3J3-4

複合非負値行列因子分解 (NM2F) による 絵本データセットからの多角的パターン抽出

Analyzing Picture Book Data with Non-negative Multiple Matrix Factorization: Extended Abstract

竹内 孝*1 Koh Takeuchi　石黒 勝彦*1 Katsuhiko Ishiguro　小林 哲生*1 Tessei Kobayashi　藤田 早苗*1 Sanae Fujita　平 博順*1 Hirotoshi Taira

*1NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

In this paper, we tackle a problem to extract patterns from a data collection of picture books. The problem is formulated as a multiple matrix factorization, and we solve it by NM2F and NMF algorithm. We examined the perfomance of NM2F and NMF by both quantitative and qualitative ways. We confirmed that NM2F reduces errors and extracts reasonable patterns from the picture book data set.

1. はじめに

絵本は一般の書籍と比較して，1 冊あたりに出現する文字数が少なく, 使用される語彙も少ないが，対象年齢に合わせて使用される語彙や表現が変化していくなどのユニークな特徴を持つ．絵本データを解析し, 対象年齢毎の絵本や単語の潜在的なクラスタ構造が抽出できれば，絵本の推薦などの応用に役立てられる．しかし, 絵本の文字列データを用いて潜在構造の抽出を行おうとすると，絵本に含まれる文字数が少なくデータがスパースになりやすいため解析が困難となる．そこで対象年齢や単語品詞情報といった, 絵本データに含まれる「文字以外」のデータを利用することで，データのスパース性を減少させるなどの工夫が必要になる．

複合非負値行列因子分解 (NM2F: Non-negative Multiple Matrix Factorizaion) [7] はインデクスの対応が取れる複数の行列形式データを同時分解する技術である．NM2F は, 複数の行列間に共通の因子を仮定することで欠損値の多いスパースな行列を分解する際に欠損値推定の精度が向上する性質を持ち，音楽視聴履歴やソーシャルメディアのデータなどスパースなデータの解析に適応され良好な性能を示している [7].

本研究の目的は, 絵本データセットを単語データ, 対象年齢データ, 品詞データの複数行列としてとらえ, NM2F による同時分解によって潜在構造を抽出することである．実験から, NM2F によって複数行列を同時に分解することで，非負値行列因子分解法 (NMF: Non-negative Matrix Factorizaion) [3] によって行列を個別に分解する場合よりも，行列の欠損値推定精度が向上することを定量的に示す．また, NM2F によって抽出された潜在構造が多角的で理解しやすいことを定性的に示す.

2. 絵本データと統計値

本研究で用いる絵本データセットは以下の通りである．絵本の総数は $J = 1,200$ 冊である．絵本の文章に文献 [9] の形態素解析器を用いて，ひらがなの形態素解析と同時に漢字表記による単語の基本形 (以下, 単語) を抽出する．データセットは $I = 21,507$ 種類の語彙と $M = 32$ 種類の品詞からなる．品詞体系は IPA 品詞体系による．単語の出現頻度の分布を図 1，絵本毎の総文字数の分布を図 2 に示す．それぞれの分布に冪乗則をフィッティングしたものを図中に赤線で示す．図より単語出現頻度, 絵本の総文字数は急速に小さくなることがわかる．絵本の総文字数については総文字数が 100 文字程度の絵本が多いことから，データは非常にスパースで有ることが分かる．また品詞毎の出現回数の分布を図 3 に示す．品詞には名詞-一般や動詞-自立が多く含まれているが, 各品詞の出現頻度は一定数以上存在していることが分かる.

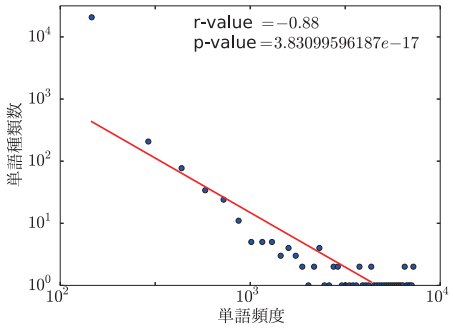

図 1　単語出現頻度分布. 横軸:単語出現頻度, 縦軸:単語種類, 赤線:冪乗則のフィッティング結果

次に絵本の対象年齢に関する統計値を図示する．絵本 1,200 冊のうち 653 冊には対象年齢の記述があり, 残りの 547 冊には対象年齢の記述が無い．絵本の対象年齢のうち 0 歳から 4 歳までを利用し $N = 5$ とする．対象年齢の記述のある絵本における対象年齢の分布を図 4 に示す．データセットには 1 歳から 3 歳を対象年齢とする絵本が多く含まれるが，各年齢を対象とする絵本が一定の数存在していることが分かる.

3. 複合非負値行列因子分解

複合非負値行列因子分解 (NM2F: Non-negative Multiple Matrix Factorizaion) とは，インデクス対応が取れる複数の非負行列を非負の因子行列に同時分解する手法である [7]．NM2F のキーアイディアは同一のインデクスに対して同一の因子行列を仮定することである．NM2F は, 1 つの観測データと 2 つの補助データを扱う*1．主観測データは I 次元からなり, 総デー

連絡先: NTT コミュニケーション科学基礎研究所, 619-0037, 京都府相楽郡精華町光台 2-4, koh.takeuchi@lab.ne.jp

*1　本稿では, ベクトルを小文字の太字, 行列を大文字の太字, 転置を T で表す.

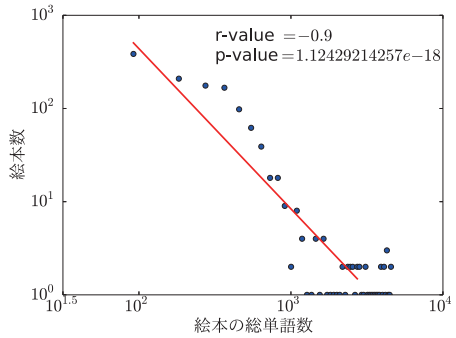

図 2　絵本の総単語分布. 横軸:絵本の総単語数, 縦軸:絵本数, 赤線:冪乗則のフィッティング結果

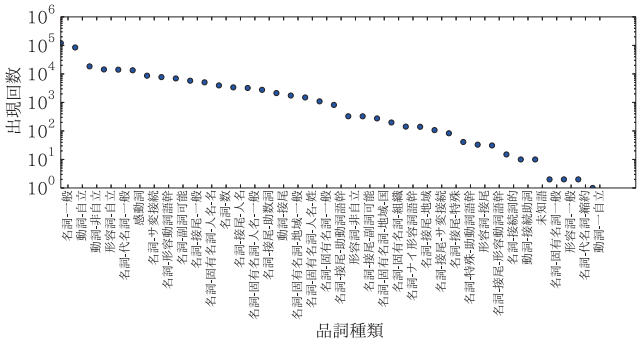

図 3　品詞分布. 横軸:品詞種類, 縦軸:出現頻度

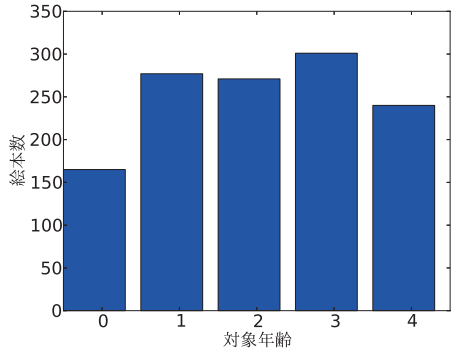

図 4　対象年齢の分布. 横軸:対象年齢, 縦軸:出現頻度

タ点数を J とする． j 点目のデータにおける次元 i の観測値を $x_{i,j}$ とおき，観測データは $\boldsymbol{X} = [x_{i,j}] \in \mathbb{R}_+^{I\times J}$ と定める． 1 つ目の補助データは, $\boldsymbol{X}$ の j 点目のデータに対応して観測される． 補助データの次元を N とする． j 点目の補助データにおける次元 n の観測値を $y_{n,j}$ とおき，補助データは補助行列 $\boldsymbol{Y}$ として，$\boldsymbol{Y} = [y_{n,j}] \in \mathbb{R}_+^{N\times J}$ と定める． 2 つ目の補助データは, $\boldsymbol{X}$ の i 次元に対応して観測される．補助データの点数を M とする． m 点目の補助データにおける次元 i の観測値を $z_{i,m}$ とし，補助データは補助行列 $\boldsymbol{Z}$ として，$\boldsymbol{Z} = [z_{i,m}] \in \mathbb{R}_+^{I\times M}$ と定める. 行列 $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ の関係を図 5 の左側に示す.

NM2F の目的は, 観測行列 $\boldsymbol{X}$ からの基底行列 $\boldsymbol{W}$ と係数行列 $\boldsymbol{H}$ の推定である． NM2F は補助行列 $\boldsymbol{Y}$ と $\boldsymbol{Z}$ を利用するため, 既存手法と比較して汎化性能の向上が期待できる. さらに補助行列からも基底行列と係数行列を抽出するため, 補助行列に潜在するパターンも同時に抽出出来る利点を持つと期待できる.

観測行列と補助行列の具体例を説明する． 絵本データセットでは，観測行列 $\boldsymbol{X}$ は単語 i が絵本 j に出現した頻度から計算し

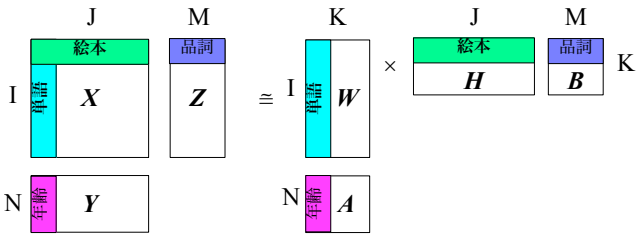

図 5　観測行列 $\boldsymbol{X}$ を基底行列 $\boldsymbol{W}$ と係数行列 $\boldsymbol{H}$ に分解する. その際, 補助行列 $\boldsymbol{Y}$ と $\boldsymbol{Z}$ を同時に分解する.

た tf-idf 特徴量， 補助行列 $\boldsymbol{Y}$ は年齢 n が絵本 j で対象とされているかの特徴量， 補助行列 $\boldsymbol{Z}$ は単語 i が品詞 m として使用された頻度である.

NM2F は, 分解後の行列が全て非負となる制約の下で観測行列 $\boldsymbol{X}$ と補助行列 $\boldsymbol{Y}, \boldsymbol{Z}$ を分解する． 分解から得られる基底行列と係数行列は, 非負制約によってスパースになり, さらに直感的に理解しやすいものとなると報告されている． 観測行列 $\boldsymbol{X}$ が高度にスパースで行列分解が困難な際に補助行列を用いることで, 利用できる情報量を増やし， NM2F は観測行列を分解するために 2 つの補助行列を利用することで, より性能の高い基底行列と係数行列の推定が可能になる.

観測行列は K 個の基底を持つと仮定する． 観測行列 $\boldsymbol{X}$ の k 番目の基底ベクトルを $\boldsymbol{w}_k = (w_{1,k}, \cdots, w_{I,k})^{\mathrm{T}} \in \mathbb{R}_+^I$ とし，基底行列を $\boldsymbol{W} = [w_{i,k}] \in \mathbb{R}_+^{I\times K}$ とおく． 行方向の補助行列の k 番目の基底ベクトルを $\boldsymbol{a}_k = (a_{1,k}, \cdots, a_{N,k})^{\mathrm{T}} \in \mathbb{R}_+^N$ とし，補助行列 $\boldsymbol{Y}$ の基底行列を $\boldsymbol{A} = [a_{n,k}] \in \mathbb{R}_+^{N\times K}$ とおく．次に観測行列 $\boldsymbol{X}$ の j 点目のデータに対応する係数ベクトルを $\boldsymbol{h}_j = (h_{1,j}, \cdots, h_{K,j})^{\mathrm{T}} \in \mathbb{R}_+^K$ とし，係数行列を $\boldsymbol{H} = [h_{k,j}] \in \mathbb{R}_+^{K\times J}$ と定める． 列方向の補助行列を $\boldsymbol{Z}$ の m 点目のデータに対応する係数ベクトル $\boldsymbol{b}_m = (b_{1,m}, \cdots, h_{K,m})^{\mathrm{T}} \in \mathbb{R}_+^K$ とし，係数行列を $\boldsymbol{B} = [b_{i,m}] \in \mathbb{R}_+^{K\times M}$ と定める.

このとき, 観測値 $x_{i,j}, y_{n,j}, z_{i,m}$ の再構成値 $\hat{x}_{i,j}, \hat{y}_{n,j}, \hat{z}_{i,m}$ は, 基底と係数の重み付き線形和と定める.

$$\hat{x}_{i,j} = \boldsymbol{w}_i{}^{\mathrm{T}}\boldsymbol{h}_j,\ \hat{y}_{n,j} = \boldsymbol{a}_n{}^{\mathrm{T}}\boldsymbol{h}_j,\ \hat{z}_{i,m} = \boldsymbol{w}_i{}^{\mathrm{T}}\boldsymbol{b}_m \tag{1}$$

.

NM2F は, 行列 $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ を行列 $\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{H}, \boldsymbol{B}$ で再構成した際の再構成誤差 $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta)$ を最小化する.

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{A},\boldsymbol{B}} \mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta)$$
$$\text{subject to } \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B} \geq 0. \tag{2}$$

このとき, α, β は $\alpha \geq 0, \beta \geq 0$ を満たすスケーリング変数であり， 再構成誤差 $\mathcal{D}$ は, 行列の要素 $x_{i,j}$ とその再構成値 $\hat{x}_{i,j}$ との間の誤差 $d(x_{i,j}|\hat{x}_{i,j})$ を用いて以下のように定める.

$$\begin{aligned}
&\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta) \\
&= \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}) + \alpha\mathcal{D}(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{H}) + \beta\mathcal{D}(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{B}) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} d(x_{i,j}|\hat{x}_{i,j}) + \alpha\sum_{n=1}^{N}\sum_{j=1}^{J} d(y_{n,j}|\hat{y}_{k,j}) \\
&+ \beta\sum_{i=1}^{I}\sum_{m=1}^{M} d(z_{i,m}|\hat{z}_{i,m}).
\end{aligned} \tag{3}$$

本稿では再構成誤差 D に一般化 KL ダイバージェンス (generalized Kullback-Leibler divergence: gKL) を用いる． その

際, 行列の要素 $x_{i,j}$ とその再構成値 $\hat{x}_{i,j}$ の間の誤差を,

$$d_{\mathrm{gKL}}(x_{i,j}|\hat{x}_{i,j}) = x_{i,j}\log\frac{x_{i,j}}{\hat{x}_{i,j}} - x_{i,j} + \hat{x}_{i,j}, \tag{4}$$

とする. なお, このとき再構成誤差 D の最小化は観測分布にポアソン分布を仮定した際の最尤推定と一致する [7].

次の乗法的更新則によって, 再構成誤差 $\mathcal{D}$ を最小化する最適な $\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}$ を求められる.

$$\begin{aligned}
w_{i,k}^{\mathrm{new}} &= w_{i,k}\frac{\left(\sum_{j\in\mathcal{J}_i}\frac{x_{i,j}}{\hat{x}_{i,j}}h_{k,j} + \beta\sum_{m\in\mathcal{M}_i}\frac{z_{k,m}}{\hat{z}_{k,m}}b_{i,m}\right)}{\sum_j h_{k,j} + \beta\sum_m b_{k,m}},\\
h_{k,j}^{\mathrm{new}} &= h_{k,j}\frac{\left(\sum_{i\in\mathcal{I}_j}\frac{x_{i,j}}{\hat{x}_{i,j}}w_{k,j} + \alpha\sum_{n\in\mathcal{N}_j}\frac{y_{n,j}}{\hat{y}_{n,j}}a_{n,k}\right)}{\sum_i w_{k,j} + \alpha\sum_n a_{n,k}},\\
a_{n,k}^{\mathrm{new}} &= a_{n,k}\frac{\sum_{n\in\mathcal{N}_j}\frac{y_{n,j}}{\hat{y}_{n,j}}h_{k,j}}{\sum_j h_{k,j}},\\
b_{k,m}^{\mathrm{new}} &= b_{k,m}\frac{\sum_{m\in\mathcal{M}_i}\frac{z_{i,m}}{\hat{z}_{i,m}}w_{i,k}}{\sum_i w_{i,k}}.
\end{aligned} \tag{5}$$

4. 実験

本研究では各行列の欠損値推定問題を扱い, 欠損値の推定精度によって各行列の潜在構造の学習性能を比較する.

4.1 絵本データの定式化

単語頻度から Term-Frequency inverse Document Frequency (tf-idf) を計算し特徴量として利用する. 単語 i の絵本 j における tf-idf を要素に持つ, 観測行列を $[x_{i,j}] = \boldsymbol{X} \in \mathbb{R}_{+}^{21,507\times 1200}$ とする. 次に絵本の対象年齢は any-of-N 表現を用いて次のベクトルとして表す. 5 次元のベクトル $\boldsymbol{y} \in [0,1]^5$ を定め, ある絵本 j が n 歳を対象年齢としている場合は $y_{n,j} = 1$ とし, 対象年齢としていない場合は $y_{n,j} = 0$ と定め, 補助行列を $[y_{n,j}] = \boldsymbol{Y} \in \mathbb{R}_{+}^{32\times 1200}$ とする. 最後に単語 i の品詞 m の使用頻度を要素に持つ, 単語 × 品詞行列を $[z_{i,m}] = \boldsymbol{Z} \in \mathbb{R}_{+}^{21,507\times 32}$ とする.

4.2 評価指標

実験では, 与えられたデータセットをトレーニングセットとテストセットに分割した後, トレーニングセットに対する再構成誤差 D を最小化するようモデルのパラメータの学習を行う. モデルの汎化性能を比較するために, テストセットの非ゼロ要素に対する一般化 KL ダイバージェンスを用いる. テストセットに対する一般化 KL ダイバージェンスが小さいモデルほど, データの潜在的な構造を捉えた良いモデルといえる. 一般化 KL ダイバージェンス f を次のように定める: $f(\boldsymbol{X}|\theta) = \frac{1}{R}\sum_{r=1}^{R} d_{\mathrm{gKL}}(x_r, \hat{x}_r)$. なお, テストセットのデータ点数を R とし, θ はモデルの推定したパラメータである. 実験では, 観測行列 $\boldsymbol{X}$ の非零観測要素を 5 つのセットに分割し, 5 交差検定法によってスケーリングパラメータの推定を行った後, テストセットの一般化 KL ダイバージェンスの平均値と標準偏差を求める.

5. 実験結果

実験結果の定量評価を行う. 行列 $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ を NMF によって個別に分解した際の場合と, NM2F によって同時に分解した場合の一般化 KL ダイバージェンスの平均値と標準偏差を表 1 に示す. NMF によって選択された最適な因子数 K は, $\boldsymbol{X}$ で $K = 2$, $\boldsymbol{Y}$ で $K = 2$, $\boldsymbol{Z}$ で $K = 2$ となった. NM2F によって選択された最適な因子数は $K = 30$, スケーリング変数は $\alpha = 10, \beta = 100$ となった. NM2F は NMF と比較して, 観測行列 $\boldsymbol{X}$ の一般化 KL ダイバージェンスを優位に減少させており, 観測行列 $\boldsymbol{X}$ の潜在構造を良く学習しているといえる. また NM2F の選択した因子数は, NMF の選択した因子数よりも増えており, 補助行列によって観測行列 $\boldsymbol{X}$ の複雑な潜在構造が抽出出来たといえる. 一方, 補助行列 $\boldsymbol{Y}$, $\boldsymbol{Z}$ の一般化 KL ダイバージェンスは NMF よりも大きくなっており, 行列 $\boldsymbol{Y}$, $\boldsymbol{Z}$ の潜在構造の学習は悪化するトレードオフが生じている. これは共通の潜在因子行列を仮定するモデルに共通する挙動である.

行列	NMF	NM2F
$\boldsymbol{X}$: 単語 × 絵本	0.186 ± 0.002	$\mathbf{0.170 \pm 0.004}$
$\boldsymbol{Y}$: 対象年齢 × 絵本	$\mathbf{22.56 \pm 1.59}$	34.02 ± 0.87
$\boldsymbol{Z}$: 単語 × 品詞	$\mathbf{215.50 \pm 3.58}$	372.62 ± 47.44

表 1 一般化 KL ダイバージェンスの平均値と標準偏差. 一般化 KL ダイバージェンスが小さいほど良いモデルといえる.

次に実験結果の定性評価を行う. NM2F によって学習された 30 個の潜在因子のうち, 特徴的な物を図 6, 7, 8 に示す. 図の左から単語, 絵本タイトル, 対象年齢, 品詞の因子ベクトルに対応する. 図には各因子ベクトルで最も高い値を取った上位の要素が示されている. 図 6 で表される因子では, “さあ”, “ああ”, “はい” などの単語と, タイトルに感動詞を持つ絵本が高い値を持つ, 対象年齢と品詞の因子ベクトルでは 1 歳と感動詞が高い値を持つ. 絵本では感動詞が頻繁に出現するが, 1 歳程度を対象にしたものに, その傾向が高いことが分かる. 一方, 同じ 1 歳向けではあるが, 図 7 で表される因子では, “いい”, “大きい”, “おいしい” などの形容詞と, “たのしいいちにち”, “くらいくらい” などの絵本が高い値を持つ, 1 歳を対象とする形容詞が頻繁に出現する因子が抽出されたといえる. 絵本のタイトルにも形容詞が含まれるものが多く, 絵本の内容をよく捉えた因子が抽出されたと考えられる. 最後に図 8 で表される因子は, “そう”, “さま”, “ぐら”, “ジャッキー” などの多様な単語と, “とこやにったライオン”, “ふうせんねこ” などの絵本が高い値を持つ. 前述の因子と異なり対象年齢は 3 歳が高い値となり, 品詞は固有名詞が高い値を持つ. 絵本では対象年齢が上がるに連れて登場キャラクターを示す固有名詞の使用される頻度が高くなると考えられる.

6. 関連研究

非負値行列因子分解 (Non-negative Matrix Factorization: NMF) [3] は教師無しの行列分解で, 非負制約下 [2] で観測データを基底の重み付き線形和で低ランク近似し再構成誤差を最小化する. NMF は, コスト関数が非凸となる問題を持つが, 基底と重みが非負値のため解釈がしやすくさらに非負制約によってスパースな解を得られるという好ましい特徴を持つ. NM2F [7] は, 複数の非負値行列を同時分解する手法で, スパースな行列の分解に優れた性能を持ち, さらに非負制約によって解釈しやすい因子行列を抽出すると報告されている. NM2F のテンソルへの拡張も提案されている [8]. スパースな行列の欠損値推定問題では, 確率的行列分解 (Probabilistic Matrix Factorization: PMF) や Collective Matrix Facotorization[5, 6] や, その拡張 [1, 4] などが提案されている. 本稿

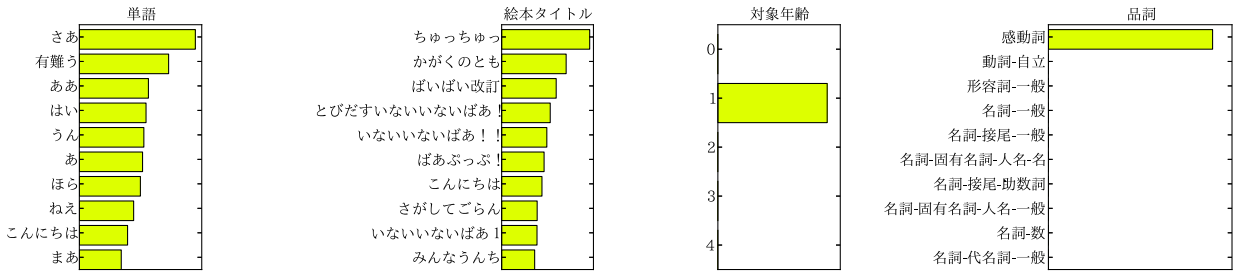

図 6　感動詞トピック

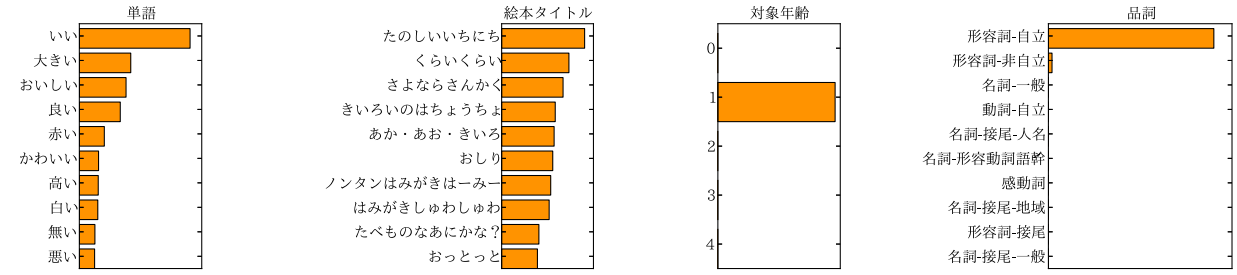

図 7　1 歳児向け形容詞トピック

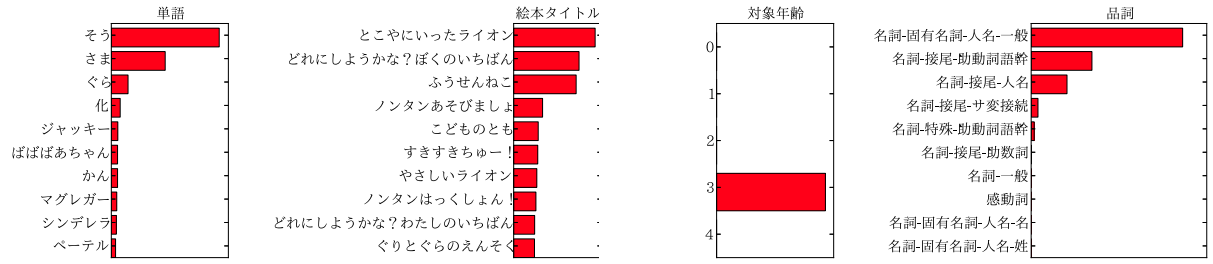

図 8　3 歳児向け名詞トピック

では, 絵本データからの理解しやすい潜在構造の抽出を目的とするため, NMF と NM2F を用いる.

7.　まとめ

本研究では複合非負値行列因子分解による絵本データからの絵本と単語の潜在構造解析を行った.　この際, 絵本の対象年齢データと単語の品詞データを補助データとして用いることで,　絵本と単語の定量的かつ定性的な精度の向上を確認した.　今後は, 絵本特有の文章構造や画像データの統計量など, より多種類のデータを用いたデータ解析による更なる精度向上が期待される.

参考文献

[1] Guillaume Bouchard, Shengbo Guo, and Dawei Yin. Convex collective matrix factorization. In *Proc. AISTATS*, 2013.

[2] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* Wiley, 2009.

[3] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[4] Mrinmaya Sachan and Shashank Srivastava. Collective matrix factorization for co-clustering. In *Proc. WWW*, 2013.

[5] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proc. SIGKDD*, 2008.

[6] Ajit P Singh and Geoffrey J Gordon. A unified view of matrix factorization models. In *Proc. ECMLPKDD*. 2008.

[7] Koh Takeuchi, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. Non-negative multiple matrix factorization. In *Proc. IJCAI*, 2013.

[8] Koh Takeuchi, Ryota Tomioka, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. Non-negative multiple tensor factorization. In *Proc. ICDM*, 2013.

[9] 藤田 早苗, 平 博順, 小林 哲生, and 田中 貴秋. 絵本のテキストを対象とした形態素解析. **自然言語処理**, 21(3), 2014.