# *A Latent Variable Model Approach to PMI-based Word Embeddings*

Sanjeev Arora+ (TACL 2016)

統計数理研究所

持橋大地

daichi@ism.ac.jp

最先端 NLP8

2016-9-11 (Sun)

Modified for ISM, 2016-9-30 (Fri)

# 概要

- word2vec, GloVe, …は PMI の行列を低ランク近似していることが知られている
- しかし, なぜ PMI を考えることが意味があるのかは不問
- 本論文では, ランダムウォークする文脈ベクトル $c_t$ からの近さに従って単語が生まれたと仮定することで,
  - 単語ベクトルの内積が相互情報量に対応すること,
  - “意味” の方向がベクトルの差と内積で書けること
    - 特に, 低次元表現により, それがノイズに埋もれないこと

  を示している.

# 注意

- 以下で誤差項 $\epsilon$ を複数含む式の等式は厳密ではありませんので,
  厳密な導出は完全版の
  http://arxiv.org/abs/1502.03520
  “RAND-WALK: A Latent Variable Model Approach to Word Embeddings” Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski
  をお読み下さい.

# はじめに

単語ベクトル $\mathbf{v}, \mathbf{w}$ について

$$\langle \mathbf{v}, \mathbf{w} \rangle \approx \mathrm{PMI}(\mathbf{v}, \mathbf{w}) \quad (1)$$

- Levy&Goldberg (2014b) は SGNS の解が (1) 式に似た,

$$\langle \mathbf{v}, \mathbf{w} \rangle \approx \mathrm{PMI}(\mathbf{v}, \mathbf{w}) - \beta \quad (2)$$

  を満たすことを示した
  - ただし高次元での話であり, 低次元表現が高精度であることの説明はない
- Hashimoto+(2016) は $p(v|w) \propto h(|\mathbf{v} - \mathbf{w}|^2)$ と仮定すると, (1) が導かれることを示した
  - ただし, 文脈ベクトルのような概念はなく, 単語だけで考えているので, “意味” の方向などは説明できない
  - 文脈ベクトルを考えることは, 意味が複数ある場合への拡張でも有効 (Arora+, arXiv 2016)

# 基本モデル

- 時刻 $t$ において, ランダムウォークする文脈ベクトル $c_t$ があり, それに従って単語 $w_t$ が下の式に従って発生すると仮定

$$p(w_t|c_t) = \frac{\exp(\mathbf{w}_t^T c_t)}{Z} \quad ; \; Z = \sum_w \exp(\mathbf{w}^T c_t) \tag{3}$$

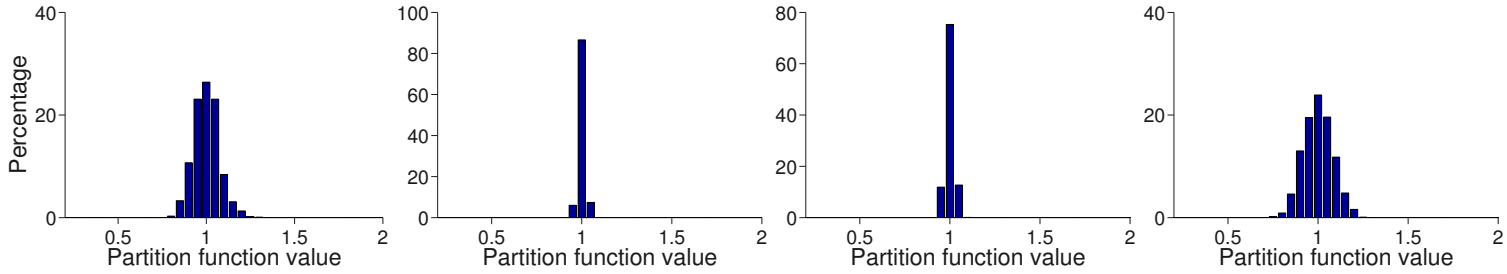
  - これは, Mnih&Hinton (2007) の Log-bilinear モデルを動的にしたもの
  - Belanger&Kakade (2015) はカルマンフィルタで単語が生成されるモデルを考えたが, 線形システムであり, 上のようなロジスティック回帰ではない
- 単語には確率が非常に大きいものもあるので, それは単語ベクトル $\mathbf{w}$ のノルムに反映される
  - $\mathbf{w} = s \cdot \hat{\mathbf{w}}$ を仮定, $\hat{\mathbf{w}}$ は単位球面 $\mathcal{C}$ 上のベクトル, $s$ は確率変数

# 定理 (0)

- 以下の計算で, 正規化定数 $Z(c) = \sum_w \exp(\mathbf{w}^T c)$ は $c$ にかかわらずほぼ 1 (!)
  - Self-normalizing log-linear model (Andreas & Klein 2014)

$$p_{c \sim \mathcal{C}} \left[ (1-\epsilon)Z \leq Z(c) \leq (1+\epsilon)Z \right] \geq 1 - \delta \tag{4}$$

ここで $\epsilon = \tilde{O}(1/\sqrt{n})$, $\delta = \exp(-\Omega(\log^2 n))$ ($n$: 単語種数)




(a) **SN** (b) GloVe (c) CBOW (d) skip-gram

Figure 1: The partition function $Z_c$. The figure shows the histogram of $Z_c$ for 1000 random vectors $c$ of appropriate norm, as defined in the text. The $x$-axis is normalized by the mean of the values. The values $Z_c$ for different $c$ concentrate around the mean, mostly in $[0.9, 1.1]$. This concentration phenomenon is predicted by our analysis.

# 定理 (1)

- 先の生成モデルと窓幅 2 の下で,

$$\log p(v,w) = \frac{|\mathbf{v}+\mathbf{w}|^2}{2d} - 2\log Z \pm \epsilon \tag{5}$$

$$\log\, p(w) = \frac{|\mathbf{w}|^2}{2d} - \log Z \pm \epsilon \tag{6}$$

これから,

$$\mathrm{PMI}\,(v,w) = \log \frac{p(v,w)}{p(v)p(w)} \tag{7}$$

$$= \frac{\mathbf{v}^T\mathbf{w}}{d} \pm O(\epsilon)\,. \tag{8}$$

# 補題 (2)

- 窓幅が $q$ 単語の場合は, 単語ペアの数がその中で $\binom{q}{2}$ 通りあるので,

$$\log p_q(v,w) = \frac{|\mathbf{v}+\mathbf{w}|^2}{2d} - 2\log Z + \gamma \pm \epsilon \tag{9}$$

$$\mathrm{PMI}_q(v,w) = \log \frac{p(v,w)}{p(v)p(w)} \tag{10}$$

$$= \frac{\mathbf{v}^T\mathbf{w}}{d} + \gamma \pm O(\epsilon) \tag{11}$$

ただし, $\gamma = \log \frac{q(q-1)}{2}$.

  - Levy&Goldberg(2014) の

$$\langle \mathbf{v}, \mathbf{w} \rangle \approx \mathrm{PMI}(\mathbf{v}, \mathbf{w}) - \beta \tag{12}$$

    の $\beta$ を説明

# 証明の概略

- 文脈 $c \sim \mathcal{C}$ と続く文脈 $c' \sim p(c'|c)$ について,

$$p(v, w) = E_{c,c'}[p(v, w|c, c')] \tag{13}$$

$$= E_{c,c'}[p(v|c)p(w|c')] \tag{14}$$

$$= E_{c,c'}\left[\frac{\exp(\mathbf{v}^T c)}{Z_c}\frac{\exp(\mathbf{v}^T c')}{Z_{c'}}\right] \tag{15}$$

  を示せばよい.

- ここで最初の定理から, $Z_c$ と $Z_{c'}$ が高い確率で $(1 \pm \epsilon)Z$ であることを用いると, それが成り立つ場合 (T1) と成り立たない場合 (T2) に分けて上の確率を評価できる
  - T2 の場合は, 確率は以下で negligible

$$|T_2| = \exp(-\Omega(\log^{1.8} n)) \tag{16}$$

# 証明の概略 (2)

- T1 の場合は,

$$T_1 = \frac{1 \pm O(\epsilon)}{Z^2} E_c \left[ \exp(\mathbf{v}^T c) E_{c'|c}[\exp(\mathbf{w}^T c')] \right] \tag{17}$$

$$= \frac{1 \pm O(\epsilon)}{Z^2} E_c \left[ \exp(\mathbf{v}^T c)\, A(c) \right] \tag{18}$$

$$= \frac{1 \pm O(\epsilon)}{Z^2} E_c \left[ \exp(\mathbf{v}^T c)\, (1 \pm O(\epsilon_z)) \exp(\langle \mathbf{w}, c \rangle) \right] \tag{19}$$

$$\simeq \frac{1 \pm O(\epsilon)}{Z^2} E_c [\exp(\langle \mathbf{v} + \mathbf{w}, c \rangle)] \tag{20}$$

- さらに

$$E_{c \sim \mathcal{C}}[\exp(\langle \mathbf{v} + \mathbf{w}, c \rangle)] = (1 \pm \epsilon) \exp\left( \frac{|\mathbf{v} + \mathbf{w}|^2}{2d} \right) \tag{21}$$

から, 定理が得られる. □

# 基本モデルに基づく単語ベクトルの学習

- $X_{vw}$ で窓の中で単語 $v, w$ が共起した回数を表すと, 提案モデルでは, $X$ の確率

$$L = \log \prod_{v,w} p(v,w)^{X_{vw}} \tag{22}$$

$$= \sum_{v,w} X_{vw} \log\, p(v,w) \tag{23}$$

を最大化するような単語埋め込みベクトル $\mathbf{w}$ を探せばよい.

# 基本モデルに基づく単語ベクトルの学習 (2)

- 準備 1: 総観測数 $N = \sum_{v,w} X_{vw}$ について

$$\Delta_{vw} = \log \frac{N\, p(v,w)}{X_{vw}} \tag{24}$$

とおけば,

$$\sum_{v,w} X_{vw}\Delta_{vw} = \sum_{v,w} X_{vw} \log \frac{N p(v,w)}{X_{vw}} \tag{25}$$

$$= \sum_{v,w} X_{vw}(\log N + \log p(v,w) - \log X_{vw}) \tag{26}$$

$$= \sum_{v,w} X_{vw} \log p(v,w) - c \tag{27}$$

よって,

$$L = \sum_{v,w} X_{vw} \log p(v,w) = \sum_{v,w} X_{vw}\Delta_{vw} + c\,. \tag{28}$$

# 基本モデルに基づく単語ベクトルの学習 (3)

- 準備 2: ここで,

$$\sum_{v,w} X_{vw} = \sum_{v,w} N p(v,w) \tag{29}$$

$$= \sum_{v,w} X_{vw} \exp\left(\log \frac{N\, p(v,w)}{X_{v,w}}\right) \tag{30}$$

$$= \sum_{v,w} X_{vw} e^{\Delta_{vw}} \simeq \sum_{v,w} X_{vw} \left(1 + \Delta_{vw} + \frac{\Delta_{vw}^2}{2}\right) \tag{31}$$

- これから,

$$\sum_{v,w} X_{vw} \Delta_{vw} \simeq -\frac{1}{2} \sum_{v,w} X_{vw} \Delta_{vw}^2 \,. \tag{32}$$

# 基本モデルに基づく単語ベクトルの学習 (4)

- よって,

$$L = \sum_{v,w} X_{v,w}\Delta_{vw} + c \tag{33}$$

$$\simeq -\frac{1}{2}\sum_{v,w} X_{vw}\Delta_{vw}^2 + c \tag{34}$$

$$= -\frac{1}{2}\sum_{v,w} X_{vw}\left(\log\frac{Np(v,w)}{X_{vw}}\right)^2 + c \tag{35}$$

$$= -\frac{1}{2}\sum_{v,w} X_{vw}(\log p(v,w) + \log N - \log X_{vw})^2 + c \tag{36}$$

$$= -\frac{1}{2}\sum_{v,w} X_{vw}\left(\frac{|\mathbf{v}+\mathbf{w}|^2}{2d} - 2\log Z \pm \epsilon + \log N - \log X_{vw}\right)^2 \tag{37}$$

# 基本モデルに基づく単語ベクトルの学習 (5)

- ゆえに, $L$ の最大化は以下の最小化と同値.

$$\min_{\mathbf{w},C} \sum_{v,w} X_{vw} \left( \frac{|\mathbf{v}+\mathbf{w}|^2}{2d} - \log X_{vw} - C \right)^2 \tag{38}$$

# 他モデルとの関係

- GloVe (Pennington+ 2014) の目的関数:

$$\sum_{v,w} f(X_{v,w})(\log X_{vw} - \mathbf{v}^T\mathbf{w} - s_v - s_w - C)^2 \tag{39}$$

(with $f(X) = \min(X^{3/4}, 100)$) において, バイアス項 $s_v, s_w$ の意味が明解に: $s_w = |\mathbf{w}|^2$.

- Word2vec (CBOW) の予測確率は

$$p(w_{t+1}|\{w_k\}_{i=1}^K) \propto \exp\left(\left\langle \mathbf{w_{t+1}}, \frac{1}{K}\sum_{k=1}^K \mathbf{w}_k \right\rangle\right) \tag{40}$$

  - $\frac{1}{K}\sum_{k=1}^K \mathbf{w}_k$ は $c_t$ の最尤推定量

# 意味方向ベクトルの学習

- Pennington+(2014) の議論によれば,

  man : woman = king : queen

  が成り立つのは, 確率で以下の関係が成り立つから.

$$\frac{p(x|\text{king})}{p(x|\text{queen})} \approx \frac{p(x|\text{man})}{p(x|\text{woman})} \tag{41}$$

  - この式が説明できればよい
- ある関係 $R$ について上の式が成り立っていたとすると, $R$ を満たすすべての単語ペア $(a, b)$ について

$$\frac{p(x|a)}{p(x|b)} = \nu_R(x) \cdot \zeta_{a,b,R}(x) \tag{42}$$

  であることが期待できる. 右辺第 2 項はノイズ項.

# 意味方向ベクトルの学習 (2)

- ここで定理 1 から, $p(v|w) = \dfrac{p(v,w)}{p(w)}$ は

$$\log p(v|w) = \frac{|\mathbf{v}+\mathbf{w}|^2}{2d} - 2\log Z \pm \epsilon - \frac{|\mathbf{w}|^2}{2d} + \log Z \pm \epsilon$$

$$= \frac{1}{2d}(|\mathbf{v}|^2 + 2\mathbf{v}^T\mathbf{w}) - \log Z \pm \epsilon \tag{43}$$

- これより,

$$\log \frac{p(v|w)}{p(v|w')} = \log p(v|w) - \log p(v|w') \tag{44}$$

$$= \frac{1}{2d}(2\mathbf{v}^T(\mathbf{w}-\mathbf{w}')) \pm \epsilon \tag{45}$$

$$= \frac{1}{d}\langle \mathbf{v}, \mathbf{w}-\mathbf{w}' \rangle \pm \epsilon\,. \tag{46}$$

  - 「$\mathbf{v}$ と $\mathbf{w}-\mathbf{w}'$ がどれだけ同じ方向を向いているか」が得られた!

# 意味方向ベクトルの学習 (3)

$$\frac{1}{d}\langle \mathbf{v}, \mathbf{w} - \mathbf{w}' \rangle + \epsilon = \log \nu_R(v) + \zeta_{a,b,R}(v) \tag{47}$$

$$\therefore \quad \langle \mathbf{v}, \mathbf{w} - \mathbf{w}' \rangle = d \log \nu_R(v) + d(\zeta_{a,b,R}(v) - \epsilon) \tag{48}$$

とおいてみる

- これが関係 $R$ を満たす任意の単語 $v$ について成り立つから, $V$ をそれらの $m$ 個の単語ベクトルを並べた $m \times d$ の行列, $\nu_R$, $\zeta_R$ を $m$ 次元のベクトルにとれば, 行列表記で

$$V(\mathbf{w} - \mathbf{w}') = d \log \nu_R + \zeta'_{a,b,R} \tag{49}$$

$$(\text{ただし}, \quad \zeta'_{a,b,R}(v) = d(\zeta_{a,b,R}(v) - \epsilon)$$

  と書くことができる.
  - これは $m \times d$ 次元の線形回帰モデル!
  - $V$ の擬似逆行列 $V^\dagger$ を左からかければ, 回帰が解ける.

# 低次元化とノイズ

- (49) 式は有用だが, ノイズ $\zeta'_{a,b,R}$ が大きい
  - 次元 $d$ が高次元の場合に, $d \log \nu_R$ を上回るほどのノイズ
- $d \ll n$ で低次元の場合に, ノイズが少なくなる!
- これが次の定理

# 定理:3

- 次の条件が成り立っているとする.
  (1) $V$ の最大固有値が, 固有値の幾何平均より $c_1$ 倍以上大きい
  (2) $V$ の左特異ベクトルは, $\zeta'_{a,b,R}$ とほぼ無相関
  (内積が高々 $c_2|\zeta'_{a,b,R}|/\sqrt{n}$)
  (3) $V$ の列のノルムの最大値は $O(\sqrt{d})$
- このとき, $d$ 次元に削減した場合のノイズ $\bar{\zeta}_{a,b,R}$ は

$$|\bar{\zeta}_{a,b,R}|_2 \lesssim |\zeta'_{a,b,R}|_2 \frac{\sqrt{d}}{2} . \tag{50}$$

  - 証明は少し面倒, 論文の Appendix A.10-A.12 参照

# 実験

- 観測された共起カウント行列 $X_{vw}$ を $\max(X_{vw}, 100)$ で置き換えた観測値に対して提案法を適用したものを, SN (Squared Norm) と呼ぶ
  - Word2vec, GloVe 等と比べ, ほとんどヒューリスティックなパラメータがない
- GloVe, CBOW, skip-gram と比較 (それぞれデフォルトパラメータ設定)

# 実験 (2)

- 単語アナロジータスクの精度
  - Google(G): 7874 個の意味的ペア, MSR(M): 10167 個の文法的ペア
  - 提案法は単語の順序を使っていないので, 後者では不利なはず
  - 生成モデルでも充分に高い精度

| | Relations | **SN** | GloVe | CBOW | skip-gram |
|---|---|---|---|---|---|
| G | semantic | 0.84 | 0.85 | 0.79 | 0.73 |
| | syntactic | 0.61 | 0.65 | 0.71 | 0.68 |
| | total | 0.71 | 0.73 | 0.74 | 0.70 |
| M | adjective | 0.50 | 0.56 | 0.58 | 0.58 |
| | noun | 0.69 | 0.70 | 0.56 | 0.58 |
| | verb | 0.48 | 0.53 | 0.64 | 0.56 |
| | total | 0.53 | 0.57 | 0.62 | 0.57 |

# 実験 (3)

- “意味方向” の存在を示す
- 回帰行列の第 1 固有ベクトルを使った場合 (1st) と, 第 2 固有ベクトルを使った場合 (2nd)

| relation | cap-com | cap-wor | adj-adv | opp |
|---|---|---|---|---|
| 1st | $0.65 \pm 0.07$ | $0.61 \pm 0.09$ | $0.35 \pm 0.17$ | $0.42 \pm 0.16$ |
| 2nd | $0.02 \pm 0.28$ | $0.00 \pm 0.23$ | $0.07 \pm 0.24$ | $0.01 \pm 0.25$ |

  - 第 1 主成分のみが強い相関, 第 2 主成分はほとんどノイズ
  - 「意味方向」の存在!

# 実験 (4)

- 「意味方向」が存在することが分かったので, たとえば “男性–女性” の方向を得るのに “king–queen” だけでなく, “boy–girl” なども使えばよりロバストに
- $k$ 個の同義ペアを使った実験 (RD: relation direction)
  - ペアは多数の質問から, $\mathbf{v}-\mathbf{w}$ を $K$ 平均クラスタリングして獲得

| | **SN** | GloVe | CBOW | skip-gram |
|---|---|---|---|---|
| w/o **RD** | 0.71 | 0.73 | 0.74 | 0.70 |
| **RD**($k = 20$) | 0.74 | 0.77 | 0.79 | 0.75 |
| **RD**($k = 30$) | 0.79 | 0.80 | 0.82 | 0.80 |
| **RD**($k = 40$) | 0.76 | 0.80 | 0.80 | 0.77 |

# まとめ

- ランダムウォークする文脈から単語が生まれる生成モデルを考えることで, 単語の同時確率が計算できる→
  - 単語の相互情報量が単語ベクトルの内積になることが説明できる
  - 「意味方向」が単語ベクトルの内積で計算できることが説明できる
- これまでの理論的研究は高次元の設定に限られ, なぜ低次元の埋め込みが必要なのか理解されていなかった
  - 低次元に埋め込むことで, 単語ベクトルが空間上一様に分布することになり, 真の値へのノイズが高次元の場合より減る→ノイズに埋もれない
- 「意味方向」の存在を回帰モデルの固有ベクトルから示し, 多数の単語ペアから求める実験で効果を実証