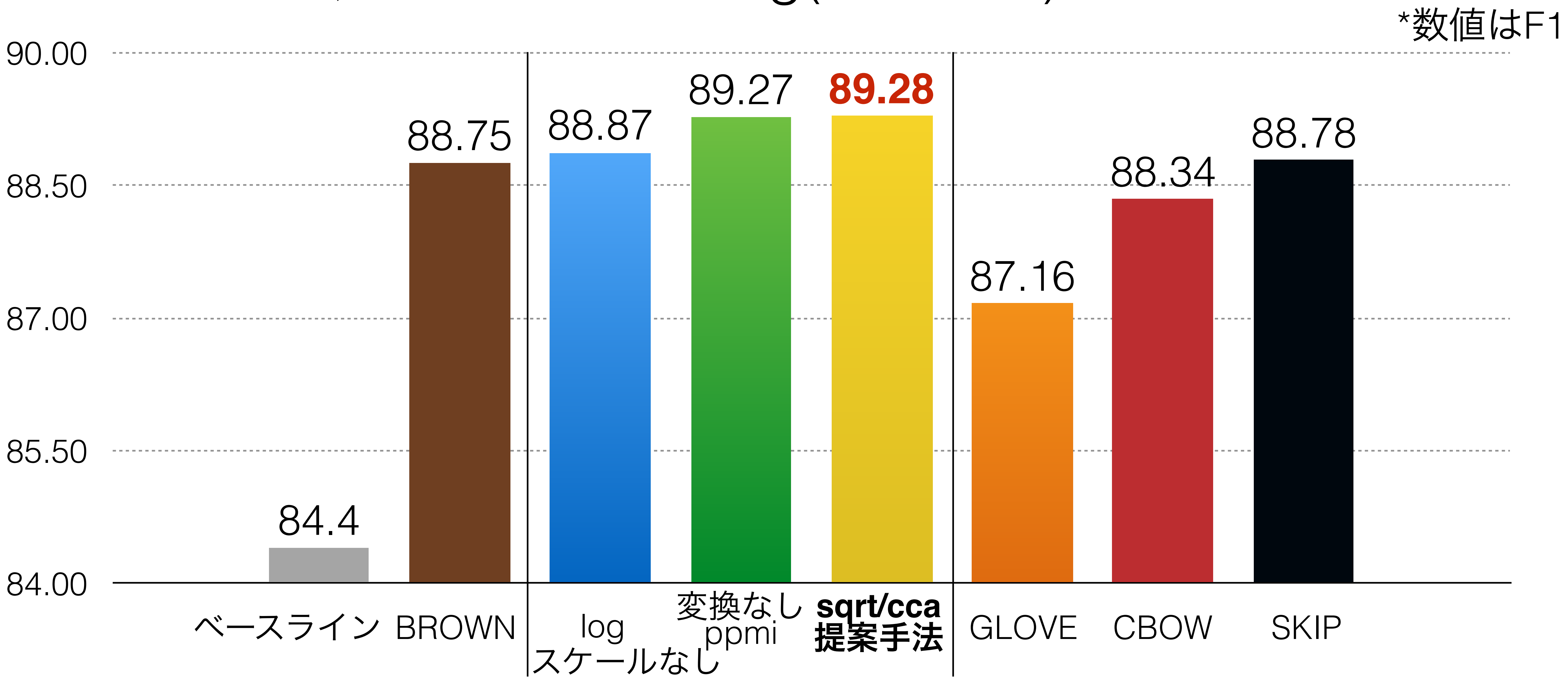


SVDモデル(上)内では提案手法(sqrt/cca)がベスト

全体ではskip-gramが最も良い結果

結果：NER (CoNLL 2003)

- 30次元, Brown clustering(BROWN)は1000クラスタ



SVDモデル(中央)がBROWN, SKIPを上回る

まとめ

- 共起行列の成分をCCAを用いて変換する手法を提案
  - その拡張としてブラウンモデルを取り入れた

- SVDモデルのテンプレートを導入した

- similarity, analogy, NERのタスクでskip-gramとcomparableな結果

Appendixes

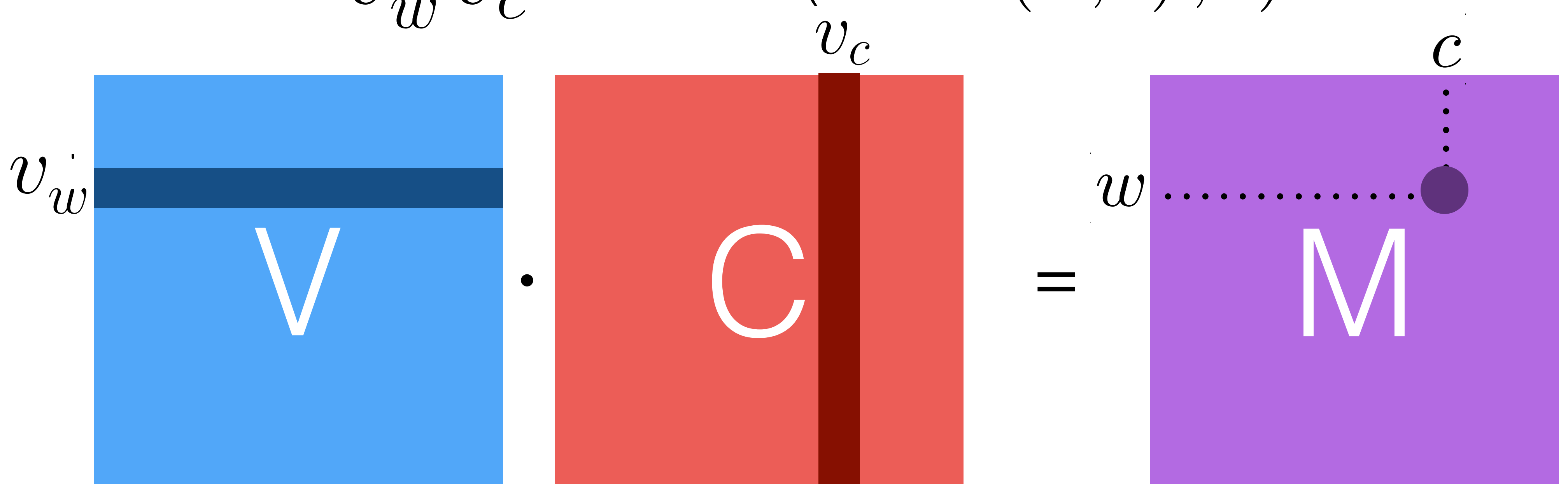
復習：skip-gram

- ある目的関数を最大化するように単語/文脈ベクトルを学習:

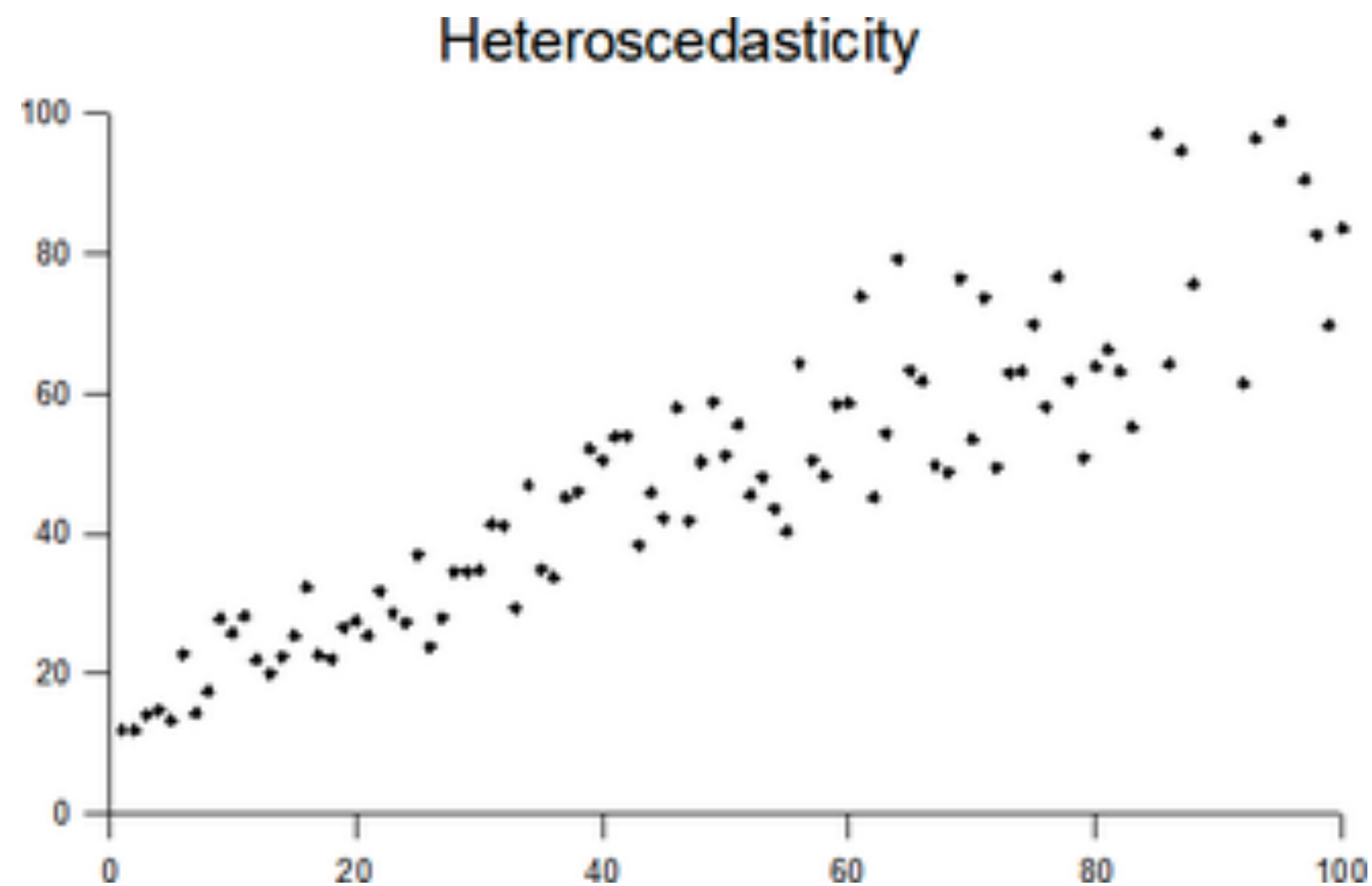
$$(v_w, v_c) = \arg\max_{u,v} J(u, v)$$

- その内積は共起頻度のPPMIである[Levy&Goldberg14]

$$v_w^\top v_c = \max\left(PMI\left(w, c\right), 0\right)$$



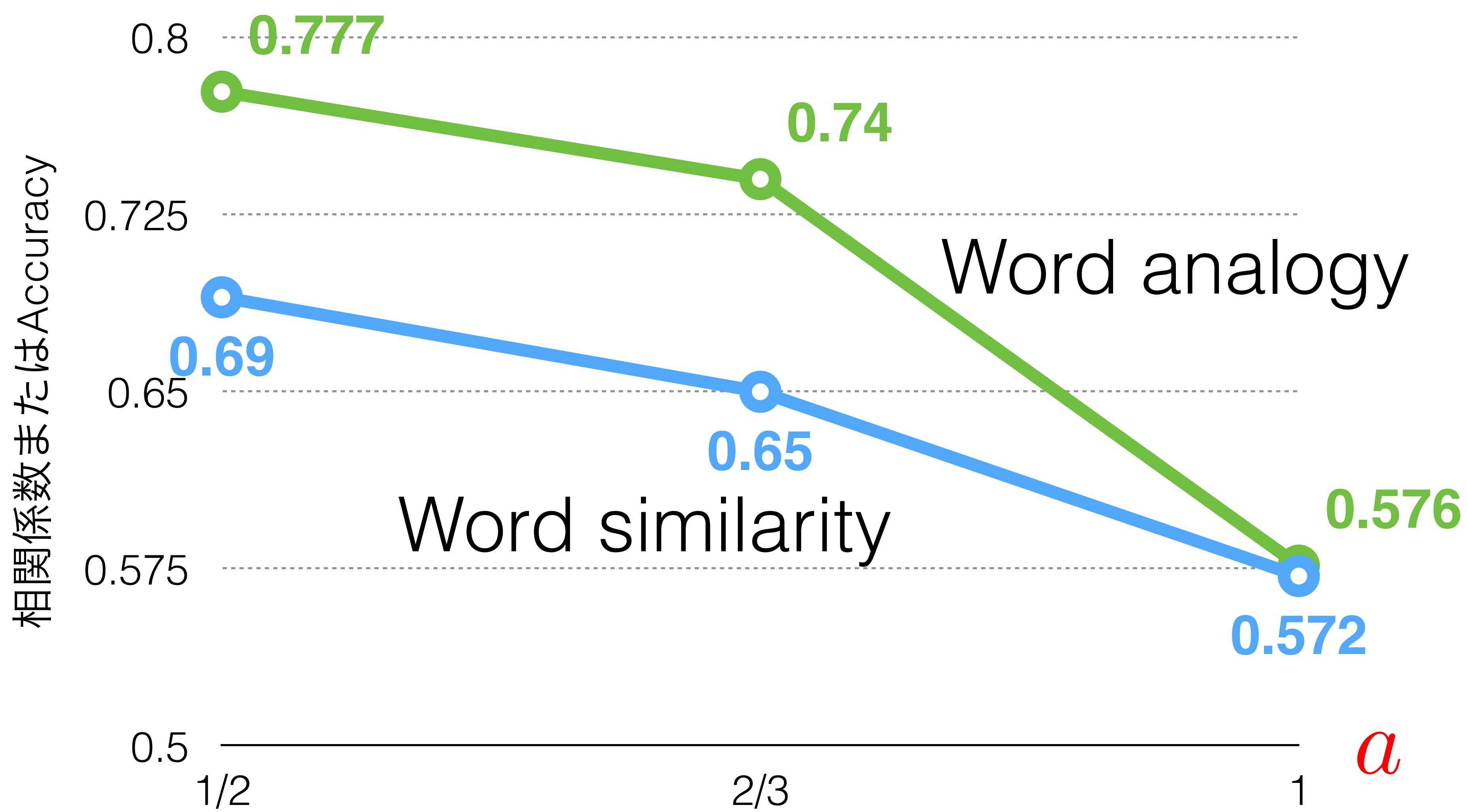
分散不均一性



予備実験(a の影響)

- 1000次元

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



Template of SVD model

SPECTRAL-TEMPLATE
Input: word-context co-occurrence counts $\#(w,c)$, dimension m, transformation method t, scaling method s, context smoothing exponent $\alpha \leq 1$, singular value exponent $\beta \leq 1$
Output: vector $v(w) \in \mathbb{R}^m$ for each word $w \in [n]$
Definitions: $\#(w) := \sum_c \#(w,c)$, $\#(c) := \sum_w \#(w,c)$, $N(\alpha) := \sum_c \#(c)^\alpha$

1. Transform all $\#(w,c)$, $\#(w)$, and $\#(c)$:

$$\#(\cdot) \leftarrow \begin{cases} \#(\cdot) & \text{if } t = \text{—} \\ \log(1+\#(\cdot)) & \text{if } t = \log \\ \#(\cdot)^{2/3} & \text{if } t = \text{two-thirds} \\ \sqrt{\#(\cdot)} & \text{if } t = \text{sqrt} \end{cases}$$

2. Scale statistics to construct a matrix $\Omega \in \mathbb{R}^{n \times n}$:

$$\Omega_{w,c} \leftarrow \begin{cases} \#(w,c) & \text{if } s = \text{—} \\ \frac{\#(w,c)}{\#(w)} & \text{if } s = \text{reg} \\ \max\left(\log \frac{\#(w,c)N(\alpha)}{\#(w)\#(c)^\alpha}, 0\right) & \text{if } s = \text{ppmi} \\ \frac{\#(w,c)}{\sqrt{\#(w)\#(c)^\alpha}}\sqrt{\frac{N(\alpha)}{N(1)}} & \text{if } s = \text{cca} \end{cases}$$

3. Perform rank-m SVD on $\Omega \approx U\Sigma V^\top$ where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_m)$ is a diagonal matrix of ordered singular values $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$.

4. Define $v(w) \in \mathbb{R}^m$ to be the w-th row of $U\Sigma^\beta$ normalized to have unit 2-norm.

Performance of SVD model

Configuration		500 dimensions			1000 dimensions		
Transform (t)	Scale (s)	AVG-SIM	SYN	MIXED	AVG-SIM	SYN	MIXED
—	—	0.514	31.58	28.39	0.522	29.84	32.15
sqrt	—	0.656	60.77	65.84	0.646	57.46	64.97
log	—	0.669	59.28	66.86	0.672	55.66	68.62
—	reg	0.530	29.61	36.90	0.562	32.78	37.65
sqrt	reg	0.625	63.97	67.30	0.638	**65.98**	70.04
—	ppmi	0.638	41.62	58.80	0.665	47.11	65.34
sqrt	cca	**0.678**	**66.40**	**74.73**	**0.690**	65.14	**77.70**

- Word similarity: 13 anotators
  - ex) (money, cash, 9.08), (king, cabbage, 0.23)

- Word analogy: [Levy&Goldberg'14]
  - a : b ~ c : x
  - $\operatorname{argmax}_{x \in V \setminus \{a,b,c\}} \cos(x, c) * \cos(x, b) / (\cos(x, a) + \varepsilon)$