発話スタイルの類似とユーザ本人による対話の好ましさの相関

沼屋 征海 [1]　守屋 彰二 [1]　佐藤 志貴 [1,2*]　赤間 怜奈 [1,3]　鈴木 潤 [1,3]
[1] 東北大学　[2] 株式会社サイバーエージェント　[3] 理化学研究所
{numaya.ikumi.t4,shoji.moriya.q7}@dc.tohoku.ac.jp
{shiki.sato.d1,akama,jun.suzuki}@tohoku.ac.jp

概要

対話システムのユーザ本人にとって好ましい応答生成に向け，ユーザの嗜好等を発話内容に反映する試みが盛んに行われている．一方，生成する発話のスタイルもシステムへの印象に影響を与える可能性が先行研究で示唆されている．本研究では，ユーザの過去発話に対するシステム発話のスタイル類似度を自動評価する手法を提案し，システム発話のスタイル類似度とユーザによるシステムの主観評価との相関を調査する．両者の間で正の相関を確認し，発話スタイルの類似がユーザ個人にとっての対話の好ましさの一側面となっていることを明らかにした．

1　はじめに

対話型大規模言語モデル (LLM) の発達を背景に，内容や一貫性のある応答を生成できる対話システムが構築可能となりつつある [1, 2, 3]．これにより，一般的な応答の質を高めるだけではなく，ユーザ本人にとって好ましい応答の生成にも注目が集まっている．特に先行研究では，長期的な文脈やユーザ情報を記憶すること [4] やユーザの感情を推定すること [5] で，ユーザ本人にとってより好ましい内容を含む応答を生成するための研究が行われている．

一方，応答の内容に限らず，システムの発話スタイルがユーザ本人にとっての対話の好ましさに影響を与える可能性も報告されている．特に，話者間の同調が対話の好ましさと相関することが示唆されている．同調とは，対話の進行に伴い，互いの発話スタイルや語彙選択，間合いなどが類似する現象である．システムの発話がユーザの発話に類似している例，していない例を表 1 に示す．Giles らは，人同士の対話において，同調が相手話者に対する印象や共感度の向上に寄与することが示した [6]．さらに，Kanezaki らは，内容と語彙のスタイルに関する同調を考慮し，シングルターンのシステムとユーザ間の対話において，発話間の類似度をもとに応答候補を選択する手法を提案した [7]．人手評価から，発話スタイルの類似性が高い応答において，応答の面白さや対話意欲が高くなる傾向が示された．

表 1　発話スタイル類似性の有無の例．有が太字．

話者	発話例
User	*Hey, my phone's broken.*
System A	***Oh no, what's up with it?***
System B	*Please explain the issue in detail.*
User	*Super slow.*
System A	***Maybe clear out some apps or junk files.***
System B	*Check storage and apps.*

本研究では，マルチターンの対話に対して，システムの発話とユーザの文脈における発話スタイルの類似度を算出し，主観評価との相関を測ることで，発話スタイルの類似度が本人による好ましさに影響を与えるかを検証する．実験の結果，両者の間に正の相関が確認でき，発話スタイルの類似がユーザ自身におけるシステムへの好ましさの一側面となっていることが明らかになった．この結果から，ユーザ個人にとって好ましい応答生成の実現には，発話内容だけではなく発話スタイルも考慮する必要があることが示唆される．

2　関連研究

2.1　同調による話者への影響

同調とは，対話の進行に伴い，互いの発話スタイルや語彙選択，間合い，声の高さなどが徐々に類似していく現象を指し，収束やエントレインメントとも呼ばれる．Giles らは，Communication Accommodation Theory (CAT) の枠組みの中で，人同士の対話における同調の具体例とその効果について言及しており [6]，同調が魅力度 [8] や親密度 [9]，信

* 東北大学の学術研究員としての成果

頼感 [10] など，相手話者に対する印象を向上させることを示している．また，Nenkova らは，人同士の対話における頻出単語に着目した同調の定量化手法を提案し，同調が対話の自然さや対話を介したタスクの成功率に影響を与えることを示した [11]．

これら先行研究を踏まえ，同調を対話システムに応用する研究がなされている [7]．本研究でも，対話システムによる同調が人に与える影響を考える．

2.2 言語的同調の定量化

Nasir らは，話者間の語彙的同調を定量化する手法として CLiD を提案した [12]．語彙的同調とは，各話者の語彙選択の傾向やその意味が類似していく現象を指す．この手法では，ある発話をアンカー発話とし，アンカー発話に対する相手話者の後続の複数発話を文脈として考慮する．アンカー発話と文脈中の発話の全ペアにおいて Word Mover's Distance (WMD) を計算し，その最小値をアンカー発話と相手話者との語彙的距離として定義する．N ターンの対話における話者 a の発話系列 $(u_1^{\mathrm{a}}, u_2^{\mathrm{a}}, \cdots\cdots, u_N^{\mathrm{a}})$ と話者 b の発話系列 $(u_1^{\mathrm{b}}, u_2^{\mathrm{b}}, \cdots\cdots, u_N^{\mathrm{b}})$ からなる対話 $(u_1^{\mathrm{a}}, u_1^{\mathrm{b}}, u_2^{\mathrm{a}}, u_2^{\mathrm{b}}, \cdots\cdots, u_N^{\mathrm{b}})$ において，話者 a の i ターン目の発話 u_i^{a} をアンカー発話とし，話者 b の k ターン分の後続発話 u_j^{b} を文脈として考慮したとき，その語彙的距離は以下のように表す：

$$\mathrm{LID}^{i}_{\mathrm{WMD}} = \min_{i \le j < j+k} \mathrm{WMD}(u_i^{\mathrm{a}}, u_j^{\mathrm{b}}) \tag{1}$$

このようにして計算した話者 a の全発話に対する距離の平均値を，話者 a による話者 b の発話への語彙的距離 CLiD (Conversational Linguistic Distance) と定義し，語彙的同調の度合いとした．この CLiD によって定量された同調とカウンセリングにおける共感度との間に正の相関があることが示された．

Kanezaki らは，単語間だけでなく，文全体の意味的類似性を考慮するため，Nasir らが提案した CLiD を拡張し，新たに BERTScore による意味的距離を導入した [7, 13]．この手法では，意味的距離を以下のように定義している：

$$\mathrm{LID}^{i}_{\mathrm{BERT}} = \min_{i \le j < j+k} (1 - \mathrm{BERTScore}(u_i^{\mathrm{a}}, u_j^{\mathrm{b}})) \tag{2}$$

また，$\mathrm{LID}^{i}_{\mathrm{WMD}}$ の算出には後述する単語スタイルベクトルを用いることで，単語のスタイル類似性も考慮した同調の定量化を行った．これらの $\mathrm{LID}_{\mathrm{BERT}}$ と $\mathrm{LID}_{\mathrm{WMD}}$ を用いて，シングルターンの対話における次の発話が持つべき同調の度合いを予測するモデルを構築し，応答候補の中から同調を重視した発話を選択する手法を提案した．その結果，従来手法と比較して，人の主観評価が高い応答候補を選択できることが示された．

本研究では，上述のシングルターンの対話における語彙や内容に基づく同調とは異なり，マルチターンの対話における発話文や文脈全体における発話スタイルの同調を定量化する新たな手法を提案する．

2.3 発話スタイルを捉えた単語ベクトル

対話を行う話者には，それぞれ固有の発話スタイルが存在する．方言や敬語などがその例である．Akama らはこのようなスタイルを定量化する手法として，スタイル類似性を捉えた単語ベクトル空間を提案した [14]．本稿では，これを単語スタイルベクトルと呼ぶ．単語の意味をモデリングしたベクトル空間である Countinuous Bag-of-Words (CBOW) [15] では，近傍の単語集合を入力として，ターゲット単語の予測確率が最大化するように学習が行われる．一方，単語スタイルベクトルでは，CBOW の考え方を逆転させ，近傍の単語集合を取り除くことで意味を排除し，発話スタイルのみを捉えることを可能にした．具体的には，ある単語 w に対し，この手法で学習したスタイルのみを捉えたベクトル $\boldsymbol{x}_w$ と CBOW と同様の手法で学習したベクトル $\boldsymbol{y}_w$ を連結することで，発話スタイルと意味の両方を反映した単語スタイルベクトル $\boldsymbol{v}_w$ を得る：

$$\boldsymbol{v}_w = \boldsymbol{x}_w \oplus \boldsymbol{y}_w \tag{3}$$

本研究では，この単語スタイルベクトルを用いて，発話スタイルの類似度を評価する．

3 話者間の発話類似度評価

2 節で述べたように，Kanezaki らの先行研究では，シングルターンの対話において，内容や語彙に基づく類似度を測定することで同調のモデリングを行った．しかし，実際の対話はマルチターンで進行するため，一つの発話に対する類似性だけでなく，相手の過去の文脈との類似性を考慮したモデリングが求められる．本研究では，発話スタイルを捉えた単語ベクトル空間を用いて，話者の発話文および文脈を埋め込みとして表現し，それらの類似度を測定することで同調の度合いを評価する．このアプローチにより，内容に依存せずユーザの話し方に基づく発話スタイルの類似性を考慮することが可能になる．

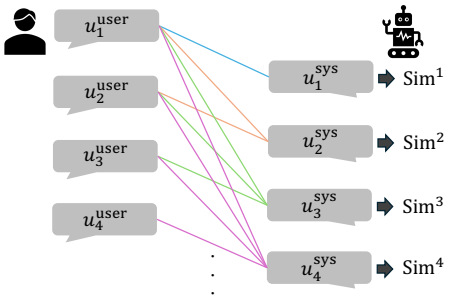

図 1　システムの各発話に対する類似度評価範囲．発話スタイルの類似度はシステムの各発話ごとに算出し，その発話以前のユーザの過去文脈全体との間で計算を行う．

具体的には，図 1 のように，着目するシステム発話とそれ以前のユーザの全発話との類似度を測定する．また，発話スタイルの類似度評価手法の比較対象として，語彙的同調を定量化する WMD および意味的同調を定量化する BERTScore を用いる．

3.1　スタイルベクトル類似度

単語スタイルベクトル（2.3 節）の平均により，着目する発話や文脈のスタイルベクトルを得る．話者 x の i ターン目の発話に含まれる全単語の集合を W_i^x，i ターン目までの発話に含まれる全単語の集合を $W_{\leq i}^x$ と表記する．また，単語集合 W_i^x，$W_{\leq i}^x$ に含まれる全単語スタイルベクトルの平均をそれぞれ $\boldsymbol{v}_{W_i^x}$，$\boldsymbol{v}_{W_{\leq i}^x}$ と表記する．

本研究では，システムの各発話に対してユーザの過去発話全体との類似度を算出する．i ターン目の対話システムの発話の類似性スコアは，$\boldsymbol{v}_{W_i^{sys}}$ とこれ以前のユーザの全文脈 $\boldsymbol{v}_{W_{\leq i}^{user}}$ に対するスタイルベクトルのコサイン類似度を用いて算出する．

$$\mathrm{Sim}^i_{\mathrm{Vec}} = \cos(\boldsymbol{v}_{W_i^{sys}}, \boldsymbol{v}_{W_{\leq i}^{user}}) \tag{4}$$

3.2　WMD，BERTScore による類似度

2 節で言及した CLiD は発話間の距離を表しており，その値が小さいほど発話間の類似性が高いことを示す．一方，本研究では，値が大きいほど類似性が高いことを表す類似性スコアを用いる．具体的には，WMD および BERTScore により算出される CLiD を 1 から引くことで，i ターン目のシステムの発話に対する類似性スコアを定義する．なお，話者 x の i ターン目の発話を u_i^x とする．

$$\mathrm{Sim}^i_{\mathrm{WMD}} = \min_{1\leq j\leq i} \mathrm{BERTScore}(u_i^{sys}, u_j^{user}) \tag{5}$$

$$\mathrm{Sim}^i_{\mathrm{BERT}} = \min_{1\leq j\leq i} (1 - \mathrm{WMD}(u_i^{sys}, u_j^{user})) \tag{6}$$

以上の手法では，ユーザとシステムの発話ペアに対して類似度を算出している．一方で，発話間の類似性だけでなく，システムの発話に対するユーザ文脈全体の類似性を考慮することも重要である．そこで，ユーザ文脈中の全発話を結合した u_i^{sys} に対する類似性スコアを導入する．1 から i ターン目までのユーザの発話文をつなぎ合わせたものを $u_{\leq i}^{user}$ と表記し，この文脈全体に基づく類似性スコアを以下のように表す．なお，BERTScore を用いて算出する場合，[SEP] トークンによって発話文を結合する．

$$\mathrm{Sim}^i_{\mathrm{BERT_{cat}}} = \mathrm{BERTScore}(u_i^{sys}, u_{\leq i}^{user}) \tag{7}$$

$$\mathrm{Sim}^i_{\mathrm{WMD_{cat}}} = 1 - \mathrm{WMD}(u_i^{sys}, u_{\leq i}^{user}) \tag{8}$$

4　実験設定

実験では，発話スタイルに関する類似性スコアとユーザ本人の主観評価との相関を測定し，このスコアがユーザ本人による好ましさに影響を与えるかを検証する．さらに，類似性スコアを好ましさに関連する他の自動評価手法の評価値に足し合わせることで，ユーザ本人による好ましさをより高精度に評価できるかを検証する．

データセット．　本研究では，2018 年に行われた ConvAI2 competition の対話データを収集した ConvAI2 データセットを使用した [16]．このデータセットには，システムとユーザによる対話のテキストデータと，ユーザが対話終了後に付与した主観評価ラベルが含まれている．評価ラベルは，fluency，consistency，engagingness の観点に基づいて 1-5 の 5 段階で総合評価したスコアである．このようにユーザが付けた評価値を用いることで，発話スタイルの類似性とユーザ本人による好ましさとの関連性を測ることができると考えられる．

ベースライン評価．　類似性スコアとユーザ本人の主観評価との相関の強さを評価する基準として，LLM-Eval [17] とユーザ主観評価との相関を算出した．LLM-Eval は，BLEU-4[18] や ROUGE-L[19] などの従来の自動評価指標と比較して，より人手評価に近い対話評価が可能であることが示されている [17]．本実験では LLM-Eval の算出に用いる LLM として GPT-4o-mini[1)] を採用した．プロンプトには出力形式，対話文脈，およびそれに続く対話システ

1)　https://platform.openai.com/docs/overview.

表 2 類似性スコアと主観評価スコアとの相関

評価スコア	スピアマン相関係数
Base	0.28
Sim_{Vec}	**0.28**
Sim_{WMD}	0.21
Sim_{BERT}	0.14
$\text{Sim}_{\text{WMD,ctx}}$	0.19
$\text{Sim}_{\text{BERT,ctx}}$	0.00

ムの発話を与え，指定した観点についてターンごとに 0 から 100 のスコアを出力させた．なお，類似性スコアの範囲に合わせるため，出力スコアを 100 で割り正規化した値を最終的な評価値（以下，ベースラインスコア）として使用した．本研究では，好ましさの一観点として発話スタイルの類似性に注目しているため，LLM-Eval を用いて各発話に対する “preference” の評価を行った．i ターン目のベースラインスコアを Base^i と表記する．

重み付きスコア． スタイルに関する類似性スコアをベースラインスコアと組み合わせることでユーザの主観評価とより強い相関を持つ評価が可能かを検証した．各発話に対しベースラインスコアと類似性スコアを重み付きで足し合わせたときのスコアを算出した．類似性スコアの重みを α としたときの i ターン目の重み付きスコアは以下の式で表される：

$$\text{Weighted\_Score}^i = \text{Sim}^i * \alpha + \text{Base}^i * (1 - \alpha) \quad (9)$$

主観評価との相関の測定方法． データセットの評価ラベルは対話単位で付与されているため，各ユーザの発話に対する類似性スコアの平均値を計算し，それを対話レベルの類似性スコアとして算出した．これを用いて，評価ラベルとの相関を測定した．なお，測定にはスピアマン相関係数を用いた．

5 結果および考察

5.1 類似度と主観評価との相関

結果． 表 2 にベースラインスコアおよび類似性スコアと対話に付与されたユーザ本人の主観評価スコアとのスピアマン相関係数を示す．ベースラインスコア Base では 0.28 の正の相関を示した．類似性スコアに関しては，スタイルベクトル類似度 Sim_{Vec} が 0.28 の正の相関を示し，ベースラインと同等であり，最も高い相関を示した．

考察． 4 節で述べたように，ベースラインとして採用した LLM-Eval によるスコアは，人手評価と高い相関を持つことが確認されている．結果から，

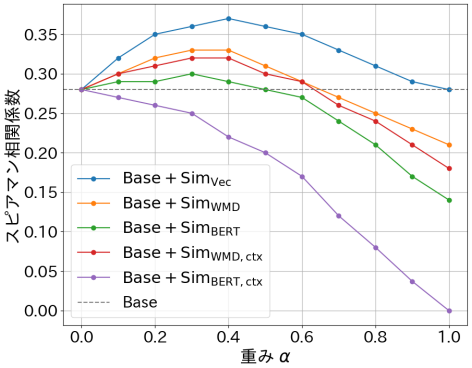

図 2 重み付きスコアとユーザの主観評価スコアの相関

今回新たに提案した発話スタイルに関する類似性スコアの相関係数はベースラインスコアと同等であり，発話スタイルの類似がユーザ本人にとっての好ましさの一側面である可能性が示唆された．

5.2 重み付きスコアと主観評価との相関

結果． 図 2 に α を変化させたときの重み付きスコアと主観評価との相関を示す．スタイルベクトル類似度 Sim_{Vec} を用いた重み付きスコアの相関を見ると，重み α が上昇するに従って相関が上昇し，$\alpha = 0.40$ 付近で最大となることが分かった．このとき，ベースラインの相関係数から 0.09 ほど向上していることが確認された．

考察． ベースラインスコアと発話スタイルによる類似性スコアを重み付きで足し合わせ，適切な重みを選択した場合，両者を単独で用いるよりも高い相関が得られた．このことから，発話スタイルの類似性を考慮することで，ユーザによる好ましさをより高精度に自動評価できる可能性が示唆された．

6 おわりに

本研究では，ユーザ本人が対話システムに対して対話の好ましさを感じるための一要素として，発話スタイルの類似性に注目した．実験では，システムの発話に対するユーザの過去文脈とのスタイル類似を定量化し，これがユーザの主観評価と相関するという結果を得た．このことは，発話スタイルの類似性がユーザ本人による好ましさに影響を与える可能性を示唆する．また，既存の自動評価指標に類似性スコアを組み込むことで，主観評価との相関がさらに向上することも確認した．今後の研究では，発話スタイルの定量化精度の向上や他モダリティの同調，話者間の類似性と異なる観点から好ましさを解明する新たなアプローチの模索を計画している．

謝辞

本研究は JSPS 科研費 JP22K17943，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の助成を受けたものです．

参考文献

[1] OpenAI. Introducing chatgpt. https://openai.com/index/chatgpt/.

[2] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pages 270–278, Online, July 2020. Association for Computational Linguistics.

[3] Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. Building a personalized dialogue system with prompt-tuning. In Daphne Ippolito, Liunian Harold Li, Maria Leonor Pacheco, Danqi Chen, and Nianwen Xue, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop**, pages 96–105, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[4] Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. **CoRR**, abs/2402.16288, 2024.

[5] Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. SEPRG: Sentiment aware emotion controlled personalized response generation. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, **Proceedings of the 14th International Conference on Natural Language Generation**, pages 353–363, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[6] Howard Giles, Tania Ogay, et al. Communication accommodation theory. **Explaining communication: Contemporary theories and exemplars**, pages 293–310, 2007.

[7] 金崎 翔大, 河野 誠也, 湯口 彰重, 桂井 麻里衣, and 吉野 幸一郎. エントレインメント尺度および戦略が対話システムの評価に与える影響の調査. In **言語処理学会第 30 回年次大会発表論文集**, pages 1384–1388, 3 2024.

[8] Richard L. Street, Robert M. Brady, and William Benjamin Putman. The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. **Journal of Language and Social Psychology**, 2:37 – 56, 1983.

[9] Marianne LaFrance. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. **Social Psychology Quarterly**, pages 66–70, 1979.

[10] Makiko Imamura, Yan Bing Zhang, and Jake Harwood. Japanese sojourners' attitudes toward americans: Exploring the influences of communication accommodation, linguistic competence, and relational solidarity in intergroup contact. **Journal of Asian Pacific Communication**, 21(1):115–132, 2011.

[11] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In **Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08**, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[12] Md. Nasir, Sandeep Nallan Chakravarthula, Brian R. Baucom, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Modeling interpersonal linguistic coordination in conversations using word mover's distance. **CoRR**, abs/1904.06002, 2019.

[13] 金崎 翔大, 河野 誠也, 湯口 彰重, 桂井 麻里衣, and 吉野 幸一郎. エントレインメントスコアを用いた応答リランキングとその自動評価. In **言語処理学会第 29 回年次大会発表論文集**, pages 1963–1968, 3 2023.

[14] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised learning of style-sensitive word vectors. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers**, pages 572–578. Association for Computational Linguistics, 2018.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2013.

[16] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). In **The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations**, pages 187–208. Springer, 2020.

[17] Yen-Ting Lin and Yun-Nung Chen. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In **Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)**, pages 47–58, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02**, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pages 74–81, 2004.

表 3　主観評価ラベル 5 の対話に対する評価の実例

話者	発話文	Base	Base+Sim_{Vec}
システム	*That is great. I want to be an accountant.*		
ユーザ	*Yes, it is a very safe job. I have broken my arm last week.*		
システム	*I am sorry to hear that.*	0.80	0.76
ユーザ	*I was skating and fell down, so pity!*		
システム	*Wow, that sounds so cool!*	0.75	0.73
ユーザ	*Not so cool as it might be.*		
システム	*I will have to try it.*	0.60	0.70
ユーザ	All my friends are skate-boarders. Do you like ...		
システム	*I like to play sports.*	0.70	0.71
ユーザ	*And do you have a hat? I wear mine everywhere and everytime!*		
システム	*I have a few, that is a great idea!*	**0.70**	**0.75**
ユーザ	*Let's take our hats and have a walk!*		
システム	*I love the smell of the trees and it is a beautiful day*	0.75	0.72
ユーザ	*Yes, sounds great! What is your job?*		
システム	*I work at a bank, and I work at a local place.*	0.70	0.71

A　具体的な対話評価例

ベースラインスコアおよびスタイルベクトル類似度を用いた重み付きスコアによる評価結果を，実際のデータセット中の対話例を用いて表 3 に示す．今回用いた重み α は，図 2 において相関係数が最大となった $\alpha = 0.4$ を採用している．

この対話では，主観評価ラベルとして最高値の 5 が付与されており，評価スコアが高くなることが望ましい．システムの発話 *I have a few, that is a great idea!*に着目すると，ベースラインスコアと比較して重み付きスコアの方が高い評価を示している．この発話以前のユーザの文脈では，"*!*" が 2 度使用されているなど，ユーザの発話とスタイルが類似していると考えられる．このような観点から，類似性スコアが向上し，重み付きスコアも上がったと考えられる．以上のように，ユーザの主観評価に近いスコアを算出できている事例が確認された．

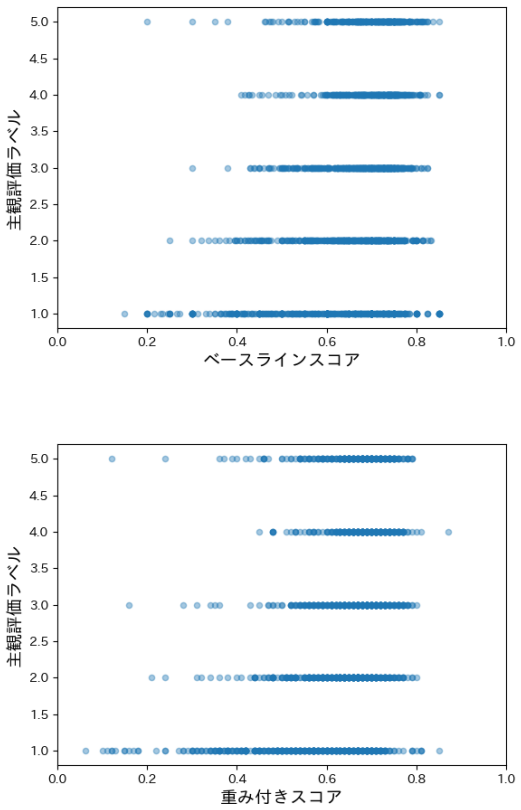

図 3　重み付け前後の主観評価との相関

B　重み付け前後の相関の変化

図 3 にベースラインスコアおよび重み付きスコアとデータセットに付与された主観評価ラベルとの相関の分布を示す．今回用いた重みは $\alpha = 0.4$ である．分布の相関係数は，ベースラインスコアが 0.28，重み付きスコアが 0.37 であり，重み付きスコアの方が高い相関を示している．

両者を比較すると，評価ラベル 1 の場合，重み付きスコアの方がより低い値に分布しており，これが相関の向上に寄与していると考えられる．一方で，重み付きスコア単体で見ると，評価ラベル 5 よりも 4 のほうが高い値に多く分布しており，この点においては想定とは異なる結果となった．