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[1,a)]　須藤 克仁[1,b)]　吉野 幸一郎[1,c)]　中村 哲[1,d)]

概要：本研究では鏡映変換に基づく埋め込み空間上の単語の属性変換を提案する．自己相互情報量 (PMI) に基づく単語埋め込みは，$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$ といったアナロジーが成立することが知られている．このアナロジーを用いて入力単語 x を “king” から “queen” に，また “queen” から “king” に変換することが可能である．一方，アナロジーによる変換は x が男性か女性かどうかで演算が変わるため，x の属性に関する知識が必要となるが，そのような知識は無数にあるため全て付与することは不可能である．そこで本研究では，そのような知識を用いることなく特定の属性を持つ単語を変換するため，理想的な性質を持つ写像である鏡映変換を導入する．鏡映変換は同じ写像でベクトルの位置を相互に反転させる変換であるため，入力単語ベクトルが目的の属性を持つかどうかにかかわらず変換できる．性別属性を変換する実験の結果，提案手法は属性の知識を用いることなく，性別単語を 45.8%の精度で相互に変換できることが示された．また性別属性を持たない単語に鏡映変換を適用した結果，最大で 99.9%が変換されず，鏡映変換は目的属性を持つ単語のみを変化させる非常に高い安定性を持つことが示唆された．

1. はじめに

分散表現 [1] はデータとデータの間にある類似性を捉えた数値表現である．自然言語処理分野においては単語の分散表現の研究が行われてきた [2], [3], [4]．word2vec のモデルの一つである負例サンプリングを用いた Skip-gram モデル（SGNS）[3] や GloVe[4] などの埋め込み手法で得られたベクトルは線形性を持つことが知られている．最も有名な例は $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$ である．このような性質（アナロジー・線形類推関係）を理論的に説明する研究も行われてきており，近年では SGNS が Shifted Pointwise Mutual Information （sPMI）[5] と等価であることが証明されている [6], [7], [8], [9], [10]（6 節）．すなわち PMI に基づいた埋め込みを行うことで線形類推関係が獲得される．このような線形類推関係を用いて分散表現の属性を変換することができる．例えば上の例では $\overrightarrow{king}$ から $\overrightarrow{queen}$ へ性別属性を変換している．本研究では埋め込み空間上で分散表現の持つ属性を制御して変換する新たな表現学習に取り組む．この技術は言語データの拡張や埋め込み空間上の推論などに様々な応用が期待できる（8 節）．例えばデータ拡張への応用では，元文の “He is a boy.” の各単語の性別属性を変換し “She is a girl.” という新たな文を作ることができる．一方でアナロジーに基づく属性変換では入力単語がどのような属性を持っているのか，変換をする前に知っておかなくてはならない（3 節）．しかし入力単語の属性に関する知識を全ての単語に付与することは困難である．そこで本研究では属性知識を用いず埋め込み空間上で単語属性を反転させる手法を提案する（4 節）．本研究の貢献は以下のとおりである

図 1　単語属性変換

- ある分散表現の特定の属性を反転した分散表現を得る新たな表現学習の枠組みの提案
- 知識を用いることなく特定の属性を持つ単語のみ変換する手法の提案

2. 埋め込み空間上の単語属性変換

本稿では単語を x, その単語の分散表現を $\mathbf{v}_x$ と表記する．なお，この分散表現は SGNS などで事前に学習されているとする [2], [3]．本研究で扱うタスクでは入力として単語 x と，変化させたい属性の one-hot ベクトル $\mathbf{z}$，そして正解単語として t が与えられる．またこれらをまとめた集合を $(x, t, \mathbf{z}) \in \mathcal{A}$ とする（例：$(man, woman, \mathbf{z}_{\text{gender}}) \in \mathcal{A}$）．本タスクでは $(x, t, \mathbf{z}) \in \mathcal{A}$ が与えられたとき，属性 $\mathbf{z}$ につ

[1] 奈良先端科学技術大学院大学 先端科学技術研究科
[a)] ishibashi.yoichi.ir3@is.naist.jp
[b)] sudoh@is.naist.jp
[c)] koichiro@is.naist.jp
[d)] s-nakamura@is.naist.jp

いて反転させる関数 $f_{\mathbf{z}}$ に x の分散表現 $\mathbf{v}_x$ を入力し出力 $\mathbf{v}_y$ を得る．そして $\mathbf{z}$ の属性を反転させた分散表現 $\mathbf{v}_t$ と $\mathbf{v}_y$ が単語埋め込み空間上での最近傍であるかで評価を行う．

$$\mathbf{v}_t \approx \mathbf{v}_y = f_{\mathbf{z}}(\mathbf{v}_x). \tag{1}$$

例えば $(man, woman, \mathbf{z}_{gender})$ が与えられたとき，性別の属性変換 $f_{\mathbf{z}_{gender}}$ によって $\mathbf{v}_{man}$ を $\mathbf{v}_{woman}$ に変換する．また変換関数 $f_{\mathbf{z}}$ の性能は変換精度と安定性スコアで評価する（5.2 節）．

3. アナロジーに基づく属性変換

SGNS[2], [3] や GloVe[4] で得られた分散表現は線形類推関係を持つことが知られている [11], [12], [13] ．例えば SGNS で得られた分散表現 $\mathbf{v}_{queen}$ は式 2 の右辺で得られたベクトルと近い位置に埋め込まれている．

$$\mathbf{v}_{queen} \approx \mathbf{v}_{king} - \mathbf{v}_{man} + \mathbf{v}_{woman}, \tag{2}$$

$$\approx \mathbf{v}_{king} - (\mathbf{v}_{man} - \mathbf{v}_{woman}). \tag{3}$$

ここで式 2 を変形すると式 3 が得られる．式 3 より，$\mathbf{v}_{king}$ から差分ベクトル $\mathbf{v}_{man} - \mathbf{v}_{woman}$ を引くことで $\mathbf{v}_{queen}$ に変換できる．このように性別属性を持つ単語ペアの差分ベクトルを用いて性別属性を変換できる．一般化のために，男性と女性のような一対一の対応を持つ単語（father と mother など）の中で，片方の属性（例：男性）を持つ単語の集合を $\mathcal{M}$, もう片方の属性（例：女性）を持つ単語の集合を $\mathcal{F}$，属性対の差分ベクトルを $\mathbf{d}$ とすると，アナロジーに基づく単語属性の変換式は次のようになる．

$$f_{\mathbf{z}}(\mathbf{v}_x) = \begin{cases} \mathbf{v}_x - \mathbf{d} & \text{if} \quad x \in \mathcal{M}, \\ \mathbf{v}_x + \mathbf{d} & \text{if} \quad x \in \mathcal{F}. \end{cases} \tag{4}$$

式 4 より，入力単語が $\mathcal{M}$ に属すか $\mathcal{F}$ に属すかによって演算が変わる．例えば性別属性の場合，x が男性を表す単語であれば差分ベクトル $\mathbf{d}$ を引き，女性を表す単語であれば $\mathbf{d}$ を加えることで x の性別を変換する．これはつまり，アナロジーに基づく変換には入力単語 x の属性に関する知識（x が $\mathcal{M}$ に属すか $\mathcal{F}$ に属すか）が必要であることを示している．このような知識は無数に存在するため，あらかじめすべての単語に属性知識を付与することは困難である．

4. 鏡映変換に基づく属性変換

4.1 知識を用いない変換

そこで属性知識を用いない理想的な変換を考える．$m \in \mathcal{M}$, $w \in \mathcal{F}$ とすると，属性知識を用いず単語の属性を変換する変換関数 $f_{\mathbf{z}}$ は

$$\mathbf{v}_m = f_{\mathbf{z}}(\mathbf{v}_w), \tag{5}$$

$$\mathbf{v}_w = f_{\mathbf{z}}(\mathbf{v}_m), \tag{6}$$

のような性質を持つことが望ましい．つまり変換関数 $f_{\mathbf{z}}$ は入力単語 m または w が $\mathcal{M}$ に属すか $\mathcal{F}$ に属すか考慮することなく，同じ写像によって変換を行う．式 5, 6 をまとめると，

$$\forall m \in \mathcal{M}, \quad \mathbf{v}_m = f_{\mathbf{z}}(\, f_{\mathbf{z}}(\mathbf{v}_m)\,), \tag{7}$$

もしくは

$$\forall w \in \mathcal{F}, \quad \mathbf{v}_w = f_{\mathbf{z}}(\, f_{\mathbf{z}}(\mathbf{v}_w)\,), \tag{8}$$

となる．したがって理想的な写像 $f_{\mathbf{z}}$ とは二回適用すると恒等写像となるような変換である．このような写像は対合と呼ばれている．ただし $f_{\mathbf{z}}$ 自身が恒等写像であるもの（例：1 をかける）は除く．例えば最も簡単な対合は $f(\mathbf{v}) = -\mathbf{v}$ である．

4.2 鏡映変換

鏡映変換は対合の一種であり，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．したがって同一の鏡による鏡映変換（$Ref_{\mathbf{a},\mathbf{c}}$）を二回繰り返すと恒等写像となる．

$$\forall \mathbf{v} \in \mathbb{R}^n, \quad \mathbf{v} = Ref_{\mathbf{a},\mathbf{c}}(\, Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v})\,). \tag{9}$$

標準内積が与えられた n 次元計量ベクトル空間 $\mathbb{R}^n$ における鏡映変換は

$$Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v}) = \mathbf{v} - 2\frac{(\mathbf{v}-\mathbf{c})^{\mathrm{T}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\mathbf{a}, \tag{10}$$

と定義される．ここで $\mathbf{a}$ および $\mathbf{c}$ はそれぞれ鏡を決定するパラメタであり，$\mathbf{a}$ は鏡に直交するベクトル，$\mathbf{c}$ は鏡が通る $\mathbb{R}^n$ 上の点である．なお $\mathbf{a}$，$\mathbf{c}$ そして $\mathbf{v}$ の次元数は等しい．

4.3 埋め込み空間における属性変換への適用

埋め込み空間上で鏡映変換を行い特定の属性（例：性別）を持つ単語の位置を反転させる．このとき鏡映変換における鏡を属性 $\mathbf{z}$ から推定する．ここで鏡は 2 つのベクトル $\mathbf{a}$，$\mathbf{c}$ によって一意に決まるため，$\mathbf{a}$ と $\mathbf{c}$ を属性 $\mathbf{z}$ から推定する．本研究では全結合の多層パーセプトロン（MLP）によって各属性ごとに $\mathbf{a}$ と $\mathbf{c}$ を推定する（式 11, 12，図 2）．

$$\mathbf{a} = MLP(\mathbf{z}), \tag{11}$$

$$\mathbf{c} = MLP(\mathbf{z}). \tag{12}$$

そして入力単語ベクトル $\mathbf{v}_x$ の属性を反転させたベクトルを鏡映変換し $\mathbf{v}_y$ を得る．

$$\mathbf{v}_y = Ref_{\mathbf{a},\mathbf{c}}(\mathbf{v}_x). \tag{13}$$

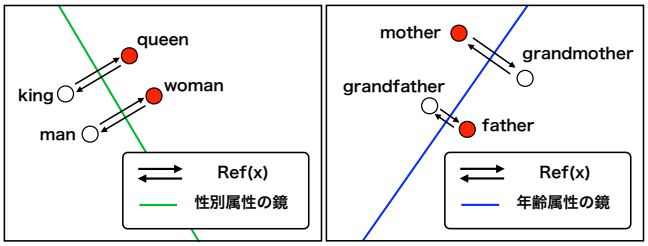

図 2 鏡映変換における鏡（超平面）を属性 $\mathbf{z}$ から推定することによって埋め込み空間上のベクトルの位置を相互に反転させることができる．

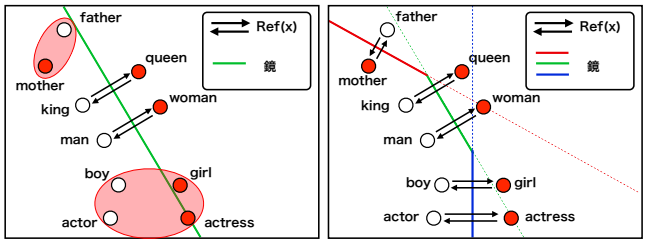

図 3 鏡の推定方法の比較．左は属性 $\mathbf{z}$ のみから鏡を推定するため線形分離不可能な単語ペアが存在する．右の Parameterized Mirror は属性 $\mathbf{z}$ と入力単語ベクトル $\mathbf{v}_x$ から鏡を逐一推定するため柔軟に変換できる．

4.4 Parameterized Mirror

式 11，12 によって鏡を推定する場合，図 3 左のように線形分離不可能なデータが存在す場合に正しく変換できない．これは属性 $\mathbf{z}$ から $\mathbf{a}$ および $\mathbf{c}$ を推定しているため鏡が一つに固定されてしまうことが原因である．この問題を解決するため，属性 $\mathbf{z}$ に加えて入力単語ベクトル $\mathbf{v}_x$ も $\mathbf{a}$ および $\mathbf{c}$ の推定に用いる（式 14，15）．$\mathbf{v}_x$ を用いることで鏡を入力単語ごとに逐次的に決定することが可能となる．

$$\mathbf{a} = MLP([\mathbf{z}; \mathbf{v}_x]), \tag{14}$$

$$\mathbf{c} = MLP([\mathbf{z}; \mathbf{v}_x]). \tag{15}$$

ここで $[\cdot;\cdot]$ はベクトルの列方向の連結を表す．このようにして鏡を学習対象（Parameterized Mirror）とすることで，未学習データを入力した際には鏡が柔軟に推定される．例えば図 3 の性別属性の変換において，$\mathbf{v}_{boy}$ から $\mathbf{v}_{girl}$ への変換が成立するような鏡（青線）を学習しておく．ここで未学習である $\mathbf{v}_{actor}$ が $\mathbf{v}_{boy}$ と類似している場合，$\mathbf{v}_{boy}$ で学習した鏡と類似した鏡が推定されるため，$\mathbf{v}_{actor}$ から $\mathbf{v}_{actress}$ へ変換することが可能になると考えられる（図 3）．

4.5 損失関数

損失関数 $\mathcal{L}$ を以下のように定義する．

$$\mathcal{L}(\Theta) = \frac{1}{|\mathcal{A}|}\sum_{(x_i,t_i,\mathbf{z}_i)\in\mathcal{A}}(\mathbf{v}_{y_i} - \mathbf{v}_{t_i})^2 + \frac{1}{|\mathcal{N}|}\sum_{x_j\in\mathcal{N}}(\mathbf{v}_{y_j} - \mathbf{v}_{x_j})^2. \tag{16}$$

ここで $\mathcal{N}$ は属性 $\mathbf{z}$ を持たない単語の集合である．例えば $\mathbf{z}$ が性別属性の場合 $apple \in \mathcal{N}$ などが考えられる．$\frac{1}{|\mathcal{N}|}\sum_{x_j\in\mathcal{N}}(\mathbf{v}_{y_j} - \mathbf{v}_{x_j})^2$ は属性を持たない単語を変換関数によって変化させないための制約である．また Θ は学習されるパラメタである．$t+1$ 回目の更新時に Θ_{t+1} を推定するため以下の最適化問題を解く．

$$\Theta_{t+1} \leftarrow \arg\min_{\Theta_t} \mathcal{L}(\Theta_t). \tag{17}$$

5. 実験

5.1 実験設定

予備実験として性別属性の変換を行った．学習済みの単語埋め込みモデルとして Google が公開している word2vec[*1][3] を使用した．この学習済みモデルの単語ベクトルは $n = 300$ 次元である．Google アナロジー test set [*2] [2] から抽出した性別属性を持つ単語対と独自に収集した単語対を加え合計 53 単語対を取得した．入力単語と目的単語を入れ替え，合計 106 単語対を用いて実験を行った（$|\mathcal{A}| = 106$）．データセット $\mathcal{A}$ は train/val/test = 58/24/24 に分割した．また学習に用いる属性なし単語集合のサイズは $|\mathcal{N}_{\text{train}}| = 4$ とした．学習データが小規模 $|\mathcal{A}_{\text{train}}| + |\mathcal{N}_{\text{train}}| = 62$ であるため，イテレーション毎に $\sigma = 0.1$ のガウシアンノイズを $\mathbf{v}_x$ へ加え学習データを拡張した．

5.2 評価方法

評価は変換精度（Accuracy）と安定性スコア（Stability）で行った．変換精度は性別属性をもつ単語が正しく変換されているかを表し（例：“boy” → “girl”），安定性スコアは性別属性を持たない単語が変換されないかを表す（例：“human” → “human”）．これらは以下の式で計算した．

$$\delta_k(t) = \begin{cases} 1 & \text{if} \quad t \in \mathcal{S}_k, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

$$\text{Accuracy@}k = \frac{1}{|\mathcal{A}_{\text{test}}|}\sum_{(x_i,t_i,\mathbf{z}_i)\in\mathcal{A}_{\text{test}}} \delta_k(t_i), \tag{19}$$

$$\text{Stability@}k = \frac{1}{|\mathcal{N}_{\text{test}}|}\sum_{x_i\in\mathcal{N}_{\text{test}}} \delta_k(x_i). \tag{20}$$

ここで，$\mathcal{S}_k = \{y_i\}_{i=1}^k$ は出力ベクトル $\mathbf{v}_y$ の埋め込み空間上での上位 k 近傍単語集合である．なおテスト時の属性なし単語集合のサイズ $|\mathcal{N}_{\text{test}}|$ は，学習済みモデルの語彙（300 万単語）から $\mathcal{N}_{\text{train}}$，$\mathcal{A}$ 中の単語を除外してランダムに 1000 単語サンプリングして使用した（$|\mathcal{N}_{\text{test}}| = 1000$）．

比較のため以下の提案手法とベースラインを用いた．

Ref
提案手法．Parameterized Mirror を用いない．鏡の推

*1 https://code.google.com/archive/p/word2vec/
*2 http://download.tensorflow.org/data/questions-words.txt

表 1　変換精度と安定性スコアの比較

Method	Knowledge	Accuracy (%)				Stability (%)			
		Mean@3	@1	@2	@3	Mean@3	@1	@2	@3
Ref		40.27	25.00	41.66	54.16	**99.53**	**99.50**	**99.50**	**99.60**
Ref + PM		**55.55**	**45.83**	58.33	62.50	96.90	96.50	96.90	97.30
MLP		19.44	8.33	8.33	33.00	0.00	0.00	0.00	0.00
Diff (+)		21.31	7.61	25.57	30.74	83.29	79.36	84.64	85.87
AvgDiff (+)		23.61	4.16	33.33	33.33	98.13	98.10	98.10	98.20
Diff (−)		21.45	8.33	25.14	30.89	76.02	69.04	78.67	80.35
AvgDiff (−)		23.61	8.33	29.16	33.33	97.17	96.90	97.30	97.30
Diff	✓	40.65	15.94	48.41	57.67	-	-	-	-
AvgDiff	✓	47.20	12.50	**62.50**	**66.66**	-	-	-	-

表 2　学習データ $|\mathcal{N}_{\text{train}}|$ を増加させた際の変換精度・安定性の変化

Method	Accuracy@1 (%)				Stability@1 (%)			
	$\|\mathcal{N}_{\text{train}}\|$=0	4	10	50	$\|\mathcal{N}_{\text{train}}\|$=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

定には 2 層の全結合多層パーセプトロンを用いた．

Ref+PM

提案手法．Parameterized Mirror を用いる．鏡の推定には 2 層の全結合多層パーセプトロンを用いた．

MLP

全結合多層パーセプトロン．$\mathbf{v}_y = MLP([\mathbf{v}_x; \mathbf{z}])$ で変換する．最も変換精度が高かった 2 層の MLP の結果を記載した．

Diff

アナロジーに基づいた手法（式 4）．差分ベクトル $\mathbf{d} = \mathbf{v}_m - \mathbf{v}_w$ を用いて変換する．ここで m,w はそれぞれ男性単語集合 $\mathcal{M}$, 女性単語集合 $\mathcal{F}$ の要素である．m,n は訓練データ $\mathcal{A}_{\text{train}}$ から一つづつサンプリングし変換精度および安定性スコアを計算することを $\mathcal{A}_{\text{train}}$ のすべての要素について行い精度とスコアを平均した．

Diff (+)

$\forall \mathbf{v}_x \in \mathbb{R}^n, f_{\mathbf{z}}(\mathbf{v}_x) = \mathbf{v}_x + \mathbf{d}$ として変換する手法．

Diff (−)

$\forall \mathbf{v}_x \in \mathbb{R}^n, f_{\mathbf{z}}(\mathbf{v}_x) = \mathbf{v}_x - \mathbf{d}$ として変換する手法．

AvgDiff

アナロジーに基づいた手法（式 4）．平均差分ベクトル $\mathbf{d}_{avg} = \frac{1}{|\mathcal{G}_{\text{train}}|} \sum_{(m_i,w_i)\in\mathcal{G}} (\mathbf{v}_{m_i} - \mathbf{v}_{w_i})$ を用いて変換する．ただし $\mathcal{G}$ は訓練データ $\mathcal{A}_{\text{train}}$ 中の男性単語 m と女性単語 w の集合である（$(w_i, m_i) \in \mathcal{G}$）．

AvgDiff (+)

$\forall \mathbf{v}_x \in \mathbb{R}^n, f_{\mathbf{z}}(\mathbf{v}_x) = \mathbf{v}_x + \mathbf{d}_{avg}$ として変換する手法．

AvgDiff (−)

$\forall \mathbf{v}_x \in \mathbb{R}^n, f_{\mathbf{z}}(\mathbf{v}_x) = \mathbf{v}_x - \mathbf{d}_{avg}$ として変換する手法．

なお Optimizer には Adam（$\alpha = 0.001$）[14] を用いた．またネットワークの重みを固定してチューニングを行うことで特定の手法が有利・不利になる状況を避けた．学習ベースの手法（Ref，Ref+PM，MLP）はチューニング時に Dropout[15] や Batch Normalization[16] などの正則化を加えたモデルも同様に学習させたが，今回は正則化を用いないモデルの変換精度が高い結果となったため記載していない．

5.3　変換精度と安定性

5.3.1　実験結果

表 1 は提案手法および比較手法の変換精度と安定性スコアを示している．Knowledge は変換の際に属性知識（例：$actor \in \mathcal{M}$）を用いるかを示している．表 1 より変換精度で最も優れている手法は Parameterized Mirror を用いた提案手法（Ref+PM）であることがわかった．知識を用いて変換する手法（Diff, AvgDiff）よりも高い変換精度を達成している．一方で，変換精度が最も低い手法は MLP であった．今回の実験においては別の空間に写像せず元の埋め込み空間上で変換を行う手法（MLP 以外）が優れている結果となった．Diff(+)(−) また AvgDiff(+)(−) では両者とも (+) より (−) の変換精度が高い．これは男性女性の単語間で単語の使われ方に偏り（bias）があるためと考えられる [17], [18].

次に安定性の評価では提案手法（Ref, Ref+PM）と平均差分ベクトル $\mathbf{d}_{avg}$ で変換するアナロジーベースの手法（AvgDiff(−), AvgDiff(+)）のスコアが高い結果となった．最も安定性が低い手法は MLP で，1000 単語中すべての単語を変換してしまっている．安定性が高い手法に注目し変換精度を比較すると Ref と同様に知識を用いないアナロジーベース手法の最高値が 8.33%，知識を用いる手法でも

表 3 文 X が与えれらたときの変換結果（“a”，“,” は未知語のため入力しない）

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**
$MLP(x)$	she herself daughter her (a) mother (,) herself she niece she he her mother her himself

表 4 エラー分析の結果

Case	#Case/$\lvert\mathcal{A}_{\text{test}}\rvert$ (%)
A	11/24 (45.8)
B	11/24 (45.8)
C	2/24 (8.3)
D	0/24 (0.0)

表 5 失敗例（括弧内の数値は $Ref(x)$ との cos 類似度）

x	t	$Ref(x)$	
		@1	@2
gentlemen	ladies	Excellencies_ladies (0.734)	MODERATOR_Ladies (0.733)
ladies	gentlemen	ladies (0.517)	gentleman (0.469)
king	queen	princess (0.756)	queen (0.731)
queen	king	prince (0.724)	queen (0.665)

最高値が 12.50%と低い．一方，鏡映変換に基づく提案手法は変換精度を維持（45.83%）しつつ 96%以上の高い安定性を持つことが示された．

属性なし単語の学習量が少ない可能性も考えられるため $|\mathcal{N}_{\text{train}}|$ を増やして安定性について再度比較を行った．表 2 はその結果を示している．$|\mathcal{N}_{\text{train}}|$ を 0 から 50 まで増やしたが MLP の安定性が向上することはなかった．一方提案手法（Ref+PM）は最高で 99.9%の安定性（$|\mathcal{N}_{\text{train}}| = 10$）を達成している．つまり $|\mathcal{N}_{\text{test}}| = 1000$ 単語中，変換されてしまった単語はただ 1 件のみであった．一方でそのモデルは性別属性を持つ単語の変換精度が 37.5%であることから提案手法は，**変換対象の属性を持つ単語が入力された場合は属性を変換し，そうでない場合は変化させない**ということが実験的に示された．

5.3.2 安定性を持つ条件についての考察

これまでの実験で鏡映変換の安定性が予想以上に高く，目的の属性を持つ単語のみ変換させることが示された．これは目的の属性を持たない単語ベクトルが鏡映変換によってほとんど移動しないことを示している．以下は推察であるが，鏡 $M1$ で移動しないベクトルは $M1$ 上に存在する必要があるため，ある 2 つの鏡 $M1$ と $M2$ によって移動しないベクトルは $M1$ と $M2$ が交差する部分空間に存在していると考えられる．例えば $\mathbb{R}^2$ の低次元 Euclid 空間上の 2 つの鏡 $M1$ と $M2$（直線）によってベクトル $\mathbf{v}$ が変化しないためには $\mathbf{v}$ が $M1$ と $M2$ の交差する範囲に存在すればよい．ただし $\mathbb{R}^2$ においてはその範囲は点であるので全ての変換対象外のベクトルがそこに存在することは難しいが，実験で用いた高次元 Euclid 空間（$\mathbb{R}^{300}$）では鏡 $M1$ と $M2$（超平面）の交差する範囲は点ではなく $300-2$ 次元の部分空間 $\mathbb{R}^{298} \subset \mathbb{R}^{300}$ であるため多くのベクトルはその範囲に存在でき変換されなかったものと考えれる．

5.4 変換例

文 $X = \{x_1, x_2, ...\}$ を与え，一単語づつ変換関数に入力した結果を表 3 示す．ただし本研究で用いたモデルの語彙にない単語（“a”，“,”）は入力しない．MLP は全ての単語を変換してしまっているが，鏡映変換は性別属性を持つ単語のみ変換させている．例えば $Ref(\mathbf{v}_{father})$ は “mother” に変換されているが，$Ref(\mathbf{v}_{when})$ は変換されず “when” のままである．また二回鏡映変換を適用した結果もとの単語に戻っていることがわかる．例えば Ref に “father” を入力した場合 “mother” に変換され，“mother” を入力した場合 “father” に変換されている．これは入力 x が男性であるか女性であるかといった知識を用いずに x の性別を反転させていることを示している．

次に提案手法（Ref+PM）のエラー分析を行った．テストデータ $\mathcal{A}_{\text{test}}$ の属性変換結果を以下の 4 つに場合分けし分析した．

- **A:** 提案手法によって x から t に正しく変換されたケース（成功）
- **B:** x から変換されなかったケース
- **C:** 性別変換は成功しているケース
- **D:** その他

表 4 はエラー分析の結果である．性別属性の変換に失敗したもののうち，入力 x から変換されなかったケース（A）がもっとも多かった．少数であったが性別変換自体には成功していたケースも存在した（C）．またそれ以外のケース（D）は存在しなかった．

表 5 は鏡映変換で変換に失敗した例である．入力 x =“gentlemen” に対して性別反転した場合の正解は “ladies” であるが，提案手法で変換して得られたベクトルの最近傍は “Excellencies_ladies”，次の最近傍は “MODERATOR_Ladies” であった．予測された単語は “ladies” そのものではないが，男性から女性に変換できており，さらに “Excellencies_ladies” と “MODERATOR_Ladies” は “ladies” を具象化した表現である．また “king” は “queen” に変換されなかったが “princess” に変換され，“queen” は “prince” に変換された．両者とも性別の変換自体には成功している．さらに変換後も “king”, “queen” に共通する貴族の属性は保持されている．

予備実験として性別以外の属性を変換した結果を表 6 に示す．元文から性別変換，年齢変換，そして時制変換を順番に適用した．性別変換時は “mother” が “father” に変換されるが，年齢変換時は “father” が “grandfather” に変換されているなど，性別以外の属性も変換可能であることがわかる．

表 6 他の属性の変換結果

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

6. 関連研究

word2vec[2], [3] や GloVe[4] で得られた単語分散表現は線形類推関係が成り立つように学習されていないにもかかわらず，そのような特性を持つ．これまで単語分散表現がなぜ線形類推関係 [11], [12] を持つのかについて理論的な説明がなされてきた [6], [7], [8], [9], [10]．この中で [6] らは強い仮定のもと SGNS と sPMI（シフト自己相互情報量）が等価であることを限定的に証明した．これまでの研究では実際には成立しない仮定が導入されていたが，近年 [10] と [9] らはそれらの仮定を導入せずに証明したとしている．

本研究ではそのような線形類推関係に着目し，埋め込み空間上で分散表現の属性を反転させる新たな表現学習を提案する．類似した研究として文のスタイル変換が存在する [19], [20], [21], [22], [23]．このタスクでは文が与えられ教師なしでその文の言い回しなどの表現を変換する．例えば formal な文から informal な文への変換などを行った研究がある [21]．[22] らは文の属性（mood, tense）などを制御し変換することに成功した．この点は本研究でも同じであるが，これらのスタイル変換はニューラル機械翻訳 [24] のフレームワークを用いているので，ある文から別の文への変換に特化したアルゴリズムとなっている．本研究で提案する属性変換はある分散表現から同一の埋め込み空間上の分散表現への変換を行うタスクである．このような基礎的な変換技術は 8 節で述べるような応用の可能性がある．

7. 結論

本研究では単語埋め込み空間上で単語の属性を変換する新たな表現学習に取り組んだ．アナロジーによる変換では入力単語が持つ属性を知識として与える必要があったが，本研究ではそのような知識を用いることなく変換するために鏡映変換を導入した．実験の結果，提案手法は特定の属性を持つ単語であれば 45.8%の精度で変換し，その属性を持たない単語であれば 99.9%を変換せず，**知識を用いることなく目的属性を持つ単語のみ安定して変化させることに成功した**．

8. 今後の展望

本研究では埋め込み空間上で鏡映変換を行うことで分散表現の属性変換が可能であることを示した．しかしながら実験では性別属性の変換にとどまっているため，今後は他の属性についても同様の実験を行い効果を検証する．そして提案手法を用いて文のデータ拡張を行いその有用性を示す予定である．

また，今回提案した埋め込み空間上の属性変換は単語埋め込み空間以外にも適用可能である．高品質な画像の生成モデルとして有名な GAN[25] では，学習された潜在空間上のベクトルには SGNS 等と同様に線形類推関係があることが知られている（例：メガネをかけた男性 − メガネをかけていない男性 + メガネをかけていない女性 = メガネをかけた女性）[26]．したがって画像変換（Visual アナロジー）[27] やスタイル変換（Style transfer, Visual attribute transfer）[28], [29] などの潜在空間上の変換タスクに応用できる可能性がある．

他には単語単位ではなく文単位の属性変換などが考えられる [19], [20], [21].

また，鏡映変換は $\mathbf{v}_{man} = Ref(\ Ref(\mathbf{v}_{man})\ )$ のように二回適用するともとに戻るので記号論理における論理否定（$man = \neg\neg\, man$）と似た性質をもつ．応用としては埋め込み空間上で任意の分散表現の否定表現の推論を考えている．例えばアナロジーが成り立つ知識グラフ埋め込み空間 [30] で，ある Entity のある Relation を否定した表現を予測するなど，様々な応用が考えられる．

謝辞 本研究は JST CREST（課題番号: JPMJCR1513）の支援を受けて行った．

参考文献

[1] Hinton, G. E., McClelland, J. L., Rumelhart, D. E. et al.: *Distributed representations*, Carnegie-Mellon University Pittsburgh, PA (1984).
[2] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013).
[4] Pennington, J., Socher, R. and Manning, C.: Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (2014).
[5] Church, K. W. and Hanks, P.: Word association norms, mutual information, and lexicography, *Computational linguistics*, Vol. 16, No. 1, pp. 22–29 (1990).
[6] Levy, O. and Goldberg, Y.: Neural word embedding as implicit matrix factorization, *Advances in neural information processing systems*, pp. 2177–2185 (2014).
[7] Arora, S., Li, Y., Liang, Y., Ma, T. and Risteski, A.: A latent variable model approach to pmi-based word em-

beddings, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 385–399 (2016).

[8] Gittens, A., Achlioptas, D. and Mahoney, M. W.: Skip-Gram- Zipf+ Uniform= Vector Additivity, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76 (2017).

[9] Ethayarajh, K., Duvenaud, D. and Hirst, G.: Towards Understanding Linear Word Analogies, *CoRR*, Vol. abs/1810.04882 (online), available from ⟨http://arxiv.org/abs/1810.04882⟩ (2018).

[10] Allen, C. and Hospedales, T. M.: Analogies Explained: Towards Understanding Word Embeddings, *CoRR*, Vol. abs/1901.09813 (online), available from ⟨http://arxiv.org/abs/1901.09813⟩ (2019).

[11] Levy, O. and Goldberg, Y.: Linguistic regularities in sparse and explicit word representations, *Proceedings of the eighteenth conference on computational natural language learning*, pp. 171–180 (2014).

[12] Mikolov, T., Yih, W.-t. and Zweig, G.: Linguistic regularities in continuous space word representations, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751 (2013).

[13] Linzen, T.: Issues in evaluating semantic spaces using word analogies, *CoRR*, Vol. abs/1606.07736 (online), available from ⟨http://arxiv.org/abs/1606.07736⟩ (2016).

[14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[15] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*, Vol. 15, No. 1, pp. 1929–1958 (2014).

[16] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167* (2015).

[17] Kaneko, M. and Bollegala, D.: Gender-preserving Debiasing for Pre-trained Word Embeddings, *arXiv preprint arXiv:1906.00742* (2019).

[18] Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W.: Learning gender-neutral word embeddings, *arXiv preprint arXiv:1809.01496* (2018).

[19] Niu, X., Rao, S. and Carpuat, M.: Multi-Task Neural Models for Translating Between Styles Within and Across Languages, *CoRR*, Vol. abs/1806.04357 (online), available from ⟨http://arxiv.org/abs/1806.04357⟩ (2018).

[20] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R. and Black, A. W.: Style Transfer Through Back-Translation, *CoRR*, Vol. abs/1804.09000 (online), available from ⟨http://arxiv.org/abs/1804.09000⟩ (2018).

[21] Jain, P., Mishra, A., Azad, A. P. and Sankaranarayanan, K.: Unsupervised Controllable Text Formalization, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6554–6561 (2019).

[22] Logeswaran, L., Lee, H. and Bengio, S.: Content preserving text generation with attribute controls, *CoRR*, Vol. abs/1811.01135 (online), available from ⟨http://arxiv.org/abs/1811.01135⟩ (2018).

[23] Dai, N., Liang, J., Qiu, X. and Huang, X.: Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation, *arXiv preprint arXiv:1905.05621* (2019).

[24] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to sequence learning with neural networks, *Advances in neural information processing systems*, pp. 3104–3112 (2014).

[25] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in neural information processing systems*, pp. 2672–2680 (2014).

[26] Radford, A., Metz, L. and Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks, *arXiv preprint arXiv:1511.06434* (2015).

[27] Reed, S. E., Zhang, Y., Zhang, Y. and Lee, H.: Deep visual analogy-making, *Advances in neural information processing systems*, pp. 1252–1260 (2015).

[28] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232 (2017).

[29] Liao, J., Yao, Y., Yuan, L., Hua, G. and Kang, S. B.: Visual Attribute Transfer through Deep Image Analogy, *CoRR*, Vol. abs/1705.01088 (online), available from ⟨http://arxiv.org/abs/1705.01088⟩ (2017).

[30] Liu, H., Wu, Y. and Yang, Y.: Analogical inference for multi-relational embeddings, *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, pp. 2168–2178 (2017).