鏡映変換に基づく埋め込み空間上の単語属性変換

石橋 陽一，須藤 克仁 ，吉野 幸一郎，中村 哲

奈良先端科学技術大学院大学

背景

- 埋め込み空間上のベクトルの属性を反転させる新たな表現学習の枠組みを提案
- アナロジーに基づく単語属性変換には事前知識が必要
- 事前知識を用いない変換を行うために鏡映変換を導入した手法を提案

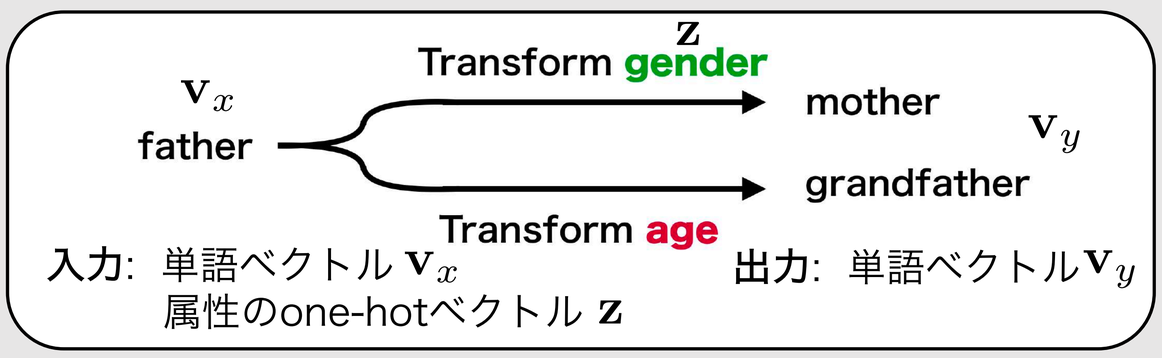
結論

- ✓ 鏡映変換に基づく手法によって事前知識を用いず45.8%の精度で単語の属性を反転させることに成功　（E.g.，girl ⇒ boy，　boy ⇒ girl）
  - ✓ 知識を用いるアナロジーに基づく手法やMLPよりも高精度
- ✓ 鏡映変換は高い安定性を持ち、属性を持たない単語は96%以上の高精度で変化させないことを示した　（E.g.，apple ⇒ apple ）

埋め込み空間上の単語属性変換

どのようなタスクか?

埋め込み空間上の単語ベクトルの特定の属性を変換



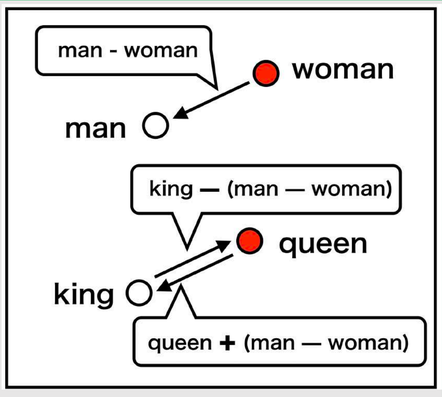
入力: 単語ベクトル $\mathbf{v}_x$
属性のone-hotベクトル $\mathbf{z}$

出力: 単語ベクトル $\mathbf{v}_y$

アナロジーに基づく属性変換

- アナロジーで単語の属性を変換できる
  - 例： $\mathbf{v}_{king} - (\mathbf{v}_{man} - \mathbf{v}_{woman}) \approx \mathbf{v}_{queen}$
- 問題点 ： 入力単語が男性・女性のどちらに属すかで変換に用いる演算が変わる ＝ 事前知識が必要

$$f_{\mathbf{z}}(\mathbf{v}_x) = \begin{cases} \mathbf{v}_x - (\mathbf{v}_{man} - \mathbf{v}_{woman}) & \text{if} \quad x \in \mathcal{M}, \\ \mathbf{v}_x + (\mathbf{v}_{man} - \mathbf{v}_{woman}) & \text{if} \quad x \in \mathcal{F}. \end{cases}$$



- ゴール：知識を用いず単語属性を変換 ＝ 同じ関数で変換

$$\left.\begin{aligned} \mathbf{v}_{man} &= f_{\mathbf{z}}(\mathbf{v}_{woman}) \\ \mathbf{v}_{woman} &= f_{\mathbf{z}}(\mathbf{v}_{man}) \end{aligned}\right\} \quad \mathbf{v}_{man} = f_{\mathbf{z}}(\, f_{\mathbf{z}}(\mathbf{v}_{man})\,)$$

このような関数が望ましい

鏡映変換に基づく属性変換

鏡映変換とは?

- 「鏡」と呼ばれる超平面によって２つのベクトルの位置を相互に反転させる写像
- 二回繰り返すと恒等写像になる（望ましい関数の条件）

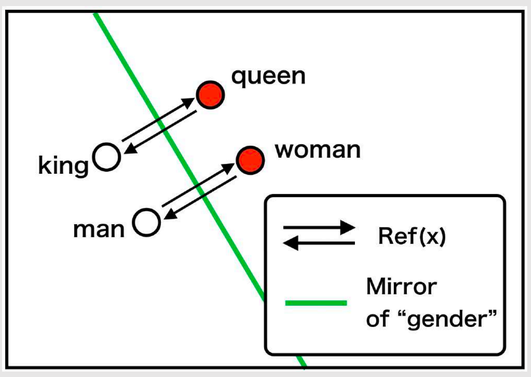
$$Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v}) = \mathbf{v} - 2\frac{(\mathbf{v}-\mathbf{c})^{\mathrm{T}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\mathbf{a}$$

- $Ref_{a,c}(x)$: 鏡映変換
- $\mathbf{v}$: 入力ベクトル
- $\mathbf{a}, \mathbf{c}$: 鏡を決定するパラメタ
- $\mathbf{v} = Ref_{\mathbf{a},\mathbf{c}}(\, Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v})\,)$

単語属性変換への適用

- 入力ベクトル $\mathbf{v}_x$ の属性 $\mathbf{z}$ を反転させたベクトル $\mathbf{v}_t$ を予測

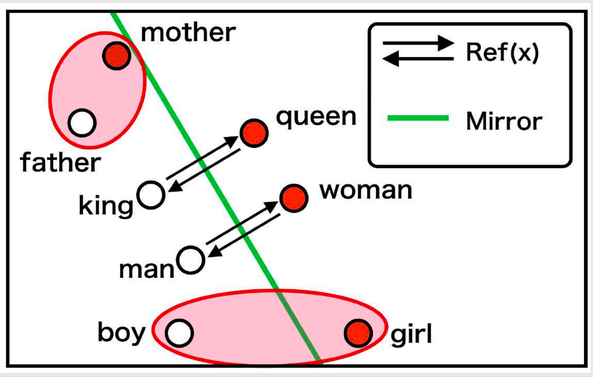
$$\mathbf{v}_t \approx \mathbf{v}_y = Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v}_x)$$



- 属性を反転させる鏡を学習
  - ２つの方法

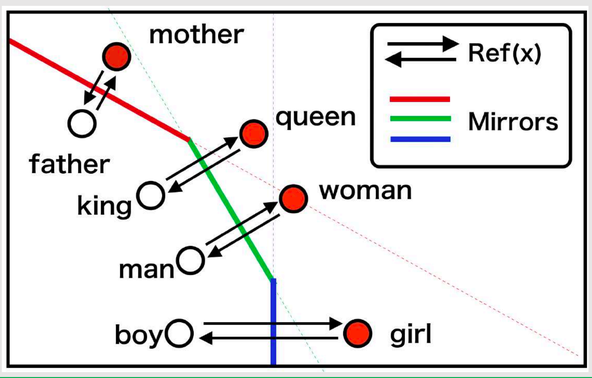
属性ベクトル $\mathbf{z}$ のみから鏡を推定
➢ 線形分離不可能

$$\mathbf{a} = MLP(\mathbf{z})$$
$$\mathbf{c} = MLP(\mathbf{z})$$



入力単語ベクトル $\mathbf{v}_x$ を用いて鏡を推定（Parameterized mirror ）

$$\mathbf{a} = MLP([\mathbf{z}; \mathbf{v}_x])$$
$$\mathbf{c} = MLP([\mathbf{z}; \mathbf{v}_x])$$



損失関数

- 属性を持つ単語は変換されるように，持たない単語は変換されないように学習

$$\mathcal{L}(\Theta) = \frac{1}{|\mathcal{A}|}\sum_{(x_i,t_i,\mathbf{z}_i)\in\mathcal{A}} (\mathbf{v}_{y_i} - \mathbf{v}_{t_i})^2 + \frac{1}{|\mathcal{N}|}\sum_{x_j\in\mathcal{N}} (\mathbf{v}_{y_j} - \mathbf{v}_{x_j})^2$$

$(\text{man}, \text{woman}, \text{gender}) \in \mathcal{A}$　　$\text{apple} \in \mathcal{N}$

実験

- データセット：性別単語 106 ペア(train/val/test = 58/24/24)
  - $|\mathcal{A}| = 58, |\mathcal{N}| = 4$　(訓練データ)
  - イテレーション毎にガウシアンノイズを入力ベクトルに加えデータを拡張
- 評価方法（変換精度と安定性） $\delta_k(t) = \begin{cases} 1 & \text{if} \quad t \in \mathcal{S}_k, \\ 0 & \text{otherwise}, \end{cases}$ $\mathcal{S}_k$ は属性変換の出力の上位 k 近傍の単語集合

$$\text{Accuracy@}k = \frac{1}{|\mathcal{A}_{\text{test}}|}\sum_{(x_i,t_i,\mathbf{z}_i)\in\mathcal{A}_{\text{test}}} \delta_k(t_i) \qquad \text{Stability@}k = \frac{1}{|\mathcal{N}_{\text{test}}|}\sum_{x_i\in\mathcal{N}_{\text{test}}} \delta_k(x_i)$$

Ref	鏡映変換
Ref + PM	Parameterized mirrorを用いた鏡映変換
Diff	アナロジーに基づく変換。１つ単語ペアの差分ベクトルを用いる
AvgDiff	アナロジーに基づく変換。訓練単語ペアの差分ベクトルの平均を用いる

結果

- 鏡映変換は入力単語がどの属性を持つかといった知識を用いずに変換できている （最高精度）
- 鏡映変換の安定性は高い（属性を持たない1000単語を99%変換しない）

Method	know ledge	Accuracy (%)			Stability (%)		
		Mean@3	@1	@3	Mean@3	@1	@3
Ref		40.27	25.00	54.16	**99.53**	**99.50**	**99.60**
Ref + PM		**55.55**	**45.83**	62.50	96.90	96.50	97.30
MLP		**19.44**	**8.33**	33.33	**0.00**	**0.00**	**0.00**
Diff (-)		21.45	**8.33**	**30.89**	76.02	69.04	80.35
AvgDiff (-)		23.61	**8.33**	33.33	97.17	96.90	97.30
Diff	✓	40.65	15.94	57.67	-	-	-
AvgDiff	✓	47.20	12.50	**66.66**	-	-	-

鏡映変換は最高で99.9%の安定性を持つがMLPは $|\mathcal{N}|$ を増やして学習しても安定しない

Method	Accuracy @1 (%)				Stability @1 (%)			
	$\|\mathcal{N}\|=0$	4	10	50	0	4	10	50
Ref	20.83	25.00	25.00	25.00	**97.10**	**99.50**	98.40	95.60
Ref + PM	**45.83**	**45.83**	**37.50**	**29.16**	35.80	96.90	**99.90**	**99.30**
MLP	**4.16**	**8.33**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

鏡映変換による変換例

x	**when my father was a boy, he had liked the lady who is an actress**
Ref (x)	**when my mother was a girl, she had liked the gentleman who is an actor**
Ref (Ref (x))	**when my father was a boy, he had liked the lady who is an actress**

他の属性への適用例

Original	she is my mother
+ Gender	**he** is my **father**
+ Age	he is my **grandfather**
+ Tense	he **was** my grandfather