大規模言語モデルを用いた対話品質評価に関する調査

赤間 怜奈　鈴木 潤
東北大学　理化学研究所
{akama, jun.suzuki}@tohoku.ac.jp

概要

高度な言語理解能力に加えて人間的な感性をも備えつつある高性能大規模言語モデルを、人手評価の代替手段として活用することに関する議論が近年広がりを見せている。しかし、大規模言語モデルによる人手評価の代替については、いまだ現行技術の限界や致命的な課題など未解明の側面も多く、言語やタスク横断的に広く知見を収集することが重要な段階にある。本研究は、日本語を対象として、対話データの品質評価における大規模言語モデルの活用について経験的な知見を提供するものである。評価軸や回答形式が異なる複数の評価設定において、モデルの評価性能や動作傾向がどのように変動するかを定性的および定量的に調査し、結果を報告する。

1　はじめに

自然言語処理分野における大規模言語モデルの急速な発展は、多くの研究領域に革新をもたらしている。膨大な知識と高い言語理解能力、論理的思考力を備えた大規模言語モデルは、分野横断的なタスクにより測定される総合的な問題解決能力において、人間の専門家に匹敵する水準に達しつつある [1, 2]。

高性能な大規模言語モデルの活用先として近年大きな注目を集めているのが「評価者」としての役割である [3]。高度な言語理解と人間的な感性に基づく判断が必要ゆえに従来は人手評価が強く推奨されてきた文生成タスクの評価においては、大規模言語モデルの活用により低コストかつ実用的な代替手段が提供される可能性があり、とくに高い関心が寄せられている [4, 5, 6]。人間の作業をより高精度に代替する方法論の開発 [7, 8] が進む一方で、大規模言語モデルによる人手評価の代替については現行技術の限界や致命的な課題など未解明の側面も多く、言語やタスク横断的に経験的知見を広く収集することは引き続き重要な取り組みの方向性となる [9, 10]。

本研究は、日本語を対象として、対話データの品質評価における大規模言語モデルの活用について経験的な知見を収集することを目的とするものである。対象とする対話品質評価タスクの実用的な応用先としては、学習データ高品質化のための対話データフィルタリングなどが挙げられる。実験では、評価軸や回答形式が異なる複数の評価設定において、モデルの評価性能や動作傾向がどのように変動するかを定性的および定量的に調査し、その結果を報告する。調査の結果、モデルとデータセットを固定し評価設定のみを変更した場合、モデルの評価性能には順位相関係数で最大 0.9 ポイント超の差が生じることが明らかとなった。また、評価スコアとして用いる数字そのものにモデルがバイアスを持っており評価結果がその影響を受けている可能性があることも示唆された。実際、この影響を回避した評価設定においては評価性能が改善されることを確認した。

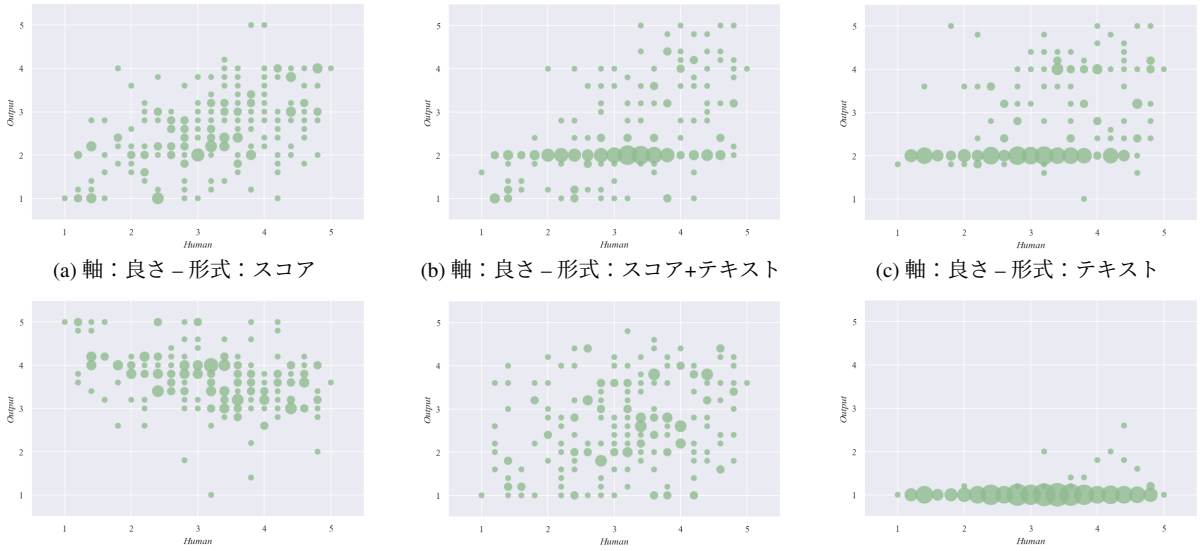
2　タスク：対話品質評価

入出力　n 組の発話–応答ペア集合を対話データ $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ とおく。本研究で対象とする対話品質評価タスクは、各発話–応答ペア $(x_i, y_i) \in \mathcal{D}$ を入力として受け取り、与えられた指示（評価軸、回答形式）の下で発話–応答ペアの品質（評価結果）を出力する形式をとる。データセット（後述）の特性上、$\mathcal{D}$ には低品質な発話–応答ペアが多く含まれている。

データセット　評価対象の対話データ $\mathcal{D}$ として大規模映画字幕コーパス OpenSubtitles [11] から作成した日本語発話–応答ペア集合（$n = 200$）を用いる。各ペアには、5 人の日本語母語話者が「連続する 2 発話が対話として許容できるか」の問いに 5 段階リッカート尺度を用いて評価したスコア[1)]が付与されている。なお、OpenSubtitles には話者情報が付与されておらず、慣習的に「字幕 1 行が 1 発話に相当する」という仮定のもと連続する 2 行を発話–応答ペアと見做して抽出する方法が用いられている [12]。当然ながら、この方法で獲得した対話デー

1）Yahoo!クラウドソーシングを利用

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 軸：良さ – 形式：スコア　(b) 軸：良さ – 形式：スコア+テキスト　(c) 軸：良さ – 形式：テキスト

(d) 軸：悪さ – 形式：スコア　(e) 軸：悪さ – 形式：スコア+テキスト　(f) 軸：悪さ – 形式：テキスト

図 1　出力されたスコア（縦軸）と人手評価スコア（横軸）の関係。左下が低評価、右上が高評価

タには対話として許容できない低品質な発話–応答ペア[2)]が含まれることが指摘されている [13, 14, 15]。

3　実験

日本語の対話品質評価における大規模言語モデルの振る舞いを観察する。手始めに、現在最も身近な高性能言語モデルのひとつといえる ChatGPT を用いて（3.1 節）、評価方法を指示する際にいくつかの観点でバリエーションを生じさせた（3.2 節）ときの評価性能ならびに動作傾向の変化を調査した。

3.1　モデル設定

本研究では、GPT-3.5 Turbo[3)]（OpenAI ChatGPT API[4)]）を観察対象の大規模言語モデルとした。生成の安定性を保持しつつ多様性も実現するために、生成時には 5 通りの温度パラメータ $T = \{0.92, 0.94, 0.96, 0.98, 1.00\}$ を用いた。つまり、全ての試行において、1 つの入力に対して 5 通りの生成結果を出力として獲得している。

3.2　評価設定：指示のバリエーション

評価時には、評価対象の対話データとともに評価方法を指示するための記述文（プロンプト）をモデルへの入力として与えた。このとき、評価の軸（全 2 通り）と回答形式（全 3 通り）についてそれぞれ異なる全ての組み合わせを指示することにより、全 6 通りの設定でモデルの評価結果を収集した。

評価軸　与えられた発話–応答ペアについて、対話としての品質が (1) どの程度「良い」か、または (2) どの程度「悪い」かのいずれかで評価するよう指示した。いずれも 5 段階リッカート尺度とした。

回答形式　5 段階リッカート尺度での評価結果は、(1) スコアのみ「1〜5」、(2) スコア+テキスト「1: 強く同意しない〜5: 強く同意する」、(3) テキストのみ「強く同意しない〜強く同意する」のいずれかで出力するよう指示した。このとき、具体的なスコアやテキストでの回答に加えて、その評価の根拠（判断理由）も合わせて出力するよう指示した。

表 1　人手評価スコアとの相関（Spearman's ρ）

設定（軸 – 形式）	出力スコア	根拠をスコア化
良さ – スコア	0.4877	**0.5570**
良さ – スコア+テキスト	**0.5072**	0.5251
良さ – テキスト	0.3898	0.4010
悪さ – スコア	–0.3778	0.4662
悪さ – スコア+テキスト	0.2568	0.2781
悪さ – テキスト	0.1867	0.1867
Akama [15]	0.3751	-
Junczys-dowmunt [16]	0.2973	-

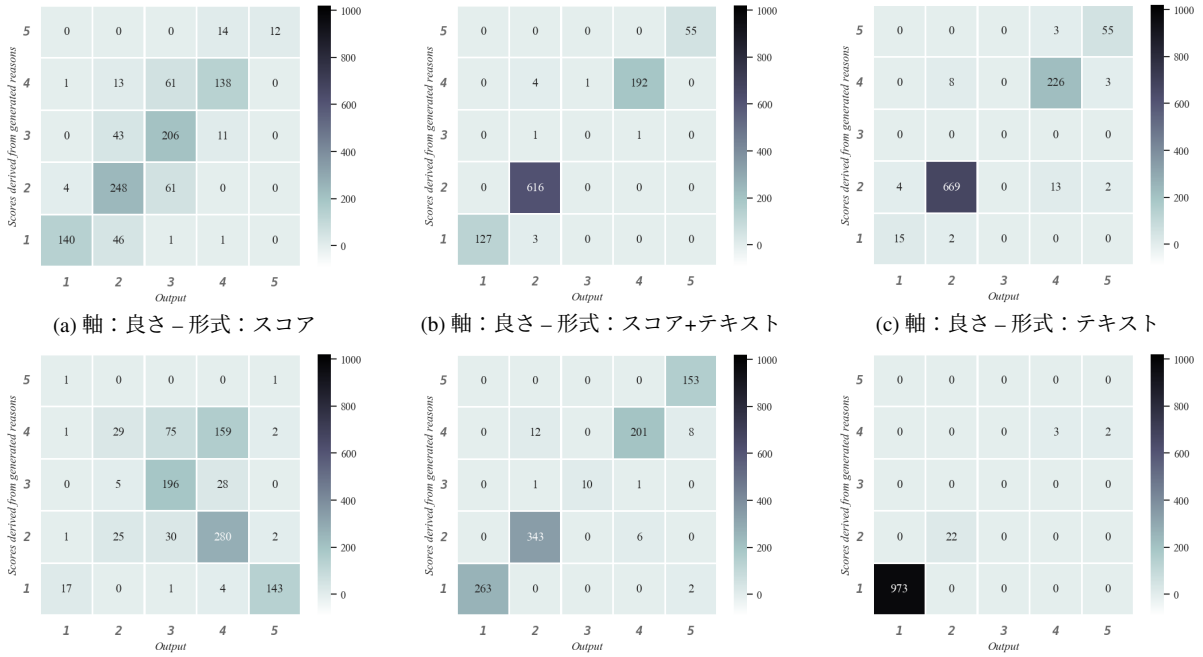
4　実験結果

4.1　評価性能：人手評価との相関

各設定におけるモデルの対話品質評価の性能を、モデルが出力した評価と人間の評価者による人手評価の相関により定量化する。表 1（左列）に、各設

2)　シーンや回想を跨ぐ無関係の 2 発話がペアとして繋がる等
3)　`gpt-3.5-turbo-0301` モデル
4)　https://openai.com/index/openai-api

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 軸：良さ – 形式：スコア　(b) 軸：良さ – 形式：スコア+テキスト　(c) 軸：良さ – 形式：テキスト

(d) 軸：悪さ – 形式：スコア　(e) 軸：悪さ – 形式：スコア+テキスト　(f) 軸：悪さ – 形式：テキスト

図 2　実際に出力されたスコア（横軸）に対する根拠に基づくスコア（縦軸）の頻度分布

定での各発話–応答ペアに対してモデルが出力した評価スコアと人手評価スコアとの Spearman の順位相関係数 [17] ρ の値を示す。このとき、モデルが出力した全ての評価は、1 を低評価〜5 を高評価とした共通の評価スコアに正規化されている。また、各ペアに対する最終的な評価スコアは人手・モデルの出力ともに 5 つのスコアの平均値とした。

最も相関が高かったのは「品質の良さを、スコアとテキストの両方で回答する」設定であった（$\rho = 0.5072$）。これは、既存の自動品質評価手法を用いた場合（表 1 下部、Akama [15] および Junczys-dowmunt [16]）よりも 0.1 ポイント以上高い値であった。反対に、最も相関が低かったのは「品質の悪さを、スコアのみで回答する」設定であった。本実験では、評価設定の異なりがモデルの評価性能に順位相関係数にして最大で 0.9 ポイント超の差を生じさせるという結果が得られた。

4.2　評価の傾向

図 1 に、各設定でモデルが出力したスコアの人手評価スコアに対する分布をそれぞれ示す。円の大きさは頻度を表す。いずれの評価軸でも共通して回答形式が「テキストのみ」の場合（図 1(c) および (f)）、図中の点が横方向に集合していることから、モデルの出力が特定のスコアに偏っていることがわかる。この傾向は、回答形式が「スコアのみ」の場合には観測されず、回答形式が「スコアとテキスト」の場合は評価軸が「良さ」の場合に中程度に観測された。

表 2　同一方法内でのスコア一貫性（Cronbach's α）

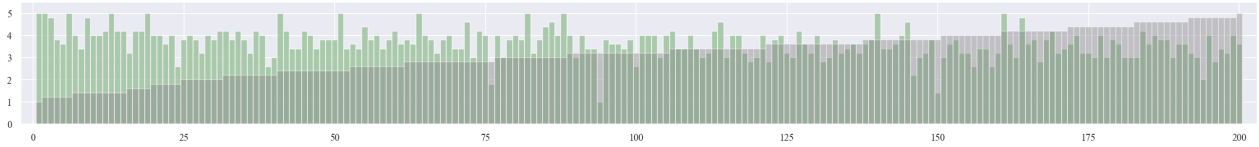
設定（軸 – 形式）	出力スコア	根拠をスコア化
良さ – スコア	0.9354	0.8783
良さ – スコア+テキスト	**0.9450**	**0.9412**
良さ – テキスト	0.9393	0.9272
悪さ – スコア	0.8025	0.7599
悪さ – スコア+テキスト	0.7054	0.7002
悪さ – テキスト	0.6352	0.6393
人間（人手評価）	0.7239	-

4.3　信頼性：評価の一貫性

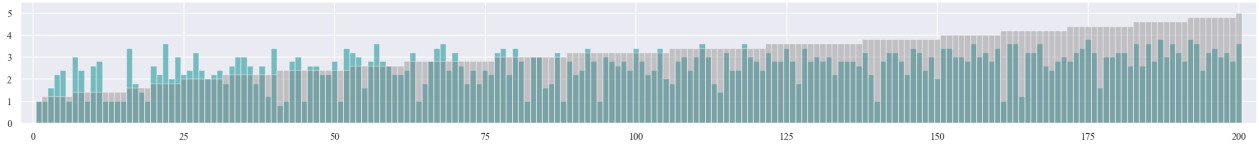
各設定におけるモデルの評価の信頼性を、評価スコアの一貫性により定量化する。表 2（左列）に、各設定で同一サンプルに対して出力された 5 つの評価スコアに基づいて算出した Cronbach の α 係数 [18] の値を示す。Cronbach の α 係数は、リッカート尺度に基づく複数人評価の内的整合性を測定するために用いられる方法のひとつで、自然言語処理分野でもいくつかの既存研究で使用されている [19, 20, 21]。

最も高い値となったのは「品質の良さを、スコアとテキストの両方で回答する」設定であった（$\alpha = 0.9450$）が、「品質の良さ」を評価軸とする場合はいずれも $\alpha = 0.9$ を超える高い値であった。最も低い値となったのは「品質の悪さを、テキストのみ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 出力されたスコア（黄緑）と人手評価スコア（グレー）



(b) 生成された根拠に基づくスコア（緑）と人手評価スコア（グレー）

図 3　各サンプルの評価スコア。評価設定「軸：悪さ – 形式：スコア+テキスト」の場合

で回答する」設定で（$\alpha = 0.6325$）、人間による評価の一貫性（$\alpha = 0.7239$）を約 0.1 ポイント、前述の最高値を約 0.2 ポイント下回る結果となった。

5　議論：スコアの選定に課題か

本実験では、モデルは 5 段階リッカート尺度評価に加えて、その評価の根拠となる評価理由も出力している（3.2 節参照）。生成された根拠の妥当性・有用性を調査するために、モデルが出力として生成した文のうち根拠に該当する部分のみを抽出し、その内容に対応すると考えられる評価スコアを再度人手で付与し直したもの（根拠に基づくスコア）と、モデルが実際に出力したスコアの一致を調査した。

5.1　出力スコアと根拠の乖離

図 2 に、各設定でのモデルが実際に出力したスコアに対する根拠に基づくスコアの頻度分布を示す。双方のスコアが完全に一致している場合は左下–右上方向の対角線上に位置するセルのみが 1 以上の値で他は 0 となるが、全ての設定でその分布にはならなかった。特筆すべきは図 2(d) の「品質の悪さを、スコアのみで回答する」場合で、根拠として生成された内容は評価スコア 2（どちらかというと低評価）を示唆するものだが実際に生成された評価スコアは 4（どちらかというと高評価）という事例が最多の 280 サンプル確認される等、根拠に対してスコアが “反転” している様子が頻繁に観察された。このことから、少なくとも本実験においてモデルは「大きい数字は “良い” 状態を表す」というバイアスを持っている可能性があることが考えられる。[5)]

5.2　根拠に基づくスコアによる性能改善

根拠に基づくスコアについても評価性能と信頼性（4.1, 4.3 節参照）を定量化した。前者について Spearman の順位相関係数 ρ の値を表 1 の右列に、後者について Cronbach の α 係数を表 2 の右列にそれぞれ示す。モデルが出力したスコアを採用した場合と比較して、全ての設定で評価性能が向上した。とくに ρ の値が大きく向上したのは「品質の悪さを、スコアのみで回答する」場合で、各サンプルに対する評価スコア（図 3）を見ると人手評価で低評価のサンプル（図の左側）に対する評価性能が大きく改善されていることが確認できた。一貫性は回答形式が「スコア」の場合はわずかに低下、他はほとんど変化なしという結果であった。

6　おわりに

本研究では、日本語の対話品質評価における大規模言語モデルの活用について経験的な知見を収集することを目的として、評価軸や回答形式が異なる複数の評価設定におけるモデルの評価性能や動作傾向の変動を定性的および定量的に調査した。モデルとデータセットを固定し評価設定のみ変更した場合でも評価性能に大きな差が生じること、モデルが特定のシンボルにバイアスを有しており評価結果が影響を受ける可能性があることなどが明らかとなった。これらの知見に着想を得て、大規模言語モデルを活用した実用的な評価方法に関するいくつかのアイディアがある。これらのアイディアの具体化と有用性の実証を今後の課題とする。

5）　経験的に、実世界では「大きい数字は “良い” 状態を表す」設定が多い。これが学習データ量の差として反映されているとすれば、このようなバイアスが存在することと辻褄は合う

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究の一部は、JSPS 科研費 JP22K17943 および JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の支援を受けたものです。

参考文献

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In **International Conference on Learning Representations**, 2021.

[2] Aakanksha Chowdhery and et al. PaLM: Scaling Language Modeling with Pathways. **Journal of Machine Learning Research**, Vol. 24, No. 240, pp. 1–113, 2023.

[3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[4] Tom Kocmi and Christian Federmann. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 193–203, 2023.

[5] Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluations? In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 15607–15631, 2023.

[6] Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. Leveraging Large Language Models for NLG Evaluation: Advances and Challenges. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 16028–16045, 2024.

[7] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 13484–13508, 2023.

[8] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[9] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT Outperforms Crowd-Workers for Text-annotation Tasks. **Proceedings of the National Academy of Sciences**, Vol. 120, No. 30, 2023.

[10] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 9440–9450, 2024.

[11] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, pp. 1742–1748, 2018.

[12] Oriol Vinyals and Quoc Le. A Neural Conversational Model. In **Proceedings of the 32nd International Conference on Machine Learning Deep Learning Workshop**, 2015.

[13] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, 2016.

[14] Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3970–3980, 2018.

[15] Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 941–958, 2020.

[16] Marcin Junczys-Dowmunt. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 888–895, 2018.

[17] C. Spearman. The Proof and Measurement of Association between Two Things, volume = 15, year = 1904. **The American Journal of Psychology**, No. 1, pp. 72–101.

[18] Lee J. Cronbach. Coefficient Alpha and the Internal Structure of Tests. **Psychometrika**, Vol. 16, No. 3, p. 297–334, 1951.

[19] Caroline Langlet and Chloé Clavel. Improving social relationships in face-to-face human-agent interactions: when the agent wants to know user‘s likes and dislikes. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing**, pp. 1064–1073, 2015.

[20] Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. Improving the Annotation of Sentence Specificity. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation**, pp. 3921–3927, 2016.

[21] Yupei Du, Qixiang Fang, and Dong Nguyen. Assessing the Reliability of Word Embedding Gender Bias Measures. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10012–10034, 2021.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).