評価の階層性に着目した雑談対話システム評価の分析

蔦 侑磨
東京大学大学院 情報理工学系研究科
tsuta@tkl.iis.u-tokyo.ac.jp

吉永 直樹
東京大学 生産技術研究所
ynaga@iis.u-tokyo.ac.jp

概要

文生成タスクでは，評価の解釈性向上のために多様な側面からの評価が期待される．複数の評価軸により評価を行う際，文評価に普遍的に見られる評価の階層性に注意する必要がある．評価軸間に階層的な依存性がある場合，依存先の評価値は依存元の評価値の影響を受けるため，誤った結論を導く可能性がある．本研究では雑談対話応答生成タスクの評価観点について，対話システム間の有意差を手掛かりに，評価軸間の階層的依存性を明らかにする試みを行った．さらには階層的な依存性を解消した場合のシステム間性能比較を再分析した．多様な観点で評価された雑談対話のデータセットを利用した実験から，階層的依存性の解消により対話システム間の性能有意差の有無が変化しうることを確認した．

1 はじめに

機械翻訳や自動要約，雑談対話などの文生成タスクでは，正解となりうる表現が様々に考えられるため，出力文を人手で評価したり，人手評価に相関する自動評価手法が提案されている．また効率的なシステム開発のために総合評価だけでなく，評価の解釈性向上のための多様な側面からの評価が期待される．特に雑談対話生成応答タスクでは，応答の関連性 [1] や，会話の一貫性 [2]，エンゲージメント [3] などの様々な側面で人手評価が行われる．自動評価手法でもこれに追随して個別の評価軸に特化した手法が提案されている [4]．

評価軸が複数存在する場合，個々の評価軸は異なる評価観点のため，それらは互いに独立した評価であることが望ましい．しかし実際には，個々の評価軸が独立せず相関が高いことが指摘されている [5]．この原因としては，評価の階層性 [6] が考えられる (図 1)．評価の階層性とは，例えば文の流暢性と応答関連性を例に挙げると，流暢性の低い文は理解できないため応答関連性を評価することすら困難であるが，一方で応答関連性が流暢性評価に影響することはない．このような評価の階層性は文レベルの評価において普遍的な特徴であり，一部の自動評価手法ではこの階層的依存関係を利用する [6, 7]．

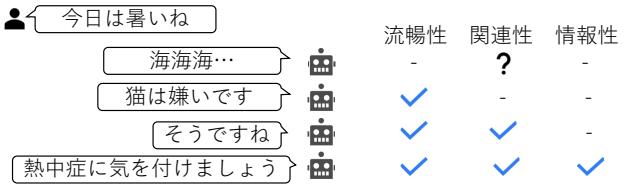

図 1 評価の階層性を表現した評価例．例えば流暢性と関連性では，応答が流暢ではないことにより応答関連性を適切に評価できないという非対称な関係性がある．

この評価軸の階層的依存性が意識されないと，個々の評価軸で誤った結論を下す可能性が考えられる．例として，対話システム A・B の性能比較を，文の流暢性と応答関連性により評価する状況を考える．そしてシステム A はシステム B と同程度に関連性の高い応答を生成する能力があるが，システム B より流暢ではないとする．流暢でない文の評価は困難であるため，応答関連性評価についてもシステム A がシステム B より劣るという誤った結論が下される可能性がある．このような状況では応答関連性評価の解釈として不適切な結論が導かれている．

そこで本研究では，評価の階層性や階層的依存性に起因する問題について分析を行う．この分析において，検定に基づく対話システム間の性能有意差を手掛かりに，対話システムペアをグループ化した対話システムペア集合を活用する．この集合を利用した評価軸間の網羅的な分析により，階層的依存関係にある評価軸ペアを明らかにする．さらに，明らかにした評価軸ペア間の階層的依存性を解消した場合の対話システム間性能有意差の再分析を行う．

実験では，DSTC9 [8] と USR データセット [7] 上での複数の対話システムへの複数の評価軸による人手評価を使って分析を行った．実験結果から，評価

軸間の階層的依存性によるバイアスは一部の対話システム間の有性能意差の有無に影響を及ぼすことが確認された．

2 関連研究

2.1 対話システム評価方法の分析

雑談対話研究の評価では，どのような会話で評価を行うか，どのように評価を行うかなど，多様な観点から妥当な評価方法に関する議論が行われてきた．例えば，対話システムの会話データ構築やアノテーションのコスト [9, 10, 11]，評価者間一致度などのアノテータの質に関する分析 [12] や高品質なアノテーションを得るための方法 [13] などである．

しかしこれらの研究では，評価の評価軸間関連性の分析までは行われていない．本研究では文レベルの評価に普遍的な評価の階層性という特徴に注目し，特に評価軸間の階層的依存性に起因するバイアスによる評価への影響がないかを観察する．

2.2 評価軸間関連性を意識した自動評価

雑談対話システムでは応答の関連性 [1] や，会話の一貫性 [2]，エンゲージメント [3] など様々な評価軸に特化した自動評価手法が提案されている [4]．その中でも評価軸間の関連性・階層性に着目して複数の評価手法を組み合わせた手法がある．Mehri らは，個々の側面に特化した自動評価評価を組み合わせることで，多様な評価軸に適応可能な自動評価手法を提案した [7]．具体的には，生成応答文の総合評価の予測のために，文の理解しやすさや応答の自然さなどの少量の個別評価データを用いて回帰分析を行い，個別の側面に特化した自動評価評価を適切に重みづけることで，総合評価を予測する自動評価手法を構築可能にした．Phy らは，評価の階層性に着目して Mehri らの手法を拡張し，それぞれの評価段階に適切な自動評価手法を割り当てる方法を提案した [6]．

本研究では，これらの研究とは異なり評価軸間の階層的依存性を単純な回帰分析や恣意的な直感に基づいて決定するのではなく，統計的仮説検定と集合理論に基づく分析の併用により，評価軸間の網羅的な分析を行い，階層的依存関係にある評価軸ペアの解明を行う．

3 評価の階層性分析

本節では，§ 3.1 で実データに基づいて評価の階層性を可視化する提案手法を説明し，§ 3.2 で評価軸間の階層的依存性を解消したシステム間性能比較分析について説明を行う．

3.1 評価階層性の可視化

評価軸間の関連性を分析するにあたり，評価軸間の評価値の相関計測や回帰分析 [6, 7]，対話システム順位類似性 [5] などの利用が考えられる．これらの方法は評価値や評価値から導出される順位を用いるため，対話システム間性能比較の方法そのままであり，妥当性の高い方法だと考えられる．しかし，対話システム間性能比較にあたり順位や評価値の差分だけでなく，その差が有意であるかも重要な指標である．それ故に，本研究では対話システム間の性能差が有意であるかをベースに分析手法を構築した．

階層的依存関係にある評価軸ペアに関して，片方の評価軸では評価が頭打ちになり対話システム間性能有意差が検出されない一方で，他方では頭打ちが無いために有意差が検出されるという状況が想定される．そこで本研究では性能有意差がある対話システムペアの集合を利用することで依存関係にある評価軸ペアを抽出する．具体的な収集方法の説明のために，以下では性能有意差のある対話システムペア（「有意差ペア」）と，その集合を用いて抽出される階層的依存関係にある評価軸ペアについて説明する．

有意差ペア　対話システム S に関する N_S 個の会話サンプルに対して，評価軸 M の観点から行われた人手による評価値集合 $\mathcal{M}(N_S)$ が得られるとする．評価軸 M に関して，対話システム S_i と対話システム S_j の性能有意差が検定 $T(\mathcal{M}(N_{S_i}), \mathcal{M}(N_{S_j}))$ により $p < 0.01$ である対話システムペアを「有意差ペア」とする．なお．実験では人手評価はリッカート尺度などの離散的な順序尺度であるため，検定 T にはマン・ホイットニーの U 検定を利用した．また，有意差ペアの性能の優劣は区別する．

階層的依存関係にある評価軸ペア　対話システム集合 $\mathbf{S}$ について評価軸 M で確認されるすべての有意差ペアの集合を $\mathbf{D}(M, \mathbf{S})$ とする．本研究では評価軸 X, Y において $\mathbf{D}(M_X, \mathbf{S}) \subset \mathbf{D}(M_Y, \mathbf{S})$ となる関係の評価軸ペアを，Y は X に依存する関係とみなして「階層的依存関係にある評価軸ペア」を抽出する．

表 1 実験に用いたデータセットの概要

データセット	会話形式	評価対象	評価サンプル数	評価者数	対話システム数
DSTC9	インタラクティブ会話	会話全体	2200	3	11
PersonaChat (USR)	既定会話文脈への応答	最終応答	300	3	5
Topical-Chat (USR)	既定会話文脈への応答	最終応答	360	3	6

3.2 評価の階層的依存性の解消

本研究では§3.1で説明した手法により可視化された評価軸間の階層関係が，個々の評価にバイアスを与え対話システム間の性能有意差の有無に影響を与えるかを調べる．そこで，評価軸間の評価値を無相関化することで評価軸間の階層的依存性を解消する方法を検討する．

無相関化処理を行う評価値列ペアは依存関係にあるため，無相関化により独立した評価値が得られるかを確認する必要がある．ここでは評価値の無相関化により評価軸間の階層的依存性が解消される過程を説明する．依存関係にある評価のうち，上位・下位の評価を高次評価・低次評価と呼ぶこととする．階層的依存性が解消された高次評価 $y \in \mathbb{R}$，低次評価 $x \in \mathbb{R}$ が独立であり，観測される高次評価を $\hat{y} \in \mathbb{R}$ とする．評価の階層性を用いる自動評価手法 [6] と同様に評価軸間の関係が線形回帰モデルにより表現可能であるとして，$\hat{y}$ が係数 $a, b, c \in \mathbb{R}$ で表現される以下の式 $f(x, y)$ に近似可能であると仮定する．

$$f(x, y) = ax + by + cxy$$

別の評価値 $\hat{y}'$ は，x, y の差分 $\Delta x, \Delta y$ を用いて

$$\hat{y}' = f(x + \Delta x, y + \Delta y)$$

と表せる．このときの $\hat{y}$ と $\hat{y}'$ 差分 $\Delta\hat{y}$ は以下のように展開される．

$$\Delta\hat{y} = (a + cy)\Delta x + (b + cx)\Delta y \tag{1}$$

観測される高次評価 $\hat{y}$ に対して，x に関する無相関化 $\frac{\delta\hat{y}}{\delta x} \to 0$ を行うことで，$a + cy$ が十分に小さな値とみなせるため，x で無相関化した $\Delta\hat{y}$ は低次評価の差分 Δx に依らない評価値 $(b + cx)\Delta y$ で算出される．

4 実験

4.1 データセット

データセットには，Yeh らが公開する [4]，DSTC9 [8] と USR [7] のデータセットを用いた．表 1 に，データセットの概要を記載する．

DSTC9 [8] このデータセットは DSTC9 コンペティションに参加した複数の対話システムを人手で絶対評価したデータセットである．本研究では 11 の対話システムとのインタラクティブな会話に対する人手評価のみを取り扱う．この評価では，システムとの会話全体への評価として，以下の 11 の評価軸で評価を行う．評価軸とその評価値範囲は Consistent([0,1]), Likable([1,3]), Diverse([1,3]), Informative([1,3]), Coherent([1,3]), Overall([1,5]), Understanding([1,3]), Flexible([1,3]), Topic Depth([1,3]), Inquisitive([1,3]) である．

USR データセット [7] このデータセットは，自動評価手法 USR [7] のメタ評価に利用されたデータセットである．PersonaChat [14] と Topical-Chat [15] の 2 つのコーパスから構成されている．この評価データセットは，事前に設定された会話文脈への応答文に対する評価である．PersonaChat コーパスでは Seq2Seq model，LSTM language model，Key-Value Profile Memory Network model により生成された応答文と人により記述された 2 種類の応答文を人手で絶対評価する．Topical-Chat コーパスではデコーディング機構のハイパーパラメタが異なる 5 つの Transformer model と人が記述した応答文の計 6 応答を人手で絶対評価する．評価軸とその評価値範囲は Understandable([0,1]), Natural([1,3]), Maintain Context([1,3]), Engaging([1,3]), Uses Knowledge([0,1]), Overall([1,5]) である．

4.2 結果

表 2 に，抽出された階層的依存関係にある評価軸ペアと，有意差ペア数を表記する．またカッコ内の数字は§3.2 での方法により階層的依存性を解消した場合の検出ペア数の変化を示す．まず，抽出された階層的依存関係にある評価軸ペアは Natural-Maintain Context, Understandable-Natural など，直感的に階層関係となりそうな組み合わせが抽出されたことが確認できる．

階層的依存性を解消した場合，USR のデータセットではほとんどの場合で検出ペア数が変化しないの

表 2 各データセットで抽出された低次評価と高次評価の組み合わせ．評価軸ペアごとに，両評価軸で共通に検出された有意差ペア数，高次評価のみので検出された有意差ペア数を記載する．カッコ内は，評価軸ペアを無相関化後の分析での有意差ペアの変化数を示す．また，無相関化した高次評価で新しく検出された有意差ペアを最右列に示す．なお，Topical Chat (USR) データセットでは Maintain Context と Engaging が，Natural の評価軸と全く同じ結果のため記載を省略した．

低次評価	高次評価	共通の有意差ペア	高次評価のみの有意差ペア	新規の有意差ペア
		DSTC9		
Diverse	Likeable	16 (±0)	13 (-5)	(+0)
Diverse	Coherent	16 (±0)	15 (-3)	(+0)
Diverse	Overall	16 (±0)	16 (±0)	(+0)
Diverse	Understanding	16 (-2)	13 (-7)	(+2)
Diverse	Flexible	16 (±0)	12 (-2)	(+0)
Diverse	Inquisitive	16 (-12)	4 (-4)	(+0)
Inquisitive	Flexible	20 (±0)	8 (-3)	(+0)
		PersonaChat (USR)		
Understandable	Maintain Context	4 (±0)	3 (±0)	(+0)
Understandable	Overall	4 (±0)	4 (±0)	(+0)
Natural	Maintain Context	5 (±0)	2 (±0)	(+0)
Natural	Engaging	5 (±0)	3 (±0)	(+0)
Natural	Overall	5 (±0)	3 (±0)	(+0)
Maintains Context	Overall	7 (±0)	1 (±0)	(+0)
		Topical-Chat (USR)		
Understandable	Natural	8 (±0)	1 (±0)	(+0)
Understandable	Uses Knowledge	6 (-2)	1 (±0)	(+0)
Understandable	Overall	8 (±0)	2 (±0)	(+0)
Natural	Overall	9 (±0)	1 (±0)	(+0)
Uses Knowledge	Overall	9 (±0)	1 (±0)	(+0)

に対し，DSTC9 ではすべての評価軸ペアで有意差ペア数の変化が起こっている．このことから DSTC9 データセットでは高次評価での性能差ではなく低次評価での性能差が原因で，高次評価上で対話システム間性能有意差が検知されたペアが存在することが確認された．

4.3 考察

表 2 に関して，式 (1) と関連性を考察する．評価軸ペアで共通の有意差ペアは，低次評価の差分 Δx により対話システム間性能有意差が検出されるため，高次評価のみの有意差ペアは Δx が相対的に小さいと考えられる．このため，無相関化により検出されなくなった有意差ペアに関して，高次評価のみの有意差ペアでは評価軸ペアで共通の有意差ペアと比較すると $a + cy$ の項の寄与が大きく，無相関化によるペア数の減少率が高次評価のみの有意差ペアで大きいのは，妥当な結果だと考えられる．また，現実的には無相関化後も $a + cy$ の項は完全には無視できないが，無相関化した高次評価のみで検出される有意差ペアは，Δx が相対的に小さいため，他の有意差ペアと比較して Δy が大きいと推測される．

5 結論

本研究では，評価の階層的依存性やそれに起因する問題について分析を行った．この分析にあたり性能有意差のある対話システムペアの集合を利用して，評価軸間の階層的依存関係を網羅的に分析し，階層的依存関係にある評価軸ペアを明らかにした．さらに階層的依存関係を解消した場合の対話システム間性能有意差の再分析を行った．DSTC9 と USR データセットを利用した分析から，階層的依存性の解消により対話システム間の性能有意差の有無が変化しうることを確認した．

今後は，既存の自動評価手法について評価軸間の階層的依存関係との関係について分析を行う．

謝辞

本研究は東京大学生産技術研究所特別研究経費および JSPS 科研費 JP21H03494 の助成を受けています.

参考文献

[1] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9230–9240, Online, November 2020. Association for Computational Linguistics.

[3] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 7789–7796, Apr. 2020.

[4] Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. In Wei Wei, Bo Dai, Tuo Zhao, Lihong Li, Diyi Yang, Yun-Nung Chen, Y-Lan Boureau, Asli Celikyilmaz, Alborz Geramifard, Aman Ahuja, and Haoming Jiang, editors, **The First Workshop on Evaluations and Assessments of Neural Conversation Systems**, pp. 15–33, Online, November 2021. Association for Computational Linguistics.

[5] Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. The glass ceiling of automatic evaluation in natural language generation. In **Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing (Findings Papers)**, pp. 178–183, Bali, Indonesia, November 2023. Association for Computational Linguistics.

[6] Vitou Phy, Yang Zhao, and Akiko Aizawa. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4164–4178, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[7] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[8] Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. Interactive evaluation of dialog track at DSTC9. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5731–5738, Marseille, France, June 2022. European Language Resources Association.

[9] Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, **Proceedings of the 4th Workshop on NLP for Conversational AI**, pp. 77–97, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Shiki Sato, Yosuke Kishinami, Hiroaki Sugiyama, Reina Akama, Ryoko Tokuhisa, and Jun Suzuki. Bipartite-play dialogue collection for practical automatic evaluation of dialogue systems. In Yan Hanqi, Yang Zonghan, Sebastian Ruder, and Wan Xiaojun, editors, **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop**, pp. 8–16, Online, November 2022. Association for Computational Linguistics.

[11] Sijia Liu, Patrick Lange, Behnam Hedayatnia, Alexandros Papangelis, Di Jin, Andrew Wirth, Yang Liu, and Dilek Hakkani-Tur. Towards credible human evaluation of open-domain dialog systems using interactive setup. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, No. 11, pp. 13264–13272, Jun. 2023.

[12] Sarah E. Finch and Jinho D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 236–245, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[13] Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. Achieving reliable human assessment of open-domain dialogue systems. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6416–6437, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[15] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In **Proc. Interspeech 2019**, pp. 1891–1895, 2019.

A　評価軸間の相関値

各データセットの評価軸間の相関を図 2, 3, 4 に示す．相関値は，Persona r により計算した．DSTC9 では，ほとんどの評価軸間の相関値が 0.5 を上回っており，全体的に 0.6〜0.7 程度の相関と相関値が高いことが確認できる．一方で，PersonaChat (USR) での評価軸間の相関は 0 に近いものもある一方で 0.7 と高い相関値もあり，評価軸間で様々である．Topical-Chat (USR) では Use Knowledge の評価軸以外の評価軸間の相関値が 0.5 を超えており，またすべてのデータセットで最大の相関値 0.85 が確認できる．

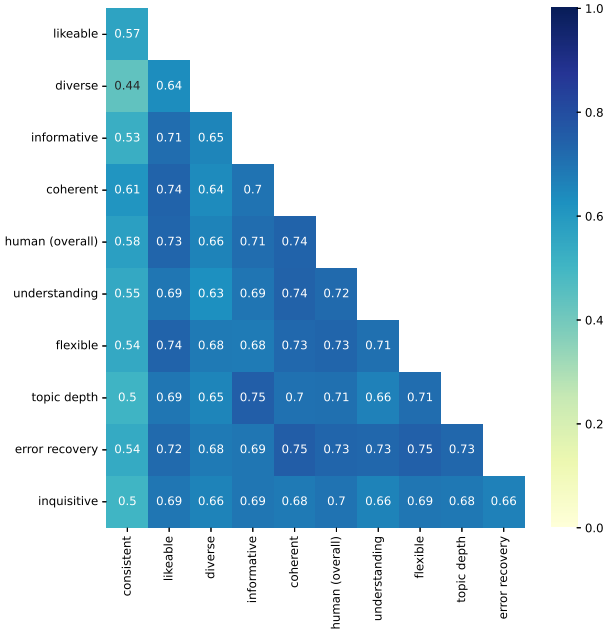

図 2　DSTC9 での評価軸間の相関値

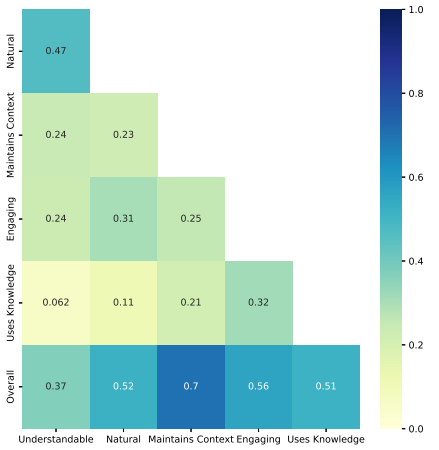

図 3　PersonaChat (USR) での評価軸間の相関値

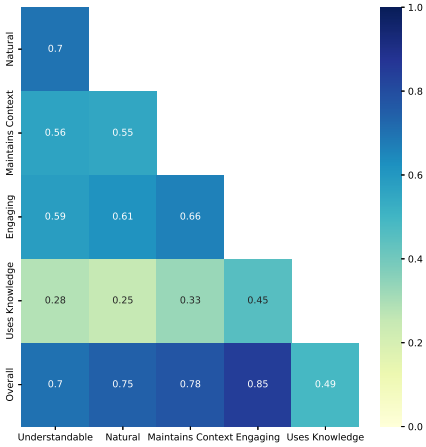

図 4　Topical-Chat (USR) での評価軸間の相関値