キャッチコピー共同作成対話コーパスにおける発話と編集および参照の分析

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

本研究では，共同作業を行う対話システムの構築に向けて，人間が共同作業を行っているデータの分析を行う．使用するデータは，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タスクに取り組んだ対話データである．分析では，発話及びキャッチコピー編集の発生有無と，キャッチコピー編集欄の位置を指し示す記号による絶対参照表現の発話内での使用有無に着目した．その結果，活発に発話やキャッチコピーの編集を行う対話は作業者の自己評価が他より高くなる傾向があることや，絶対参照が多く使用されるほど評価が高い傾向があることが分かった．

1　はじめに

対話システムが普及し，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] 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コーパスを収集した [9, 10]．このコーパスを分析することで，人間同士の共同作業についての知見が得られ，共同作業を行う対話システムの構築に役立つと考えられる．

本稿では，発話とキャッチコピー編集の発生有無に着目した分析および，発話内でのキャッチコピー編集欄を示す記号による参照表現の使用有無に着目した分析を行う．

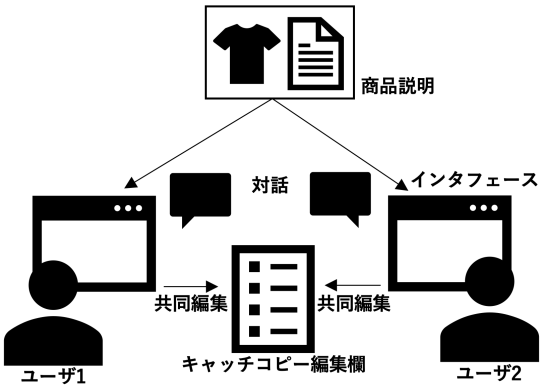

図 1　キャッチコピー共同作成タスクの概略図

2　キャッチコピー共同作成対話コーパス

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）を収集した [9, 10]．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．表 1 は収集された対話から抜粋した対話例である．この対話例では焼酎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

表 1　対話例. U_1, U_2 はそれぞれ作業者を指す. 灰色の行はキャッチコピー編集欄の編集を示す.

U_2	(キャッチコピー編集欄 F を編集) 香りの余韻。時間の余韻。
U_2	F, もう一個余韻で何か重ねたいなと思うのですが、何かいい案ありますかね
U_1	うーん、味で何か重ねられますかね？
U_2	甘味の余韻。…とか
U_1	いいんじゃないでしょうか！
U_2	(キャッチコピー編集欄 F を編集) 香りの余韻。時間の余韻。甘みの余韻。
U_2	ありがとうございます！

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. 表 2 はアンケートの質問項目である.

3　発話と編集の流れの分析

キャッチコピー共同作成タスクの作業時間中は, キャッチコピーの編集と発話という 2 種の操作を行うことができる. 我々は 2 種の操作が作業時間中にどのように発生することで作業が進んでいるかを明らかにするために, ある時間区分中での 2 種の操作両方の発生有無を考慮した作業の流れのクラスタリングを行った. さらに, 成果物であるキャッチコピーに対する評価値や作業者自身の作業に対する評価を発話と編集の流れのクラスタごとに調べ, どのような関連が見られるかを調査した.

まず, 発話と編集の流れの分析のために, 30 分のデータを 1 分ごとに区切り, それぞれの 1 分間にどのようなログが存在するかに着目した. ここで, 1 分間に存在するログの種別ごと（発話 (chat) またはキャッチコピーの編集 (edit)）について, それぞれ何人のログが存在するか（0 人, 1 人, 2 人）の組み合わせをその 1 分間の状態として定義した. つまり, 1 分間の状態の種類は, 3 の 2 乗で 9 種類となる. 例えば, ある 1 分間に二人ともがキャッチコピーを編集し, 一人はさらにチャットも送信していた場合, その 1 分間の状態は chat が一人で edit が二人のため ch1 ed2 と表現される. 1 作業は 30 分のため, 各対話は, このような状態が 30 個並んだ形の状態ベクトルで表現される. この状態ベクトルを k-modes [11] でクラスタリングし, 発話と編集の流れの分類を行った. クラスタ数はエルボー法 [12] により 6 クラスタと決定した.

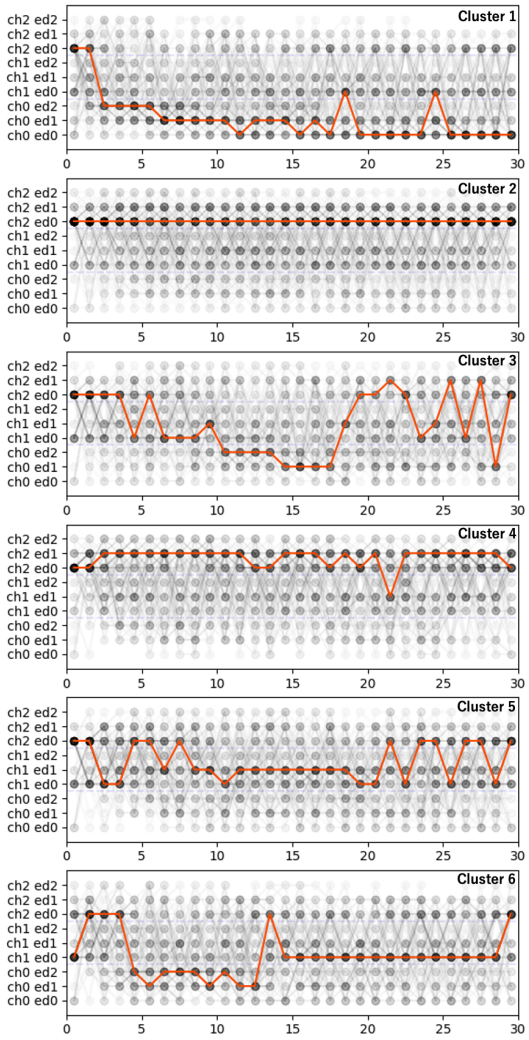

図 2　発話と編集の流れのクラスタリング結果. X 軸は開始からの経過時間（分）を示し, Y 軸の値はそれぞれの 1 分間の状態を表す. 円形のマーカは, 当該経過時間の区切りにおいて Y 軸の状態である作業の割合を, 色の濃さによって示している. 色が濃いほど, 同一クラスタ中でそのマーカに当てはまる作業数の割合が多いことを意味する.

図 2 は, クラスタリングの結果および, それぞれのクラスタに分類された作業の状態ベクトルを示している. 赤線は各クラスタの centroid を表す.

各クラスタがどのような流れで発話と編集を行っていると考えられるかを説明する.

Cluster 1:　開始直後 2 分ほど会話したのち, 15 分程度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. その後は最後まで 1 分間ログがない状態が起こる程度の遅いペースでチャットや編集を行っている.

Cluster 2:　30 分間を通して両者が速いペースで

表 2 アンケートの質問項目

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
Q3. （合意）今回の共同作業では，話し合いによって合意に至ることができましたか
Q4. （異論）今回の共同作業では，あなたと作業相手の意見が分かれることはありましたか
Q5. （アイデア）今回の共同作業では，あなただけでは思いつかなかったようなアイデアが出ましたか
Q6. （満足）今回の共同作業に対して満足できましたか
Q7. （親しみ）作業相手に親しみを感じますか
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

チャットを行っており，並行してキャッチコピーの編集を行っている．

Cluster 3: 開始から 10 分程度の間はチャットを行っている．その次の 10 分程度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．最後の 10 分程度は主にチャットを行いながらしばしばキャッチコピーの編集を行っている．

Cluster 4: クラスタ 2 と同様に 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．キャッチコピーの編集はクラスタ 2 よりさらに頻繁に行われている．

Cluster 5: 開始から 5 分程度の間はチャットを行う．その後 15 分程度はクラスタ 4 よりも遅いペースでチャットとキャッチコピーの編集を並行して行っている．そして，最後の 10 分程度は主にチャットを行いながらキャッチコピーの編集を行っている．

Cluster 6: 開始直後 4 分ほどチャットを行ったのち，10 分程度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．その後は最後まで主にチャットを行いながら，散発的に編集を行っている．

作業者にとって満足度が高い発話と編集の流れはどのようなものなのかを明らかにするため，クラスタごとの作業後アンケートの各項目の平均スコアを計算した．さらに，Steel-Dwass の多重比較 [13] を用いて，クラスタ間の差が有意かを調べた．

表 3 にクラスタごとの結果を示す．作業者の自己評価の値は，クラスタ 2 とクラスタ 4 において高かった．この 2 つのクラスタはどちらも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う流れで作業を進めている．こまめに意見やアイデアの主張，合意の確認などを行っており，意見が異なることに気づく回数も増え，Q1-Q5 の自己評価が大きい値になったと考えられる．また，他のクラスタと比べて作業相手とチャットをしている時間が長くなるため，満足度が高く，作業相手へ親しみを抱きやすくなり Q6 や Q7 が高い値になったと考えられる．キャッチコピーの評価に関しては，Q9 はどのクラスタ間にも有意な差は見られなかったが，Q8 ではクラスタ 2 とクラスタ 5，クラスタ 4 とクラスタ 5 の間に有意な差が見られた．

以上の分析より，発話と編集の流れはいくつかのクラスタに分かれることが分かった．また，クラスタごとの自己評価は作業中にチャットとキャッチコピーの編集を並行して頻繁に行っている場合に高くなることが分かった．

4 絶対参照の発生パターンの分析

キャッチコピー共同作成タスクの一つの特徴は，作成中のキャッチコピーを参照しながら発話を行っている点である．参照には，例えば作成途中のキャッチコピーを示してそれについて議論する参照や，チャット内の単語から着想を得てその表現を含むキャッチコピーを作成する，といったものがある．ここでは，参照の中でも自動抽出が容易な，キャッチコピー欄を表すアルファベットである A から H を発話中に含んでいる参照に着目し，これを絶対参照と定義する．この絶対参照と対話中の出現時間を考慮した分析を行う．

各対話の絶対参照の発生パターンは，3 節での分析と似た手法で作成する．1 対話の 30 分の対話時間を 10 分ごとに区切り，それぞれ 10 分間に絶対参照を含む発話が 1 回でも起こっている場合は 1，起こっていない場合は 0 とする 3 ビット（8 種類）で表した．例えば「011」は最初の 10 分間は絶対参照を含む発話が無く，中盤 10 分間と終盤 10 分間は両方とも絶対参照を含む発話が存在することを示す．782 対話を 8 種のパターンに分類した．

表 4 は絶対参照の発生パターンに該当する作業の統計量および作業者の自己評価の平均スコアである．多くの対話が絶対参照を含み，どちらかという

表 3　各クラスタの平均スコア．クラスタ番号の横の数値は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．上付き文字はこのクラスタ番号のクラスタよりも有意に値が高いことを示す（数字が一つのとき $p < 0.05$，2 つのとき $p < 0.01$）．

	クラスタ					
質問項目	1 (14.3)	2 (35.9)	3 (12.3)	4 (11.6)	5 (14.1)	6 (11.8)
Q1. 自分主張	4.40	**4.65**^{11,5}	4.48	4.53	4.49	4.53
Q2. 相手主張	4.28	**4.64**^{11,3,5,66}	4.44	4.60^{11,6}	4.44	4.33
Q3. 合意	4.25	4.54^{11,5}	4.43	**4.55**	4.36	4.42
Q4. 異論	1.70	**2.08**^{11}	1.84	2.00^{11}	1.97^{11}	1.85
Q5. アイデア	4.12	4.45^{11}	4.28	**4.52**^{11,6}	4.33	4.21
Q6. 満足	4.27	**4.54**^{11,5,66}	4.39	4.47	4.36	4.29
Q7. 親しみ	4.18	4.54^{11,33,55,66}	4.33	**4.56**^{11,3,66}	4.36	4.21
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

表 4　各絶対参照パターンの統計量および平均スコア．パターンの説明は 4 節を参照．太字は各項目における最大値．

	000	001	010	011	100	101	110	111
全作業に占める割合 (%)	12.0	16.1	9.1	25.3	3.5	6.5	4.0	23.5
対話平均発話数	34.63	42.88	38.46	50.66	41.37	48.41	46.90	**55.57**
発話平均文字数	20.01	**21.50**	19.63	20.36	20.38	19.83	20.02	20.07
平均単語数	349.41	461.48	373.82	515.20	420.41	475.16	460.87	**554.72**
平均キャッチコピー編集欄書き込み文字数	227.93	239.44	**253.83**	233.29	227.70	215.65	246.55	237.86
平均キャッチコピー編集欄削除文字数	83.32	98.90	106.15	99.41	87.00	85.37	**108.35**	104.86
Q1. 自分主張	4.35	4.54	4.54	4.59	4.57	4.44	4.60	**4.61**
Q2. 相手主張	4.31	4.41	4.39	4.53	4.48	4.38	**4.69**	4.66
Q3. 合意	4.27	4.35	4.44	4.53	4.37	4.45	**4.60**	4.51
Q4. 異論	1.86	1.98	1.65	1.96	1.91	2.00	1.69	**2.10**
Q5. アイデア	4.11	4.19	4.28	4.43	4.39	4.36	**4.47**	4.46
Q6. 満足	4.20	4.33	4.30	4.50	**4.59**	4.33	4.55	4.53
Q7. 親しみ	4.21	4.32	4.32	4.45	4.56	4.37	**4.61**	4.48
Q8. 興味	4.02	3.91	4.16	4.20	4.22	4.17	**4.37**	4.30
Q9. 想像	4.03	3.96	4.18	4.15	4.07	4.21	**4.34**	4.23

と序盤よりも終盤に存在する傾向があると分かる．統計量からは，作業全体を通して絶対参照を行っている作業は平均発話数が増える傾向があることや，序盤は絶対参照を行わず終盤に絶対参照を行う作業は 1 発話の平均の長さが長いこと，中盤に絶対参照を行う作業はキャッチコピー編集欄への書き込みや削除の文字数が多く，編集が盛んに行われていることが分かる．作業者の自己評価からは，序盤に絶対参照を行う作業は多くの項目で評価が高い傾向にあることが分かる．Q6 と Q7 は序盤に絶対参照があるとスコアが高くなる傾向があるのに対し，Q4 は終盤に近い時点に絶対参照があるほどスコアが高くなる傾向がみられた．

絶対参照の発生パターンの分析により，おおむね絶対参照が全体の広い範囲で行われているほど自己評価が高い傾向があり，適宜キャッチコピーの状態にグラウンディングをしながら対話することが重要であることが示唆された．また，絶対参照が存在した場合でも序盤か終盤かに応じて評価の傾向が異なることが分かった．

5　おわりに

本研究では，キャッチコピー共同作成タスクにおける対話データの分析を行った．発話と編集の流れの分析では，パターンがどの程度あるか，チャットと編集，参照がどのようなパターンの時にどのように満足度が変化するかなどを分析した．その結果，チャットとキャッチコピーの編集を並行して頻繁に行っている作業の場合に，作業者の自己評価が高くなりやすいという知見を得た．絶対参照発生パターンの分析からは，絶対参照が広く使用されるほど自己評価が高い傾向があるという知見を得た．

今後は，本研究で得られた知見を生かし，人間と対話しながら共同でキャッチコピーを作成することのできる対話システムを構築したい．たとえば，GPT-4 [14] などの大規模言語モデルを使用し，チャットや編集を並列して行い，作成中のキャッチコピーへの参照を利用することのできるモデルの開発を行いたいと考えている．

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.

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] 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.

[10] 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.

[11] 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.

[12] Robert L. Thorndike. Who belongs in the family? **Psychometrika**, Vol. 18, pp. 267–276, 1953.

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

[14] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.