キャッチコピー共同作成対話コーパスにおける第三者評価と自己評価の関係分析

周旭琳　市川拓茉　東中竜一郎
名古屋大学大学院情報学研究科
{zhou.xulin.j3@s.mail,ichikawa.takuma.w0@s.mail,higashinaka@i}.nagoya-u.ac.jp

概要

我々は，先行研究において，人間同士が対話しながらキャッチコピーを共同で作成する対話システムの構築を目的としてキャッチコピー共同作成対話コーパスを構築してきた．しかし，作成されたキャッチコピーについて，第三者による評価が行われておらず，共同作業の内容と作成された成果物の質の関係が不明確であった．そこで，本研究ではコーパスに含まれるキャッチコピーに対して第三者評価を付与し，共同作業の内容と作成された成果物の質の関係を明らかにし，それを踏まえ，キャッチコピーを共同で作成する対話システムが持つべき指針について考察する．

1　はじめに

対話システムのシステムが進展し，社会に広まるにつれて，人間と共同作業する対話システムの構築が盛んになってきた [1, 2, 3]．しかし，その多くは問題解決のための対話 [4, 5] やユーザからの命令にシステムが従うシステム [6, 7, 8, 9] が中心であり，ユーザと創造的な共同作業を行うシステムに関する研究はいまだ少ない．

我々はこれまで，共同作業が可能な対話システムの構築を目指し，そのための基礎データとして共同作業を行う際の人間のコーパスを収集してきた．具体的には，対話をしながら共同でキャッチコピーの作成を行う対話からなるコーパスを収集した [10, 11, 12]．さらに，人間同士の共同作業についての知見を得るため，このコーパスについて，発話とキャッチコピー編集の時系列に着目した分析や，発話やキャッチコピーの間での参照表現の分析などを行った．

しかし，作成されたキャッチコピーについて，第三者による評価が行われておらず，共同作業の内容と作成された成果物の質の関係が不明確であった．そこで，本稿では，作成されたキャッチコピーに対する第三者評価の収集を行い，共同作業の内容と作成された成果物の質の関係を明らかにする．

2　使用したコーパスの概要

我々は，創造的な共同作業としてキャッチコピー共同作成タスクを設定し，人間同士の共同作業のコーパス（キャッチコピー共同作成対話コーパス）を構築した [10, 11]．キャッチコピー共同作成タスクでは，提示される両作業者共通の商品説明を参考に，テキストチャットを用いた対話を行いながら，作業相手とテキストボックスを共同編集しキャッチコピーを作成する．付録における表 5，表 6 はコーパスから抜粋した対話例である．これらの対話例では，それぞれ，メロン，小型受球ネットについてキャッチコピーを作成している．

本コーパスには，のべ 105 人のクラウドワーカによって実施された，782 対話が含まれている．一対話の制限時間は 30 分である．作業者にはキャッチコピーの対象となる商品の商品説明を提示し，二人で合計 3 つ以上のキャッチコピーを作成するよう教示した．

作業者のインタフェースには，8 つのテキストボックスからなるキャッチコピー編集欄を設け，両作業者が共有・編集しあうことを可能にした．作業者が特定のテキストボックスを示しやすいよう，テキストボックスの左には，それぞれ A〜H のアルファベットのラベルを付けた．テキストボックス内に文字を入力すると，その変更が作業相手の見ている画面にも即座に反映され，また，編集が記録された．

作業後には作業自体や作成したキャッチコピーについての自己評価をアンケートにより調査した．このアンケートでは 5 段階のリッカート尺度を利用し

表 1　キャッチコピーに対する第三者評価 (1-10 の 10 段階)．一致率は Fleiss' Kappa を表す．

質問項目	平均	標準偏差	一致率
Q8-2.（興味）このキャッチコピーは目にした人の興味を引くと思いますか	4.81	2.40	0.019
Q9-2.（想像）このキャッチコピーは見る人の想像を膨らませることができると思いますか	4.70	2.51	0.024

表 2　キャッチコピーとそれに対する第三者評価の例 (ave は平均，sd は標準偏差を示す)

商品	キャッチコピー	第三者評価	ave	sd
小型受球ネット	限りなく打ち込め！球拾いゼロ、時短練習でライバルに差をつける	（興味）10,6,9,9,9	8.6	1.36
メロン	糖度 18 度以上、異次元の美味さ。	（想像）6,10,9,10,9	8.8	1.47
フェンス	あなたの心は閉ざしません	（興味）1,1,1,1,1	1.0	0.0
コンシーラー	ナチュらる？	（想像）1,1,1,1,1	1.0	0.0
シルバーカー	最近、背筋伸びたみたい	（興味）1,8,10,3,3	5.0	3.41
模型飛行機	コレクターも納得の本格戦闘機航空機ダイキャスト合金モデル	（想像）10,8,7,1,4	6.0	3.16

た．アンケートの質問項目には，「Q1.（自分主張）今回の共同作業では，あなたの考えや意見を主張することができましたか」，「Q2.（相手主張）今回の共同作業では，作業相手の方は意見や考えを主張していましたか」などに加え，キャッチコピーの内容に関する，「Q8.（興味）今回の共同作業で二人で作成したキャッチコピーは，目にした人の興味を引くと思いますか」，「Q9.（想像）今回の共同作業で二人で作成したキャッチコピーは，見る人の想像を膨らませることができると思いますか」が含まれる．キャッチコピーに関する質問項目の作成にあたっては，先行研究 [13] を参考にした．また，完成していると考えているキャッチコピー編集欄がどれか，作成した中で一番良いと考えているキャッチコピーがどれかについても収集した．

3　キャッチコピーの第三者評価

キャッチコピー共同作成対話コーパス内に含まれるキャッチコピーについて，第三者評価を収集した．

3.1　キャッチコピーの第三者評価の収集

クラウドソーシングを用い 5 人の評価者を集め，キャッチコピーに対する第三者評価を行った．5 人はそれぞれが全てのキャッチコピーを評価した．評価者には各キャッチコピー作成対象商品の商品名，商品説明，商品画像，その商品について作成されたキャッチコピーを提示した．評価項目は，コーパス構築時に収集したアンケートに含まれるキャッチコピーに関連した質問と同じ内容を問う表 1 の 2 項目（Q8-2, Q9-2）とした．コーパス構築時のアンケートでは 1 作業で作成したすべてのキャッチコピーについて一つの評価値で評価しているのに対し，第三者評価では 1 作業で作成された複数のキャッチコピーのそれぞれに対し個別にスコアを付与していることに注意されたい．キャッチコピー編集欄に書き込まれている文字列は完成したキャッチコピーだけでなく途中段階のものも含まれるため，評価対象とするキャッチコピーは，各作業の作業後アンケートで少なくとも 1 人の作業者に完成しているとみなされたキャッチコピーとした．その結果，評価対象となるキャッチコピーの合計数は 5,365 個であった．なお，より粒度を細かくすることで詳細な評価ができると考えたため，自己評価で用いた 5 段階でなく，1-10 の 10 段階で評価を行った．その際，評価者には評価値の最小が 1，最大が 10 となるように評価基準を定め，全体を一貫した基準で評価するよう教示した．また，その評価基準をテキストで回答させた．キャッチコピーの表示順は評価者ごとに異なるようランダムとした．

3.2　第三者評価の収集結果

表 1 はキャッチコピーに対する第三者評価の結果である．Q8-2（興味）の平均は 4.81，Q9-2（想像）の平均は 4.70 となった．Fleiss' Kappa の値は Q8-2（興味）が 0.019，Q9-2（想像）が 0.024 となり，評価者間の一致率は低かった．

第三者評価が高いものは，キャッチコピー作成者の自己評価も高い可能性がある．そこで，キャッチコピーに対する第三者評価と作成者の自己評価の相関の有無を調べるために，自己評価アンケートにおける作成した中で一番良いとされたかどうかとキャッチコピーに対する第三者評価の合計値との相関比を算出した．相関比の値はどちらの評価項目についてもほぼ 0 であった．また，それぞれの評価項目について，各対話の第三者評価の平均と自己評価の平均についてのピアソンの相関係数も算出した，どちらの評価項目についても相関係数はほぼ 0 で

表 3 各クラスタのキャッチコピーに関するスコアの平均．括弧内の数値はそれぞれのクラスタに分類された作業の割合 (%) を示す．太字は各質問における最大値．各質問において，上位 2 つのスコアについて下線を引いている．上付き文字は Steel-Dwass 検定でその番号のクラスタよりも有意に値が高いことを示す（$p < 0.05$）．

質問項目		クラスタ 1 (14.3%)	2 (35.9%)	3 (12.3%)	4 (11.6%)	5 (14.1%)	6 (11.8%)
自己評価（1 − 5）	Q8. 興味	4.13	4.23[5]	4.09	**4.24**[5]	4.05	4.05
	Q9. 想像	4.11	4.19	4.03	**4.22**	4.08	4.09
第三者評価（1 − 10）	Q8-2. 興味	**4.92**[4]	4.78	4.83	4.71	4.84	4.79
	Q9-2. 想像	**4.91**[245]	4.64	4.73	4.62	4.63	4.72

あった．

表 2 はキャッチコピーとその評価の例である．評価者の記述した評価基準には，「商品の良いところがわかりやすく伝わってくるか、商品を買うことによって得られる良い点が的確に表現されているか、といった点を重視して評価を行いました。」「日本語的な表現が正しくないものは低評価としています」といった表現が見られた．「限りなく打ち込め！球拾いゼロ、時短練習でライバルに差をつける」「糖度 18 度以上、異次元の美味さ。」は良いところがわかりやすい表現のため，評価が高くなったと考えられる．「あなたの心は閉ざしません」に関して述べると，フェンスは囲って閉ざすものだが，このフェンスはおしゃれで趣味を反映させられるため心は閉ざさない，という連想が必要だと推測され，評価が低くなったと考えられる．「ナチュらる？」は，「ナチュラル」をひねっており，日本語的な表現に違和感を持たれた可能性がある．「最近、背筋伸びたみたい」は直接商品の効果につながるわけではないため，「コレクターも納得の本格戦闘機航空機ダイキャスト合金モデル」は説明的で長いため，評価者によって評価が分かれたと推測される．第三者評価では直接的でないキャッチコピーは好まれづらく，説明的なものや連想を多く必要とするものは評価が分かれる，または，低い傾向にあった．

まとめると，自己評価と第三者評価の間の相関は見られず，第三者評価の間の一致率も低い．このことから，創造的な作業の結果であるキャッチコピーの評価はそれぞれの主観の影響が大きく統一的な評価が難しいことが示された．

4 分析

本節では，我々が先行研究 [12] で行った，発話とキャッチコピーの編集という 2 種の操作に関する時系列クラスタリングの結果を使用し，第三者評価と共同作業の内容の関係を明らかにする．また，特徴的なクラスタの対話事例について分析し，共同作業の内容と作成された成果物の質の関係を明らかにする．

なお，先行研究 [12] では，782 の対話のそれぞれを発話と編集の時系列データとし，k-modes [14] で 6 クラスタにクラスタリングした．その結果，クラスタ 1 は散発的にチャットを挟みながら主にキャッチコピーの編集を行っていることに加え終盤のペースが遅い作業，クラスタ 2 と 4 は 30 分間を通して両者が速いペースでチャットを行いながら，並行してキャッチコピーの編集を行っている作業，クラスタ 3 と 6 は散発的にチャットを挟みながら主にキャッチコピーの編集を行う時間がある作業，クラスタ 5 はクラスタ 4 より遅いペースでチャットとキャッチコピーの編集を行う時間がある作業が主に含まれていた．

4.1 クラスタごとの第三者評価の値

第三者評価の高い作業の流れを明らかにするため，クラスタ間の評価値の差が有意かを調べた．具体的には，クラスタ毎の対話に含まれる各キャッチコピーについて第三者評価の評価値を用い，クラスタ全体の平均値を計算した．さらに，Steel-Dwass の多重比較 [15] を用いて，クラスタ間の評価値に差があるかを検定した．

表 3 にクラスタごとの結果を，先行研究での自己評価の結果と共に示す．キャッチコピーに対する第三者評価の最大値の平均は Q8-2 と Q9-2 の両方でクラスタ 1 において最も高く，2 番目に高いのは Q8-2 はクラスタ 5，Q9-2 はクラスタ 3 であった．Steel-Dwass の多重比較では Q8-2 ではクラスタ 1 と 4，Q9-2 ではクラスタ 1 と 2，1 と 4，1 と 5 に有意差（$p < 0.05$）が見られた．キャッチコピーに対する第三者評価は作業者の自己評価の高低とは異なる傾向が見られる．

クラスタ 2 や 4 のように 30 分間を通して両者が速いペースでチャットを行いながら並行してキャッチコピーの編集を行っている作業では，対話を通じ

表 4　クラスタごとの特徴的な表現（$p < 0.01$）

クラスタ 1	埋まりましたね
クラスタ 2	そうなんです，な感じですか，てしまいました 感じですかね，てみました，いたのです なんですよ，がいいですか，みたいな感じです よろしくお願いいたします
クラスタ 3	てもいいかも，もいいかもしれ
クラスタ 4	てみました，変えてみまし，のはどうでしょ しましょうか，はどうでしょう，どうでしょうか たいのですが，みたんです，しました！ 良いですね！
クラスタ 5	はありますが，みたいですね。
クラスタ 6	気になるところ

て商品に関する連想やひねりを加えやすく，3.2 節で示したように，第三者評価が高くなりづらいような直接的でないキャッチコピーを作成しやすい傾向があるのではないかと考えられる．

4.2　クラスタごとの特徴的な表現

第三者評価が高い作業と関連する表現を分析するため，各クラスタに含まれる特徴的な表現を抽出し，それらと第三者評価の関係を考察する．

各クラスタの対話に偏って多く出現している特徴的な頻出表現を抽出し，分析を行った．抽出対象はチャットの頻出上位 500 個の 4-gram とした．その際，助詞および助動詞のみで構成される 4-gram は対象外とした．偏りの検定にはフィッシャーの正確確率検定を利用した．単語の抽出には形態素解析器として IPADIC 辞書を適応した MeCab[1] を使用した．

表 4 は抽出された特徴的な表現を示している．第三者評価が高いクラスタ 1 は「埋まりましたね」という表現が多く使われており，キャッチコピー編集欄 8 つ全てにアイデアを書き出すことを目標として作業を進めていると考えられる．自己評価は高いものの第三者評価が高くないクラスタ 2，4 は「どうでしょうか」のように，試行錯誤を行っている表現が多く使われている．試行錯誤を行う表現を多く使っていることから，対話を通じてキャッチコピーにひねりや連想を加えていることが確認できた．

4.3　対話事例

表 5 は第三者評価が高いキャッチコピーを作成しているクラスタ 1 の対話例である．この作業は同じペアの 2 回目の対話で，作業開始後 02:27〜06:05 にかけて，キャッチコピー編集欄にアイデアを書き出している．評価が特に高かったキャッチコピーはこの時間に編集欄 H に書き出され，最後まで修正は行われていない．

表 6 は自己評価が高いキャッチコピーを作成しているクラスタ 4 の対話例である．この作業では，対話により商品の認識をすり合わせながらキャッチコピーの作成をしている．評価が高かった編集欄 G の「限りなく打ち込め！球拾いゼロ、時短練習でライバルに差をつける」というキャッチコピーができる過程に注目すると，まず，「打ち込め！」というフレーズは 06:38 時点では B 欄に書き込まれているが，09:16 までの間に削除されている．それを U_1 が 09:27 に良かったと評し，11:41 に D 欄に「限りなく打ち込め！」とフレーズを追加して書き込んでいる．その後も対話しながら編集を続け，24:04 に G 欄の「球拾いゼロ、時短練習で」に H 欄の「正確なスマッシュでライバルに差をつけよう！」の表現が加えられ，U_1 は「借りました」と U_0 が考えた表現を使ったことを述べている．その後，最終的に D 欄のフレーズが G 欄に加えられ，キャッチコピーが完成している．このように，対話しながら複数の案を組み合わせ，キャッチコピーを作成している．

これらの事例分析から，クラスタによって発話と編集の流れに違いがあることが確かめられた．また，キャッチコピーが工夫されていく過程が明確になった．

5　おわりに

本研究では，キャッチコピー共同作成対話コーパスに含まれるキャッチコピーに対して第三者評価を付与し，作業者による自己評価と第三者評価との関係を分析することで共同作業の内容と作成された成果物の質の関係を明らかにした．第三者からの評価では，直接的でないキャッチコピーは好まれづらく，評価者同士の相関や評価者と作成者の間の評価の相関は低いことがわかった．発話と編集の流れの分析では，作業者の自己評価が高いクラスタとキャッチコピーの第三者評価が高いクラスタは異なることがわかった．

これを踏まえ，キャッチコピーを共同で作成する対話システムが持つべき指針について検討すると，その相関から，第三者評価を高めることについては限界があると考えられる．それであれば，親密度や新たな発想が生まれやすい傾向 [12] のある，やり取りを重視するシステムを目指すことが有意義だと考えられる．

1）　https://taku910.github.io/mecab/

謝辞

本研究は科研費「モジュール連動に基づく対話システム基盤技術の構築」(課題番号 19H05692) の支援を受けた．また，本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものである．

参考文献

[1] 市川拓茉, 東中竜一郎. マルチエージェント強化学習に基づく共同作業を自律的に行う対話システムの最適化. 言語処理学会第 29 回年次大会発表論文集, pp. 1383–1387, 2023.

[2] 江連夏美, 稲葉通将. 個人の特性に基づくブレインストーミング対話の分析. 人工知能学会 第 99 回 言語・音声理解と対話処理研究会, pp. 134–138, 2023.

[3] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In **Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems**, pp. 355:1–:34, 2023.

[4] Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga, and Sen Yoshida. Dialogue collection for recording the process of building common ground in a collaborative task. In **Proceedings of the 13th Conference on Language Resources and Evaluation**, pp. 5749–5758, 2022.

[5] Daniel Fried, Justin Chiu, and Dan Klein. Reference-centric models for grounded collaborative dialogue. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2130–2147, 2021.

[6] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in Minecraft. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5405–5415, 2019.

[7] Charles Rich, Candace L. Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human-computer interaction. **AI Magazine**, Vol. 22, No. 4, p. 15, 2001.

[8] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. AI as an active writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In **Joint Proceedings of the IUI 2022 Workshops**, Vol. 10, pp. 56–65, 2022.

[9] Siyuan Hu, Mingyu Ouyang, Difei Gao, and Mike Zheng Shou. The dawn of gui agent: A preliminary case study with claude 3.5 computer use, 2024.

[10] 周旭琳, 市川拓茉, 東中竜一郎. キャッチコピー共同作成タスクにおける対話の収集と分析. 人工知能学会第 36 回全国大会論文集, pp. 2A6GS603–2A6GS603, 2022.

[11] 周旭琳, 市川拓茉, 東中竜一郎. 人間と共同でキャッチコピーを作成する対話システムの試作. HAI シンポジウム, 2023.

[12] Xulin Zhou, Takuma Ichikawa, and Ryuichiro Higashinaka. Collecting and analyzing dialogues in a tagline co-writing task. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 3507–3517, 2024.

[13] 大岩直人. 太宰治のコトバと言葉、その物語。：現代の広告クリエイティブの視点から考察する太宰治のコピーライティング術. コミュニケーション科学, No. 51, pp. 3–43, 2020.

[14] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. **Data mining and knowledge discovery**, Vol. 2, No. 3, pp. 283–304, 1998.

[15] Meyer Dwass. Some k-sample rank-order tests. **Contributions to probability and statistics**, pp. 198–202, 1960.

A　付録

表 5　クラスタ 1 の対話事例．U_1，U_2 はそれぞれ作業者を表し，赤色は U_1，青色は U_2 が関わっていることを示す．網掛けの箇所はキャッチコピーの編集を示す．

00:51	U_2	思いつきやすそうなテーマですね！
01:49	U_2	美味しそう、、、
02:01	U_1	さっきのよりは全然いいですね！
02:27〜	A	未体験の味わい　田中さんがこだわりぬいた自信作
	B	あなたはまだ　この味を知らない
	C	メロンであって　メロンじゃない　衝撃のマルセイユメロン
	F	お手頃価格の
	G	じゅわっととろける。
06:05	H	糖度 18 度以上、異次元の美味さ。
06:07	U_2	赤肉メロンっていう品種があるんですかねえ？？
06:08〜	D	カボチャも
07:03	F	(削除)
07:06	U_1	どうなんですかね？？
07:17〜	D	見た目はカボチャ　中身はメロン　生産者田中
	E	田中さんが作った、橙色の宝石。
10:48	F	こんなメロン、食べたことない。
11:39	U_2	食べてみたいですね
...		
14:13	U_2	お手頃な価格って書いてありますけど、どれくらいなんでしょう
...		
21:20	D	抜群の糖度　抜群の味わい　これぞ抜群のコスパ
...		

表 6　クラスタ 4 の対話事例．U_1，U_2 はそれぞれ作業者を表し，赤色は U_1，青色は U_2 が関わっていることを示す．網掛けの箇所はキャッチコピーの編集を示す．

00:23	U_2	これは、テニスの練習用のネットですね
00:29	U_1	卓球の練習用の品物ってことであっているでしょうか
...		
05:34	U_1	玉を拾わなくていいのは楽で良いですね。下手だとはみでそうですけど汗
06:12	U_2	まあまあ、狙ったところに打つ練習ですからね
06:38	B	打ち込め！ネットが
06:55	U_1	確かに…入らなかった分は諦めないとですね
07:04〜	B	ネットに打ち込んでらくらく球拾い
	C	練習熱心なあなたに
09:16	D	一歩差をつける
09:27	U_1	B に書きかけていた打ち込め！が　あっ良い！って思いましたが、後に続くフレーズが難しいですね
10:30	U_2	狙ったところに打ち込んで、ネットが自動で球を拾ってくれるイメージですよね
11:05	D	(削除)
11:11	U_1	そうですよね。
11:41	D	限りなく打ち込め！
...		
22:25〜	G	球拾いゼロ、時短練習で
23:23	H	正確なスマッシュでライバルに差をつけよう！
23:53	U_2	ライバルに差をつける、、、いいのでたじゃないですか
24:04	G	球拾いゼロ、時短練習でライバルに差をつける
24:11	U_2	借りました
...		