線形判別分析の PU 学習による朝日歌壇短歌の分析

加藤真大 [1]　浦川通 [2]　田口雄哉 [2]　新妻巧朗 [2]　田森秀明 [2]　羽根田賢和 [4]　持橋大地 [3]
[1] 東京大学総合文化研究科　[2] 朝日新聞社　[3] 統計数理研究所　[4] 東北大学
mkato.csecon@gmail.com　{urakawa-t,taguchi-y2,niitsuma-t,tamori-h}@asahi.com
haneda.kento.t6@dc.tohoku.ac.jp　daichi@ism.ac.jp

概要

本研究では，朝日歌壇に掲載されている短歌の特徴を，Fisher の線形判別分析を用いて調査する．どのような短歌が朝日歌壇に掲載されているのかを調査するために，比較の対象として生成モデルによって作成された短歌を用意する．生成短歌には，もし朝日歌壇に投稿されていたら掲載されるような短歌 (正例) から，掲載されないような短歌 (負例) まで，多様な短歌が幅広く含まれている．こうした朝日歌壇短歌と生成短歌に対して，本研究では従来の線形判別分析を PU 学習の枠組みに拡張した手法を提案し，朝日歌壇短歌を正例データ，生成短歌を正例と負例が混在するラベルなしデータとみなして，これらが混在する PU 学習の枠組みで分析を行った。

1　はじめに

朝日新聞には，一世紀を超える短歌の投稿欄である「歌壇」があり，毎週，読者から送られる短歌のなかから選ばれた作品を掲載している．本稿では，生成モデルにより生成された短歌を Fisher の線形判別分析（Linear Discriminant Analysis, LDA）を通じて比較することにより，朝日歌壇短歌の特徴を分析する．これには選に洩れた短歌との比較が必要になるが、投稿された歌のうち選ばれなかった短歌は非公開である。よって短歌言語モデルから生成した歌との比較を行うが、その中には本来は朝日歌壇に載ってもよい歌も含まれていると考えられる。そこで、本研究では正例およびラベルなしデータから識別学習を行う PU 学習 [1] の枠組みを LDA に拡張し、朝日歌壇および選者がどのような歌を選んでいるか、という軸を定量的に明らかにすることを試みた。

2　使用する短歌データ

短歌の選定は，歌壇を代表する選者たちによって行われており，同じ短歌が複数の選者に選ばれる場合もある．各選者はそれぞれ一年間でおおよそ 450 首程度を選んでいる．本研究では，選者のうち永田和宏氏と馬場あき子氏によって 2006 年以降に選ばれた短歌に着目し，ひらがなとカタカナが 90%を占める短歌を評価から除外した 23,743 首（永田氏は 9,454 首，馬場氏は 14,432 首）の短歌を用いた．これらの短歌は各年ごとに出版されている書籍「朝日歌壇」から引用しており，各短歌には該当年の書籍が明記されている．

これら朝日歌壇の短歌と比較するために，生成モデルによって作成された短歌を用意した．生成短歌は [2] のモデルを用いて生成する．この生成モデルから生成された 10,000 首のうち，ひらがなとカタカナが 90%を占めているものと，短歌としての体裁をなしていないものを除いた 9,694 首の短歌を分析に用いた．

3　短歌の文埋め込み

短歌を定量的に分析するために，文埋め込みの手法を用いる．文埋め込みの手法として，[3] が提案している Smoothed Inverse Frequency （SIF）が挙げられる．SIF では未知のパラメータをハイパーパラメータとして設定する必要があるが，その設定もアルゴリズムに含めた unsupervised SIF (uSIF)[4] が提案されている．本稿では，この uSIF を用いて文埋め込みを実行する．

私たちが用いる uSIF では，文中の各単語を単語埋め込みベクトルに変換し，その単語を加重平均することで文埋め込みベクトルを得る．その単語埋め込みベクトルとして，私たちは朝日新聞単語ベクトルを用いる（詳細は付録 B）.

4　朝日歌壇短歌と生成短歌の LDA

本研究では，uSIF による文埋め込みベクトルを用いて，どの短歌が選者によって選ばれたものか，あるいは生成されたものであるかという分類問題を

LDA を用いて解く．これにより，各選者が選ぶ短歌と生成短歌の特徴の違いを調査する．

まず，分類問題を定式化する．短歌 i の文埋め込みベクトルを $\boldsymbol{x}_i$ とする．クラス k で条件づけられた埋め込みベクトルの分布の密度関数を $p(\boldsymbol{x}|k)$ とし，周辺化された密度を $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|k)$ とする．ここで，π_k はクラス事前分布（class prior）と呼ばれるスカラー値であり，$\sum_{k=1}^{K} \pi_k = 1$ かつ $\pi_k \geq 0$ を満たすとする．このクラス事前分布は既知とする．

選者 k が選んだ短歌の集合を $\mathcal{C}_k$ とし，以下の K 種類のデータセットを定義する：

$$\mathcal{C}_{k'} \coloneqq \{\boldsymbol{x}_i^{(k')}\}_{i=1}^{n_{k'}},\ \boldsymbol{x}_i^{(k')} \sim p(\boldsymbol{x}|k=k'),\ k' = 1, \ldots, K-1,$$
$$\mathcal{C}_U \coloneqq \{\boldsymbol{x}_i^{(U)}\}_{i=1}^{n_U},\ \boldsymbol{x}_i^{(U)} \sim p(\boldsymbol{x}),$$

ここで，クラス 1 からクラス $K-1$ のデータセットは正例データ，クラス 1 からクラス K のデータが混ざったデータセット $\mathcal{C}_U$ はラベルなしデータと呼ばれる．本研究では，正例データは選者によって選ばれて朝日歌壇に掲載された短歌，ラベルなしデータは生成された短歌に相当する．

4.1 LDA の概要

ここでは，PU 的な構造を考慮しない基本的な LDA について説明する [5]．LDA では，線形分類器 $f(\boldsymbol{x}) \coloneqq \arg\max_{k \in [K]} g(k|\boldsymbol{x})$ を考える．ここで，$g\colon [K] \times \mathcal{X} \to \mathbb{R}$ はスコア関数である．LDA では，線形なスコア関数 $g(x) \coloneqq \boldsymbol{W}^\top \boldsymbol{x}$ を用いて分類を行う．

このパラメータ $\boldsymbol{W}$ を求めるために，クラス内共分散行列 S_W とクラス間共分散行列 S_B を定義する．短歌全体の $\boldsymbol{x}_i$ の平均ベクトルを $\boldsymbol{m}$，ある選者 k が選んだ短歌全体の $\boldsymbol{x}_i$ の平均ベクトルを $\boldsymbol{m}_k \coloneqq \mathbb{E}_k[\boldsymbol{x}]$ とする．このとき，S_W と S_B は，$S_W \coloneqq \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in C_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)(\boldsymbol{x}_i - \boldsymbol{m}_k)^\top$ と $S_B \coloneqq \frac{1}{n}\sum_{k=1}^{K} |\mathcal{C}_k| (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top$ と定義される．ここで，K は選者数（本研究では 3）である．また，$K = 2$ のときは，$S_B \coloneqq (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$ とする．

LDA では，データを低次元に射影する行列 $\boldsymbol{W}$ を，射影後の空間においてクラス内の分散を最小化し，クラス間の分散を最大化するように求める．すなわち，$J(\boldsymbol{W}) \coloneqq \mathrm{tr}\left((\boldsymbol{W}^\top S_W \boldsymbol{W})^{-1}(\boldsymbol{W}^\top S_B \boldsymbol{W})\right)$ を最大化するように $\boldsymbol{W}$ を求める．

4.2 2 クラス分類における PU-LDA

LDA を PU 学習の設定に拡張する．最初に，2 クラス分類問題において，正例データとラベルなしデータが与えられている状況での LDA について考える．すなわち，$\mathcal{C}_1 = \{\boldsymbol{x}_i^{(1)}\}_{i=1}^{n_1}$ と $\mathcal{C}_U = \{\boldsymbol{x}_i^{(U)}\}_{i=1}^{n_U}$ の二つのデータセットを観測できる．ここで，データセット $\mathcal{C}_1$ は正例データ，クラス 1 とクラス 2 が混ざった $\mathcal{C}_U$ はラベルなしデータである．

このとき，平均 $\boldsymbol{m}_1 = \int \boldsymbol{x} p(\boldsymbol{x}|k=1)\mathrm{d}\boldsymbol{x}$ については，正例データのサンプル平均によって近似できる．課題は $\boldsymbol{m}_2$ を計算することである．通常の LDA と異なり，クラス 2 のデータセット $\mathcal{C}_2$ を観測できず，クラス 1 とクラス 2 が混ざったラベルなしデータセット $\mathcal{C}_U$ しか観測できない．本研究では，この $\boldsymbol{m}_2$ を正例データとラベルなしデータから識別・推定する方法を考える．

PU 学習でしばしば用いられる変換を通じて，$\boldsymbol{m}_2$ はラベルなしデータの平均から正例データの平均を引くことで計算できる．すなわち，

$$\boldsymbol{m}_2 = \left(\int \boldsymbol{x} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi_1 \int \boldsymbol{x} p(\boldsymbol{x}|k=1)\mathrm{d}\boldsymbol{x}\right)/\pi_2$$

として識別できる．なぜなら，$p(\boldsymbol{x}) = \pi_1 p(\boldsymbol{x}|k=1) + \pi_2 p(\boldsymbol{x}|k=2)$ であるからである．ここで，π_1 はクラス事前分布であり，本研究では既知とする．

この $\boldsymbol{m}_2$ をサンプルで置き換えたものを $\widehat{\boldsymbol{m}}_2^{\mathrm{PU}}$ とする．すなわち，

$$\widehat{\boldsymbol{m}}_2^{\mathrm{PU}} \coloneqq \left(\frac{1}{n_U}\sum_{i \in \mathcal{C}_U} \boldsymbol{x}_i^{(U)} - \pi_1 \frac{1}{n_1}\sum_{i \in \mathcal{C}_1} \boldsymbol{x}_i^{(1)}\right)/\pi_2$$

である．この $\widehat{\boldsymbol{m}}_2^{\mathrm{PU}}$ を従来の LDA における $\widehat{\boldsymbol{m}}_2$ の代わりに用いることで，クラス間距離 $\widehat{\boldsymbol{m}}_1 - \widehat{\boldsymbol{m}}_2^{\mathrm{PU}}$ を計算できる．また，およびクラス間共分散行列とクラス内共分散行列は次式で計算できる：

$$\widehat{S}_B^{\mathrm{PU}} = \left(\widehat{\boldsymbol{m}}_2^{\mathrm{PU}} - \widehat{\boldsymbol{m}}_1\right)\left(\widehat{\boldsymbol{m}}_2^{\mathrm{PU}} - \widehat{\boldsymbol{m}}_1\right)^\top,$$
$$\widehat{S}_W^{\mathrm{PU}} = \pi_1 \frac{1}{n_1}\sum_{i \in \mathcal{C}_1} (\boldsymbol{x}_i^{(1)} - \widehat{\boldsymbol{m}}_1)(\boldsymbol{x}_i^{(1)} - \widehat{\boldsymbol{m}}_1)^\top + \left(\frac{1}{n_U}\sum_{i \in \mathcal{C}_U} (\boldsymbol{x}_i^{(U)} - \widehat{\boldsymbol{m}}_2)(\boldsymbol{x}_i^{(U)} - \widehat{\boldsymbol{m}}_2)^\top - \pi_1 \frac{1}{n_1}\sum_{i \in \mathcal{C}_1} (\boldsymbol{x}_i^{(1)} - \widehat{\boldsymbol{m}}_2)(\boldsymbol{x}_i^{(1)} - \widehat{\boldsymbol{m}}_2)^\top\right).$$

4.3 多クラス分類における PU-LDA

次に，2 クラス分類を一般化して，多クラス分類問題を考える．ここでは，K 個のクラス（クラス 1 からクラス K）を考える．私たちは，クラス 1 からクラス $K-1$ に属するラベル付きデータと，クラス K に属するデータとそれ以外のクラスに属

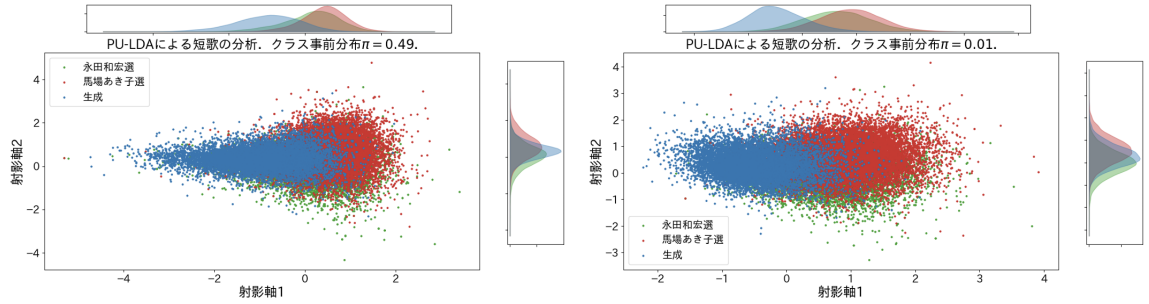

図 1: PU-LDA を用いる朝日歌壇短歌の分類．左図はクラス事前分布 $\pi = 0.49$，右図はクラス事前分布 $\pi = 0.01$ で分析．x 軸は LDA によって得られた射影軸 1 に，y 軸は射影軸 2 に対応．クラス事前分布 $\pi = 0.49$ の場合，生成短歌の多くは永田氏もしくは馬場氏によって選ばれると仮定されており，分析結果も生成短歌（青点）の多くが永田選（緑点）・馬場選（赤点）と重なるように位置している．逆に，クラス事前分布 $\pi = 0.49$ の場合，生成短歌の多くは永田氏にも馬場氏にも選ばれないと仮定しており，分析結果でも生成短歌は永田選・馬場選から離れた位置に分布している．

するデータが混ざったラベルなしデータを観測できるとする．この設定では，クラス 1 からクラス $K-1$ までのラベル付きデータは，2 クラス分類の正例データに対応する．私たちは 2 クラス分類と同様に，クラス K の $\boldsymbol{m}_K$ について $\boldsymbol{m}_K := \mathbb{E}_{k=K}[\boldsymbol{x}] = \left(\int \boldsymbol{x} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \sum_{a\in[K-1]} \pi_a \int \boldsymbol{x} p(\boldsymbol{x}\,|\,k=a)\mathrm{d}\boldsymbol{x}\right)/\pi_K$．と識別できる．この $\boldsymbol{m}_K$ をサンプルで置き換えたものを $\widehat{\boldsymbol{m}}_K^{\mathrm{PU}}$ とする．すなわち，$\widehat{\boldsymbol{m}}_K^{\mathrm{PU}} := \left(\frac{1}{n_U}\sum_{i\in\mathscr{C}_U} \boldsymbol{x}_i^{(U)} - \sum_{a\in[K-1]} \pi_a \frac{1}{n_a}\sum_{i\in\mathscr{C}_a} \boldsymbol{x}_i^{(a)}\right)/\pi_K$ である．この $\widehat{\boldsymbol{m}}_K^{\mathrm{PU}}$ を，従来の LDA における $\widehat{\boldsymbol{m}}_K$ の代わりに用いることで LDA を実行できる．ここで，$k = 1, 2, \ldots, K-1$ について $|\mathscr{C}_k|/n \approx \pi_k$ である．

5 PU-LDA を用いる分析

5.1 クラスとクラス事前分布の定義

PU-LDA を用いて朝日歌壇短歌と生成短歌を分析する．以下の 3 クラスへの分類問題を考える：(クラス 1) 永田選短歌；(クラス 2) 馬場選短歌；(クラス 3) 永田氏にも馬場氏にも選ばれない短歌．

クラス事前分布 π_k $(k = 1, 2, 3)$ は，生成短歌にクラス k の短歌が含まれる割合を意味する．例えば，$\pi_1 = 0.1$，$\pi_2 = 0.1$，$\pi_3 = 0.8$ であれば，生成短歌全体のうち 10%の短歌が永田選に，10%の短歌が馬場選に，残りの 80%の短歌が永田氏にも馬場氏にも選ばれない短歌であることを意味する．

本研究では，簡単化のために永田選と馬場選のクラス事前分布が等しく π であると仮定する．

5.2 分類結果

図 1 に，クラス事前分布が $\pi_1 = \pi_2 = 0.49$ の場合と $\pi_1 = \pi_2 = 0.01$ の場合における PU-LDA の適用結果を可視化する．前者の事例（$\pi_1 = \pi_2 = 0.49$）は，生成短歌のほとんど（98%）が朝日歌壇に掲載されることを仮定している．後者の事例（$\pi_1 = \pi_2 = 0.01$）は，生成短歌の限られた一部（2%）のみが朝日歌壇に掲載されることを仮定している．

分析結果を確認すると，$\pi_1 = \pi_2 = 0.49$ の場合では，LDA で得られた射影軸上において生成短歌の群が他の選者の短歌の群に重なるように分布している．具体的な短歌については表 1 に示す．

一方で，$\pi_1 = \pi_2 = 0.01$ の場合では，生成短歌の群が他の選者の短歌の群から離れるように分布している．LDA で得られた射影軸上において，生成短歌の群が他の選者の短歌の群から離れるように分布している．具体的な短歌については表 2 に示す．

クラス事前分布による影響をより詳しく確認するために，付録の図 2 において，$\pi_1 = \pi_2$ が 0.40，0.30，0.20，0.10 である事例を掲載している．

5.3 射影軸ごとの分析

次に，射影軸ごとに分析結果を確認する．図 1 より，射影軸の値の大きさとクラス分類について以下の結果が得られる：

- 射影軸 1 について：
  - 値が大きい場合：永田選もしくは馬場選に分類．
  - 値が小さい場合：掲載されない短歌に分類．
- 射影軸 2 について：

表 1: クラス事前分布を $\pi_1 = \pi_2 = 0.49$ と設定した場合．射影軸 1 と 2 が大きい/小さい短歌の上位 5 首．

作者	短歌	選者	軸 1	軸 2
宮本陶生 [6]	手土産が軍備拡大わが国の首相訪米歓待受ける	永田	3.408	-1.172
諏訪兼位 [7]	ミサイルを打ち落とすとう兵器らしセールスマンも大物らしき	永田	3.175	0.777
森本忠紀 [8]	翁長知事の承認取り消し報じたる琉球新報紙面いきいき	永田	2.860	-3.600
三浦礼子 [7]	おもむろにうやうやしくも校長の白き手袋　教育勅語	永田	2.809	-0.615
猪野富子 [9]	蝶や蜂トカゲ・カエル・ヘビ源五郎生きもの調査員われいよよ忙し	馬場	2.746	1.166

作者	短歌	選者	軸 1	軸 2
岸本靖子 [10]	当たり前なんてどこにもないじゃないもう信じない当たり前など	永/馬	-5.312	0.371
甲斐みどり [11]	人生って多分そんなに甘くないけれど夕日はこんなにきれい	永田	-5.218	0.355
生成	暑いけどなんか幸せでうれしいな夏だからとか関係ないけど	生成	-4.733	0.434
生成	僕にも分からない僕らにも分からないけれど僕には分かる	生成	-4.710	-0.013
生成	幸せは遠いところにあって誰でも行けるけどちょっと遠い	生成	-4.438	-0.400

作者	短歌	選者	軸 1	軸 2
藤林正則 [11]	百頭の牛の畜舎にアライグマ飼料穀物のトウモロコシ食む	馬場	1.464	4.794
川名興 [6]	食草をジャコウアゲハが捜しおりウマノスズクサみつけ産卵	永田	1.275	3.651
白井澄江 [12]	ワカケホンセイインコ来て寒風に四十雀のエサむさぼりつくす	馬場	0.950	3.649
吉田孝 [6]	ペニテングダケを食するエゾシカの生きの摂理に驚かさるる	馬場	0.072	3.529
松井惠 [13]	開拓之碑鎮め三方原ばれいしょ畑に穀雨そぼ降る	馬場	0.683	3.507

作者	短歌	選者	軸 1	軸 2
由良英俊 [14]	法案は皆多数決異論無視民主制下の総理専制	永田	0.868	-4.323
森本忠紀 [8]	翁長知事の承認取り消し報じたる琉球新報紙面いきいき	永田	2.860	-3.600
森谷弘志 [15]	政権が逆にテレビを監視する内閣広報室の夕暮	永田	1.748	-3.404
中務進 [8]	閣僚の奥に控える官僚に答弁メモを渡す人いて	永田	0.821	-3.364
黒田祐花 [16]	銀行のカードの暗証番号は合格通知の受験番号	永田	2.111	-3.244

表 2: クラス事前分布を $\pi_1 = \pi_2 = 0.01$ と設定した場合．射影軸 1 と 2 が大きい/小さい短歌の上位 5 首．

作者	短歌	選者	軸 1	軸 2
内山豊子 [11]	近畿圏に五人のきょうだいそろったとばんざいをするみんな老人	馬場	3.917	0.024
米窪千加代 [17]	星取表モンゴルグルジアブルガリア・ロシアモンゴルエストニ..	馬場	3.841	0.618
森本忠紀 [8]	翁長知事の承認取り消し報じたる琉球新報紙面いきいき	永田	3.812	-1.996
三浦礼子 [7]	おもむろにうやうやしくも校長の白き手袋　教育勅語	永田	3.531	-0.518
東海正史 [9]	棲息するヒグマもろ共買い上げし大間原発段丘の森	馬場	3.322	1.789

作者	短歌	選者	軸 1	軸 2
生成	熱帯夜クーラーに消される街灯の火で夕涼みする	生成	-2.228	0.208
生成	この夏もやっぱり線香花火がいちばん好き夏の終わり	生成	-2.177	0.256
生成	暑いねと交わすだけ少し暑いからと返すだけで終わる夏	生成	-2.147	0.512
生成	寝苦しい夜と夏バテで寝苦しくて冷房つけずに寝る	生成	-2.118	1.687
生成	梅雨空の晴れ間は短いからそんな日におにぎりを握る	生成	-2.117	0.889

作者	短歌	選者	軸 1	軸 2
藤林正則 [11]	百頭の牛の畜舎にアライグマ飼料穀物のトウモロコシ食む	馬場	2.233	4.160
白井澄江 [12]	ワカケホンセイインコ来て寒風に四十雀のエサむさぼりつくす	馬場	0.763	3.605
松井惠 [13]	開拓之碑鎮め三方原ばれいしょ畑に穀雨そぼ降る	馬場	1.975	3.305
太田忠夫 [18]	高粱の束積む騾馬は高粱の穂をもぎりおり高粱畑	永田	1.514	3.255
生成	手塩にかけ育てた青大豆収穫の日が近づくとヒヨドリ来る	生成	0.602	3.199

作者	短歌	選者	軸 1	軸 2
由良英俊 [14]	法案は皆多数決異論無視民主制下の総理専制	永田	1.282	-3.274
村上敏之 [19]	会見の総理の脇でひたすらに原稿なぞる秘書官の指	永田	1.483	-2.888
中務進 [8]	閣僚の奥に控える官僚に答弁メモを渡す人いて	永田	1.040	-2.735
徳山富雄 [14]	平成と印刷されし領収書訂正印押す令和元年	永田	1.697	-2.684
関龍夫 [8]	質疑者も首相も原稿読んでいてまもなく行う強行採決	永田	1.702	-2.643

– 値が大きい場合：馬場選に分類．
– 値が小さい場合：永田選に分類．

これらの観察に基づき，射影軸の値ごとにどのような短歌が特徴的であるかを調査する．

5.3.1　射影軸 1 に関する結果の確認

射影軸 1 は，短歌を朝日歌壇に掲載されない短歌と掲載される短歌（永田選または馬場選）に分類する軸であると考えられる．観察の結果，射影軸 1 の値が小さい短歌の特徴として，短歌内で使用される語彙が少ないことや，同じ単語が繰り返されることが挙げられる．例えば，「僕にも分からない僕らにも分からないけれど僕には分かる」という生成短歌は「僕」を 3 度繰り返している．また，「寝苦しい夜と夏バテで寝苦しくて冷房つけずに寝る」という生成短歌も「寝苦しい」や「寝」を繰り返している．

もう一つの特徴として，短歌内の話題の広がりが乏しいことがある．例えば，「熱帯夜クーラーに消される街灯の火で夕涼みする」という生成短歌は，「熱帯夜」「クーラー」「涼み」といった同じ課題に関連する語が繰り返されている．一方で，射影軸 1 の値が大きい短歌（永田選と馬場選）は，短歌内で話題の広がりがあり，使用される単語も豊富である．

さらに，射影軸 1 の値が小さい短歌には「夏」に関連する短歌が多かった．この原因として，生成短歌が「夏」に関する語を生成しやすい傾向があることが考えられる．一方で，永田選には政治的な内容が多く，馬場選には動物や畜産に関する内容が多いという傾向がある．これらの偏りが分析結果に影響を与えている可能性がある．

5.3.2　射影軸 2 に関する結果の確認

射影軸 2 は，短歌を馬場選と永田選に分ける軸であると考えられる．射影軸 2 の値が大きい短歌には動物や自然に関する単語が多く含まれており，これは馬場選の特徴であると考えられる．一方，射影軸 2 の値が小さい短歌には政治に関する単語が多く，これは永田選の特徴であると考えられる．

掲載されない短歌は，特定の話題への偏りが少なく，生成短歌を掲載されない短歌として定義していることが影響していると考えられる．すなわち，生成短歌に特定の話題への偏りがないため，本研究における掲載されない短歌も偏りが少ない短歌となっている可能性がある．今後は，生成短歌に特定の話題（例えば動物や畜産）を多く含む場合の分析結果についての検証などが必要である．

6　結論

本研究では，文埋め込みを施された短歌データに対して，PU-LDA 文埋め込みの手法を用いて分析を行った．私たちは分析を通じて，朝日歌壇に掲載されている短歌と掲載されない短歌の間の特徴を明らかにすることができた．掲載されない短歌の特徴として，語彙や話題の広がりの乏しさがある可能性を観察した．また，その分析の過程で，PU 学習における LDA の新手法を開発した．

参考文献

[1] C. Elkan and K. Noto, “Learning classifiers from only positive and unlabeled data,” in **International Conference on Knowledge Discovery and Data Mining**. Association for Computing Machinery, 2008, p. 213–220.

[2] 羽根田 賢和，浦川 通，田口 雄哉，田森 秀明，坂口 慶祐，“RLHF を用いた「面白い」短歌の自動生成の試み”，言語処理学会第 30 回年次大会論文集，3 2024.

[3] S. Arora, Y. Liang, and T. Ma, “A simple but tough-to-beat baseline for sentence embeddings,” in **International Conference on Learning Representations (ICLR)**, 2017.

[4] K. Ethayarajh, “Unsupervised random walk sentence embeddings: A strong but simple baseline,” in **Proceedings of the Third Workshop on Representation Learning for NLP**, 2018, pp. 91–100.

[5] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” **Annals of Eugenics**, vol. 7, no. 2, pp. 179–188, 1936.

[6] 朝日新聞社編，朝日歌壇 2023　朝日新聞出版，2024.

[7] ——，朝日歌壇 2017　朝日新聞出版，2018.

[8] ——，朝日歌壇 2015　朝日新聞出版，2016.

[9] 朝日新聞東京本社学芸部編，朝日歌壇 2002　朝日ソノラマ，2002.

[10] 朝日新聞社編，朝日歌壇 2013　朝日新聞出版，2013.

[11] ——，朝日歌壇 2013 1-12 月　朝日新聞出版，2014.

[12] ——，朝日歌壇 2018　朝日新聞出版，2019.

[13] 朝日新聞東京本社学芸部編，朝日歌壇 2005　朝日ソノラマ，2005.

[14] 朝日新聞社編，朝日歌壇 2019　朝日新聞出版，2020.

[15] ——，朝日歌壇 2020　朝日新聞出版，2021.

[16] ——，朝日歌壇 2010　朝日新聞出版，2010.

[17] ——，朝日歌壇 2011　朝日新聞出版，2023.

[18] 朝日新聞東京本社学芸部編，朝日歌壇 2006　朝日ソノラマ，2006.

[19] 朝日新聞社編，朝日歌壇 2022　朝日新聞出版，2023.

[20] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, “Towards universal paraphrastic sentence embeddings,” in **International Conference on Learning Representations (ICLR)**, 2016.

[21] H. Yamagiwa, M. Oyama, and H. Shimodaira, “Discovering universal geometry in embeddings with ICA,” in **Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2023.

[22] 近藤泰弘，“和歌集の歌風の言語的差異の記述 – 大規模言語モデルによる分析 –”，日本語の研究，vol. 19，no. 3，pp. 105–117，Dec. 2023.

[23] 田口雄哉，田森秀明，人見雄太，西鳥羽二郎，菊田洸，“同義語を考慮した日本語単語分散表現の学習”，情報処理学会第 233 回自然言語処理研究会，vol. 2017-NL-233，no. 17，Oct. 2017，pp. 1–5.

A　関連研究

文埋め込み [20] は自然言語の文をベクトル表現で表すことにより，定量的に扱うことが可能となる手法である．しかし，一般的に埋め込み表現で与えられるベクトルそのものを解釈することは難しいとされている．解釈のために，独立成分分析を用いる手法などが提案されている [21]．短歌の研究においては，[22] は主成分分析を用いて分析している．

B　朝日新聞単語ベクトル

この単語ベクトルは，朝日新聞社が保有する 1984 年 8 月から 2017 年 8 月までに掲載された記事のうち，約 800 万記事（延べ 23 億単語）を用いて学習しされている [23]．単語分割には MeCab を使用し，辞書は IPADIC-2.7.0 を用いている．単語ベクトルのモデルは，Skip-gram と CBOW を word2vec で学習されているほか，GloVe によって学習したモデルを提供する．朝日新聞単語ベクトルは 300 次元の単語ベクトルであるため，それを用いて得られる文埋め込みベクトルも 300 次元のベクトルになる．

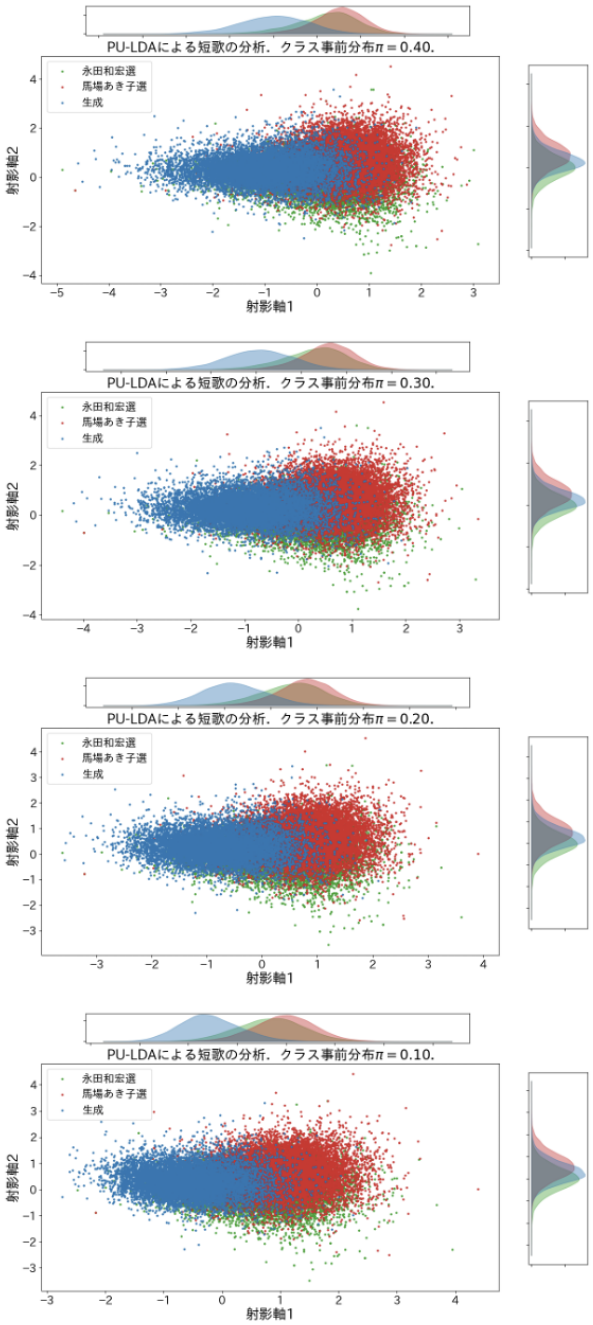

図 2: PU-LDA を用いる朝日歌壇短歌の分類．上から順にクラス事前分布 $\pi = 0.40, 0.30, 0.20, 0.10$ で分析．x 軸は LDA によって得られた射影軸 1 に，y 軸は射影軸 2 に対応．