Collecting and Analyzing Dialogues in a Tagline Co-Writing Task

Xulin Zhou, Takuma Ichikawa, Ryuichiro Higashinaka
Graduate School of Informatics, Nagoya University, Japan
{zhou.xulin.j3@s.mail, ichikawa.takuma.w0@s.mail, higashinaka@i}.nagoya-u.ac.jp

Abstract

The potential usage scenarios of dialogue systems will be greatly expanded if they are able to collaborate more creatively with humans. Many studies have examined ways of building such systems, but most of them focus on problem-solving dialogues, and relatively little research has been done on systems that can engage in creative collaboration with users. In this study, we designed a tagline co-writing task in which two people collaborate to create taglines via text chat, created an interface for data collection, and collected dialogue logs, editing logs, and questionnaire results. In total, we collected 782 Japanese dialogues. We describe the characteristic interactions comprising the tagline co-writing task and report the results of our analysis, in which we examined the kind of utterances that appear in the dialogues as well as the most frequent expressions found in highly rated dialogues in subjective evaluations. We also analyzed the relationship between subjective evaluations and workflow utilized in the dialogues and the interplay between taglines and utterances.

Keywords: Data collection, Creative collaboration, Co-writing, Dialogue system, Tagline

1. Introduction

As dialogue systems gain more popularity and become ever more prevalent in society (Roller et al., 2021; Chi et al., 2022), many studies on building such systems that can collaborate with humans have emerged (Fried et al., 2021; Mitsuda et al., 2022). However, most have focused on problem-solving dialogues or those in which the system responds to commands from users, and relatively little research has been conducted on systems that can creatively collaborate with users (Wang et al., 2023). Considering the current advancements in artwork generated by artificial intelligence (Cheng et al., 2020; Dhariwal and Nichol, 2021; Kim et al., 2022; Koley et al., 2023), we feel that the potential usage scenarios of dialogue systems can be greatly expanded if they are able to participate in more creative collaboration with users.

The scarcity of dialogue systems that can engage in such collaboration with users can be attributed to the limited availability of datasets that involve creative collaborative tasks carried out through dialogue. Therefore, in this study, we devised a tagline co-writing task in which two interlocutors collaborate to create taglines via text chat, and collected the dialogue data of people performing the task along with their subjective evaluations through a questionnaire on the created taglines. Figure 1 shows a conceptual diagram of the data collection process. We then analyzed the collected data to clarify the utterances and expressions used in the dialogues, the relationship between the subjective evaluations and interactions in the dialogues, and the interplay between taglines and utterances.

The contributions of this study are three-fold:

- We collected 782 dialogues in Japanese on the tagline co-writing task to determine how two individuals converse while referencing and critiquing in creative collaboration.
- We categorized the interaction patterns in creative collaboration and successfully identified interaction patterns that yield high satisfaction for the individuals.
- Through the analysis of referential expressions, we gained insights into the interplay between utterances and the creation being made.

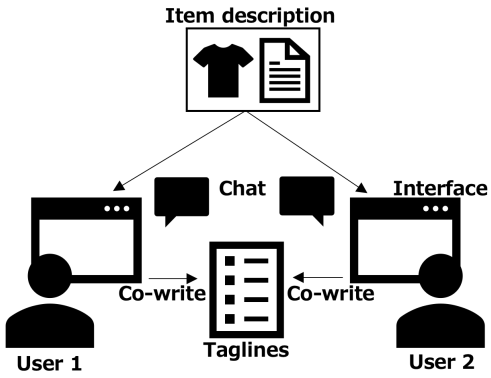

Figure 1: Conceptual diagram of tagline co-writing process

2. Related Work

There are many studies related to dialogue systems that can collaborate with users. However, most are limited to systems that perform a particular task or respond to commands given by users (Rich et al., 2001; Narayan-Chen et al., 2019;

LREC-COLING 2024, pages 3507–3517
20-25 May, 2024.

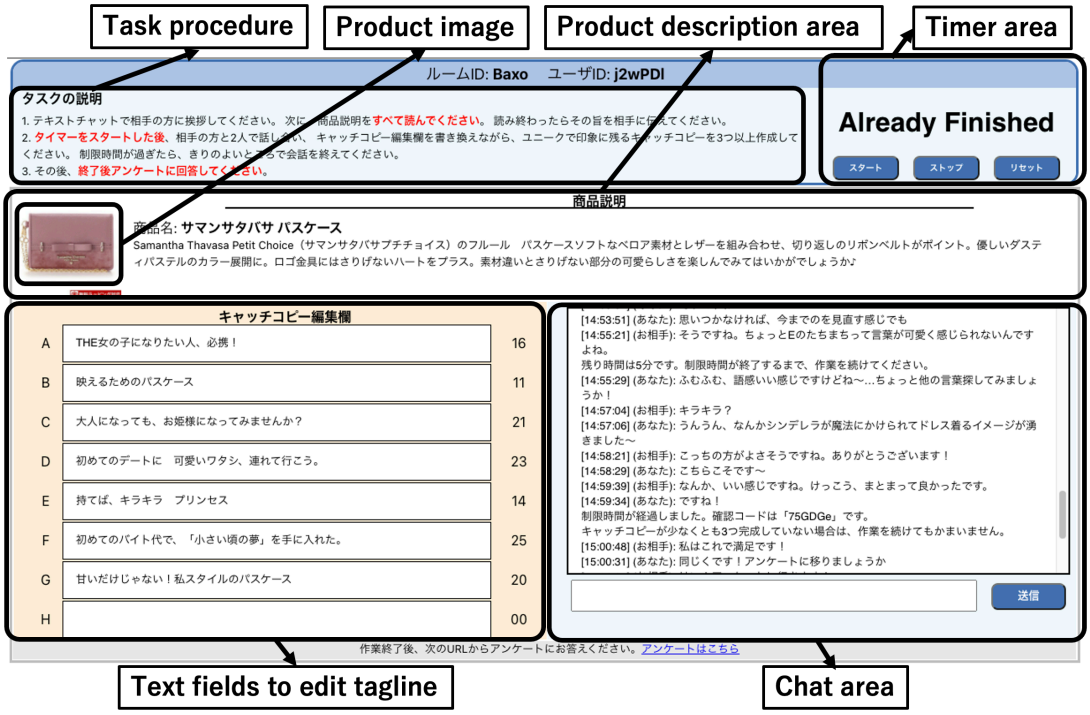

Figure 2: Interface for tagline co-writing task.

Jayannavar et al., 2020; Padmakumar et al., 2022; Roush et al., 2022; Wu, 2022).

In dialogues with collaborative tasks between two parties, establishing common ground is important (Benotti and Blackburn, 2021). Some works has focused on building common ground with users in a human-machine collaboration, but still, they are limited to particular tasks such as finding a specific object (Fried et al., 2021; Udagawa and Aizawa, 2021) or laying out objects in a specific environment (Clark et al., 2021; Lachmy et al., 2022; Mitsuda et al., 2022). One exception may be the work by Ichikawa and Higashinaka (2022), which aimed at building an agent that can engage in creative collaborative work on Minecraft. Our study differs in that we focus on tagline co-writing, which is more related to natural language. Additionally, our work focuses on the process of creating multiple taglines and enabling the individuals to compare various tagline candidates, and making the resulting dialogue contain the process through which a new artifact is created based on existing ones.

Prior research on producing creative works in collaboration with systems includes a number of studies on narrative creation with the system (Fang et al., 2023). Wordcraft is equipped with an interface that utilizes large language models (LLMs), allowing for the generation of alternative or subsequent narrative texts (Mirowski et al., 2023). Yang et al. (2022) developed a tool where humans and artificial intelligence (AI) alternately create narratives section by section and where the user has the ability to edit or regenerate these sections. However, these studies focused on the deliberation process of individuals, and dialogues between multiple parties were not considered.

3. Collection of Tagline Co-Writing Dialogues

To explore and better understand how dialogues work in the tagline co-writing task, we collected the dialogues of people working on such tasks. Specifically, we created an interface on which users perform the tagline co-writing tasks and then collected data on the human dialogues and editing operations. This section describes the tagline co-writing task and the data collection experiments we conducted, as well as the results of self-evaluation questionnaires administered to participants.

3.1. Tagline Co-Writing Task

In the tagline co-writing task, participants worked in pairs to discuss and edit the text fields collaboratively and to create taglines while referring to the provided product description. Each participant was in a different location and interacted with the partner via text chat. All information displayed on the interface was shared among the participants. Participants were free to take any action during the 30-minute period, including editing the tagline that the other participant had written earlier. The

	タイマーをスタートしました。作業を開始してください。制限時間は 30 分です。	*Timer started. Please start working. The time limit is 30 minutes.*
U_2	かわいいパスケースですね。	*Cute passcase.*
U_1	はい、可愛いですね〜THE 女の子って感じです	*Yes, it's cute—it's like THE girl.*
U_2	その　THE 女の子っていいですね。	*I like that THE girl.*
U_2	(Fill in tagline edit field A) THE 女の子	*THE girl*
U_1	お、ありがとうございます！	*Oh, thank you!*
U_2	なんか、持ってたら見せたくなっちゃいそうです。	*I feel like, if I had it, I'd want to show it to somebody.*
U_1	わかります。	*I know.*
U_2	(Fill in tagline edit field B) 映えるためのパスケース	*Passcases to look good*
U_1	なんだろう、お嬢様とかお姫様っぽい気分になれそうというか	*I don't know how to express this exactly, but it might make the person who has it feel like a young lady or a princess*

Table 1: Example dialogue. U_1, U_2 respectively represent the two users. The shaded rows indicate the edits made to the tagline. The utterances were originally in Japanese and have been translated into English by the authors.

two participants were asked to create at least three taglines in total.

3.2. Interface

To collect the dialogue data, we created a Web-based interface that enables two participants in a remote place to edit taglines collaboratively.

Figure 2 shows a screenshot of the interface. A timer was placed at the top right of the screen to display how much time was remaining until the end of the task. The product name, description, and image were displayed in the upper middle for the product that was the target of the tagline. Eight text fields (marked A–H) were provided at the bottom left of the screen, where participants could share and edit taglines. When a participant wrote something in a text field, the change was reflected on the screen of the other participant immediately. Each text field could contain up to 30 characters. The bottom right of the screen was a chat area where the participants could communicate with each other via text chat.

The interface was created using the Web application framework React[1], which enables simultaneous editing of taglines and text chat. Messaging and tagline-editing operations were managed using a back-end server that recorded the entire text chat history and tagline-editing operations with time stamps. A small-scale preliminary experiment was conducted to ascertain the usability of the interface for data collection.

3.3. Product Description Data

We prepared 100 different products for our data collection. The products were selected to cover the top genre list on Rakuten[2], one of the largest online shopping sites in Japan. Examples of items in the top genre list include Fashion, Gourmet & Beverages, and Daily Goods & Healthcare. We selected the products uniformly from these categories to ensure a balanced representation.

The product descriptions were created by manually extracting parts of the product information obtained using the Rakuten Product Search API[3]. The product images were also obtained using the same API.

No. of participants	105
No. of participant pairs	398
No. of dialogues	782
Chat	
No. of utterances in a dialogue	46.91
No. of characters in an utterance	20.31
No. of words in a dialogue	475.04
Mean no. of different words in a dialogue	170.14
Text fields to edit a tagline	
No. of fields filled in a dialogue	7.60
No. of characters entered in a dialogue	235.76
No. of characters deleted in a dialogue	98.30
No. of words per field	8.34
No. of completed taglines	6.86
No. of words in a completed tagline	8.51
No. of unique words in completed taglines	8.10

Table 2: Statistics of the collected data

3.4. Data Collection Procedure

For the data collection, first, each participant pair accessed the collaboration interface (Section 3.2) and checked the product description. After both

[1] https://ja.react.dev/

[2] https://www.rakuten.co.jp/category/

[3] https://webservice.rakuten.co.jp/documentation/ichiba-item-search

Questionnaire item	Mean	SD
Q1. Were you able to assert your thoughts and opinions?	4.54	0.71
Q2. Did the person you were working with assert their opinions and ideas?	4.50	0.80
Q3. Were you able to come to an agreement through discussion?	4.45	0.85
Q4. Were there times when you and your partner disagreed?	1.95	1.15
Q5. Did you two come up with ideas that you would not have thought of on your own?	4.34	0.96
Q6. Were you satisfied with this collaborative work?	4.42	0.89
Q7. Do you feel you became familiar with the person you were working with?	4.40	0.90
Q8. Do you think the taglines you created will attract the interest of the people who see them?	4.16	0.86
Q9. Do you think the taglines you created will capture the imagination of the viewer?	4.13	0.90
Q10. How easy was it to use the collaboration page?	4.44	0.83

Table 3: Results of questionnaire (5-point Likert scale). SD denotes standard deviation.

participants were finished with this check, one of them started the timer, and collaborative work began. The participants were instructed to engage in dialogue with their partners, revising the text fields to edit taglines and creating three or more unique and memorable taglines in one 30-minute session. After the 30 minutes were up, the participants were encouraged to wrap up the conversation, conclude their work, and answer the questionnaire.

In the questionnaire, participants were asked to evaluate their work subjectively on a five-point Likert scale (1 indicating "disagree" and 5 "agree"). The questionnaire items (listed in Table 3) were primarily inspired by the research conducted on the data collection of collaborative dialogue using Minecraft (Ichikawa and Higashinaka, 2022). In addition, the items related to tagline evaluation were created by referencing books and literature on tagline creation published in Japan. The participants also indicated which taglines were completed because we wanted to distinguish the completed taglines from those they had not finished editing. After both participants completed the questionnaire, they performed another task session, but this one for a different product. The same pair could only perform two task sessions, and each participant could perform the task session a maximum of 20 times over the entire data collection process.

Participants were recruited using the Lancers crowdsourcing service[4] in Japan, and pairs were created from among them. Participants were required to be native Japanese speakers with touch-typing ability, familiarity with text chatting, and a consistent text input speed (at least 200 characters per minute as a guideline). We aimed for a balanced gender ratio, resulting in an approximate ratio of 2:3 for males to females. Age was not a criterion. For each of the 100 products, approximately eight dialogues were collected. From 105 recruited participants, we obtained 801 dialogues in all. After removing dialogues containing technical errors, we ended up with 782 dialogues for our analyses. The data collection experiment was approved by the ethics committee of our institution with regards to the data collection procedure and the handling of personal information.

Table 1 shows an example of a dialogue collected in the interaction shown in Fig. 2. Participants exchanged their thoughts and collaboratively filled in the text fields for the taglines.

4. Collected Data

Table 2 shows the statistics for the 782 dialogues collected in this study. The average number of fields entered in per dialogue was 7.60. As there were eight text fields in total (A–H) for editing taglines, this means the tagline edit field was used up to the maximum limit in many dialogues. The mean number of characters entered in the tagline edit field in one dialogue was 235.76, and the mean number of deleted characters was 98.30. This indicates that about 42% of the input characters were eventually deleted, suggesting that a lot of trial-and-error activity took place in the tagline edit fields.

Table 3 shows the results of the questionnaire, where the answers were collected on a five-point scale. To summarize the results briefly, from Q1 and Q2, we were able to collect data where the workers were able to convey their thoughts to their partners. From Q3 and Q4, we can see that many tasks reached a consensus in the end, and that few instances occurred in which the workers had differing opinions. From Q6 and Q7, we can observe that the workers were satisfied with the tasks and developed a sense of familiarity with their partners. Q8 and Q9 indicate that the workers could create taglines that they personally found satisfactory.

Note that Q4 is about disagreement and is not an item in which higher/lower indicates better/worse. The high values for all items other than Q4 indicate that we successfully collected a fair amount of collaborative work with good self-evaluations by the participants. In particular, the high value of Q5

[4]https://www.lancers.jp

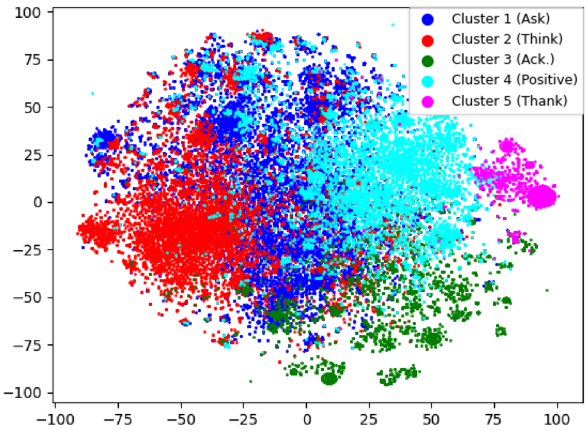

Figure 3: Utterance clustering results

(New ideas) demonstrates that two people working together to create a tagline can contribute to a wider range of ideas. The high value for Q10 (Interface) also indicates that our Web-based interface was easy for the participants to use, even though they did not receive training prior to the experiment.

5. Analysis

We analyzed the collected data to clarify which interactive factors are important with regards to the quality of the tagline co-writing task.

5.1. Analysis of utterances and expressions

We first conducted clustering of the utterances to investigate which utterances and expressions were used in the collaborative dialogue for tagline creation. Further, to understand which specific expressions were particularly used in highly rated dialogues (in subjective evaluations), we extracted expressions that appeared predominantly in highly rated dialogues and analyzed them.

5.1.1. Clustering of Utterances

First, we transformed each utterance into a vector using Sentence-BERT (Reimers and Gurevych, 2019). We chose to use sentence-BERT because we were not certain about the dialogue acts (intentions) that we would observe in the interactions, and Sentence-BERT was deemed capable of accommodating a wide variety of utterances. We then utilized K-means (Lloyd, 1982) to classify these vectors into clusters. We used K-means, which utilizes hard clustering, in the hope that we could distinctly categorize the underlying dialogue intentions for our analysis. On the basis of the silhouette plot, we determined the number of clusters to be five. To inspect the distribution of clusters visually, we conducted dimensionality reduction using t-SNE (Van der Maaten and Hinton, 2008), converting the vectors into two dimensions.

We utilized the scikit-learn package for the K-means clustering. The initial parameters were set by greedy K-means++. The K-means algorithm used a convergence tolerance of 0.0001 for the termination criterion, and the Lloyd's algorithm was implemented. Euclidean distance was used. The number of clusters for the silhouette method was evaluated within the range of 2 to 10. For t-SNE visualization, perplexity was set to 50, the maximum number of iterations for optimization was 1000, PCA was used for embedding initialization, and the learning rate was set based on previous works (Belkina et al., 2019; Kobak and Berens, 2019).

Figure 3 displays the results plotted with distinct colors for each cluster. The proportion of utterances classified into each cluster was 36.3% for Cluster 1, 24.3% for Cluster 2, 13.3% for Cluster 3, 21.6% for Cluster 4, and 4.5% for Cluster 5.

Cluster 1 contained many utterances with a querying nuance towards the partner, such as "Gの後半なんですが、見かけによらず丈夫です、にするのはどうでしょうか？ (About the second half of G, how about making it 'They are tough despite their look'?)," "なにか思いつくことはありますか？ (Do you have any ideas?)," and "女性に限定しなくてもいいのか..... (Does it have to be limited to women?....)." Cluster 2 had many expressions conveying personal thoughts and declarations of intending to think, such as "わたし的にはいじるとこないかなって思いました。やっぱりFがお気に入りです (In my opinion, there is no room for improvement. F is still my favorite.)" and "私ももう少し考えてみますね (I'll give it some more thought too)." Cluster 3 had many acknowledging utterances containing affirmative interjections and backchannels, such as "そうですね, C, D は完成されたかなって思ってます (Yes, I think C and D are completed.)" and "なるほど、確かに言われてみればそうですね (Well, come to think of it, that's true.)." Cluster 4 had many utterances containing expressions that highlight positive evaluations, such as "一番いいかもしれません (That might be best.)" and "分かります。軽いし持ち運びに便利なのも良いですね。(I see. It's great that it's light and convenient to carry around.)." In Cluster 5, many utterances contained expressions of gratitude, such as "ありがとうございます。(Thank you.)" From Fig. 3, we can see that Cluster 1 is proximate in position to Clusters 2 and 4, and that the boundaries between these clusters are not very clear. This can be attributed to the fact that many utterances seem to encompass elements characteristic of multiple clusters. For example, utterances that first provide an affirmative response to another person's

statement and then proceed to one's own thoughts would fit this description. The isolation observed in Cluster 5 can be attributed to the tendency for expressions of gratitude to be made separately from other utterances.

The most prevalent cluster overall was Cluster 1, suggesting that in the tagline co-writing task, knowledge and idea sharing mainly occur through dialogues, facilitated by inquiries for information collection and knowledge sharing. This process occurs together with other communicative acts, such as declaration, affirmation, and thanking.

5.1.2. Expressions in Highly Rated Dialogues

In addition to the clustering of utterances, to analyze the phenomena in dialogue in more detail, we focused on mining frequently used expressions in our data, especially those used in high-rated dialogues. As the target of this analysis, we extracted the top 500 most frequently occurring 4-grams in the collected chat data. We focused on 4-grams due to their length, which facilitates meaningful interpretation without excessive sparsity.

For all questionnaire items other than Q4 (Disagreement), we examined the expressions that appeared more frequently in highly rated dialogues, where the total score of the two participants' answers was 10 (i.e., both participants chose "agree"). For Q4, we examined the expressions that appeared more frequently in the dialogues where the total score of the two participants' answers was more than 6 (i.e., the median value of the scale). Fisher's exact probability test, which can be used for samples of infrequent occurrences, was utilized to identify expressions that occurred significantly more often in the highly rated dialogues. For word segmentation, we used MeCab[5], a Japanese morphological analyzer.

Excluding Q4, the results showed that expressions such as "ありがとうございます！ (Thank you!)", "いいですね！ (Great!)", "と思います！ (I think)", etc. had a frequent occurrence rate in the highly rated dialogues. In Q1 (Own opinion) and Q2 (Partner opinion), the expression "な感じですか (Is it like that?)" was frequently used. This expression is used when one person expresses understanding and asks for confirmation from the other person. Thanking, positive evaluations, expressing one's own understanding, and asking for agreement all seem to be important in the tagline co-writing task.

The expression "てもいいかも (Might be good to do)" was used particularly frequently in Q3 (Agreement). It was often utilized to bring up an idea for tagline editing, such as "A と B 合わせてもいいかも (A and B might be good to combine)". The use of such expressions is considered to lead to agreement. In Q4 (Disagreement), the question "ありますか？ (Do you have any?)" was used when asking about experience with using the product, as in "こういった商品使ったことありますか？ (Do you have any experience with using these products?)", or when asking about the evaluation or recognition of the created taglines, as in "今のリストの中でお好みとかありますか？ (Do you have any preferences among the current taglines?)". It seems that the use of expressions such as the latter, which asks about the evaluation and perception of the taglines, provides a particularly good opportunity when disagreement occurs to reconcile differences in opinion that had not been noticed.

5.2. Clustering of Workflow

Having analyzed individual utterances and expressions, we next conducted an analysis of the interactions. Specifically, we performed clustering on the workflow, considering both the edits of the tagline and the utterances exchanged, to elucidate how the tagline co-writing task progressed over the 30-minute period. We also examined the relationships between the clusters and their subjective evaluations, investigating what interactions were more satisfactory.

First, to analyze the workflow, the 30-minute data was divided into one-minute segments, focusing on what type of log existed for each minute. Here, for each minute, the type of log (chat or tagline editing) and the number of people (0, 1, or 2) involved were considered to define the state of that minute. Therefore, $3^2 = 9$ possible states exist for each minute. For example, during one particular minute where both participants edited the tagline and one of them also sent a chat, that minute's state was labeled as "chat1 person, edit2 people" (or "ch1 ed2" for short). Since each dialogue spans 30 minutes, each dialogue can be represented as a state vector composed of 30 such states. To cluster these state sequences, we utilized k-modes clustering (Huang, 1998). The k-modes algorithm was designed to cluster data considering categorical attributes. We utilized a package available on PyPI for the k-modes method, with initialization using Huang's method (Huang et al., 1997). The optimal number of clusters was determined to be six based on the elbow method. The number of clusters for the elbow method was considered within the range of 2 to 10.

Figure 4 presents the clusters. The red lines indicate the centroids of each cluster. The proportion of the state sequences assigned to each cluster was 14.3% for Cluster 1, 35.9% for Cluster 2, 12.3% for Cluster 3, 11.6% for Cluster 4,

[5]https://taku910.github.io/mecab/

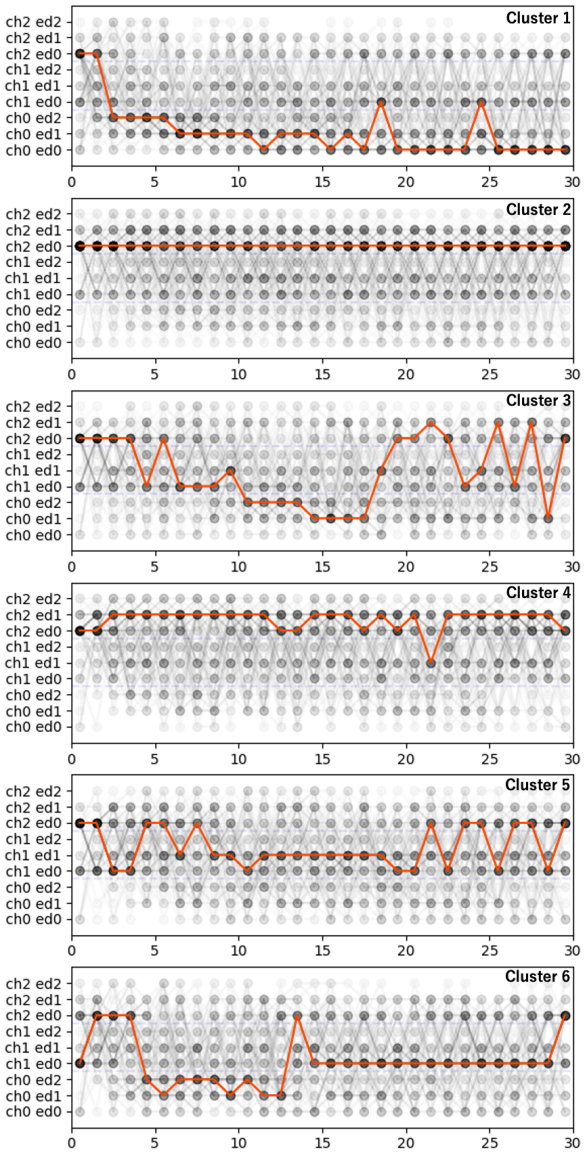

Figure 4: Workflow clustering results. X-axis represents the elapsed time from the start (in minutes), and values on the Y-axis represent the state for each minute. For example, “ch1 ed2" indicates that one of the individuals sent a chat message and that both of them made an edit. The darker the marker's color, the higher the proportion of that state.

14.1% for Cluster 5, and 11.8% for Cluster 6. Note that, in the figure, the opacity of each plotted sequence reflects the proportion, where a darker marker indicates a higher proportion with that particular state. This enables analyzing the detailed workflows within each cluster.

The characteristics of each cluster were as follows.

Cluster 1: Immediately after commencement, there was about two minutes of conversation, followed by roughly 15 minutes characterized by primarily tagline-editing activities with sporadic chatting. Subsequent to this, until the end of the dialogue, a slower pace of chat and edits occurred.

Cluster 2: Throughout the entire 30-minute duration, both parties engaged in rapid-paced chatting while concurrently editing the taglines.

Cluster 3: The workflow began with approximately ten minutes of chatting. This was followed by another ten minutes where the primary activity was tagline editing with sporadic chats. The final ten minutes had a predominant focus on chatting, with sporadic tagline edits.

Cluster 4: Mirroring the pattern of Cluster 2, continuous rapid-paced chatting occurred throughout the entire 30 minutes. However, tagline editing happened more frequently than in Cluster 2.

Cluster 5: The workflow started with around five minutes of chat. Subsequently, for roughly 15 minutes, a combination of chat and tagline edits occurred, albeit at a pace slower than in Cluster 4. The final ten minutes mainly involved chatting while editing taglines.

Cluster 6: The process started with about four minutes of conversation. This was followed by approximately ten minutes of mostly tagline editing, occasionally interrupted by chats. The remainder of the time was dominated by chatting with sporadic editing activities.

To elucidate the workflow that yielded the highest satisfaction for workers, the average subjective evaluation scores were calculated by using the dialogues for each cluster. We also performed a Steel-Dwass multiple comparison test (Dwass, 1960) to determine the significance of the differences between the scores of the clusters. The results are presented in Table 4.

Clusters 2 and 4 showed high self-evaluation scores from the workers. Both of these clusters had workflows in which chatting and concurrent tagline editing occurred. This suggests frequent sharing of opinions and ideas, affirmation of consensus, and an increased frequency of recognizing differing views, leading to higher scores in the self-evaluations (Q1–Q5). Moreover, relative to other clusters, these two had prolonged durations of chat exchanges with their work partners. This might have fostered a higher degree of satisfaction and affinity towards their partners, resulting in higher scores for Q6 and Q7.

From the aforementioned cluster analysis, we can see distinct workflows regarding how to carry out the two activities—editing taglines and exchanging chats—within the 30-minute dialogue.

Questionnaire item	Cluster 1	2	3	4	5	6
Q1. Own opinion	4.40	**4.65**^{11,5}	4.48	4.53	4.49	4.53
Q2. Partner opinion	4.28	**4.64**^{11,3,5,66}	4.44	4.60^{11,6}	4.44	4.33
Q3. Agreement	4.25	4.54^{11,5}	4.43	**4.55**	4.36	4.42
Q4. Disagreement	1.70	**2.08**^{11}	1.84	2.00^{11}	1.97^{11}	1.85
Q5. New ideas	4.12	4.45^{11}	4.28	**4.52**^{11,6}	4.33	4.21
Q6. Satisfaction	4.27	**4.54**^{11,5,66}	4.39	4.47	4.36	4.29
Q7. Familiarity	4.18	4.54^{11,33,55,66}	4.33	**4.56**^{11,3,66}	4.36	4.21
Q8. Interest	4.13	4.23^{5}	4.09	**4.24**^{5}	4.05	4.05
Q9. Imagination	4.11	4.19	4.03	**4.22**	4.08	4.09

Table 4: Average questionnaire score of each cluster. **Bold** indicates the highest values for each item. Underscores indicate the top two scores for each item. Numbers in superscript indicate that the superscripted score is significantly better at $p < 0.05$ (single number) or $p < 0.01$ (double number) than the clusters' scores indicated by these cluster numbers.

Ratio of taglines with references (%)	88.48
Absolute references per tagline	0.610
Expression references per tagline	5.203
Absolute & expression references per tagline	0.120
No. of absolute references (5-min before edit)	0.103
No. of expression references (5-min before edit)	1.168
No. of absolute references (5-min after edit)	0.043
No. of expression references (5-min after edit)	0.178

Table 5: Statistics on absolute and expression references for taglines.

Additionally, subjective evaluation scores tended to be higher in clusters where chatting and tagline editing were conducted concurrently throughout the collaborative work. This is likely because a consistent exchange of thoughts between participants enabled them to share and understand each other's perspectives frequently. This is in line with our findings in the frequently used expressions in highly rated dialogues.

5.3. Interplay between utterances and taglines

One notable feature of our data is the interplay between utterances and taglines. This occurs as references in utterances to the taglines; such references involve discussing a tagline under construction or drawing inspiration from a word in the chat to create a tagline incorporating that expression. By examining how the workers utilize references, we can understand the relationship between dialogues and the artifacts being created.

To investigate the referential relationship between chats and taglines, we carried out an extraction of utterances that contain references. We considered two types of references: absolute references, which include alphabets A to H representing the tagline fields, and expression references, which refer to the taglines using words included in the taglines. Note that this method has its limitations; it might not have captured the instances where expressions were paraphrased, and it might have extracted utterances not intended to reference taglines because we only used word-matching.

In the extraction procedure, we first performed morphological analysis on all the taglines in the data using MeCab. Then, after removing stop words, the resulting words were considered as potential words for referencing the taglines. Subsequently, morphological analysis was performed on all the utterances, and those that had either an absolute reference (corresponding to alphabet letters A–H) or an expression reference with each tagline were automatically extracted.

Table 5 provides statistics on the extracted results. Approximately 88% of the overall tagline editing had references, either with words (expression references) or with absolute references. About 20% of the absolute references also included other absolute references within the same utterance, that is, taglines were being compared.

Table 6 shows an example of the extracted utterances and the tagline edits. The product in focus for the tagline co-writing was "sweet potato shochu" (a type of distilled Japanese spirit). The initial two utterances refer to the product description, and the participants are talking about its characteristic aroma. Then, at 12:44, User 2 writes a tagline related to the aroma. This tagline contains information about the product's distinctive aroma, which was also mentioned in the utterances right after the start. At 20:27, User 2 makes an utterance that includes the absolute reference, asking for ideas. At 22:09, User 2 suggests an expression (expression reference) to be included in the tagline, and by 23:00, that expression has been added.

We found many absolute and expression references in the collaborative dialogues. Although more analyses will be needed, our current analysis suggests the complexity of interactions and the

00:46	U_2	そうですね、紅茶の香りがするんでしょうか？。	*Yes, does it smell like tea?*
01:42	U_1	紅茶まんまじゃなくても、近い香りがするんじゃないでしょうか	*If not exactly like tea, it probably has a close aroma.*
12:44	U_2	(Fill in tagline edit field F) 香りの余韻。時間の余韻。	*Afterglow of aroma. Afterglow of time.*
20:27	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	*I'd like to layer something else in F using "afterglow." Any suggestions?*
22:09	U_2	甘味の余韻。…とか	*For example, "Afterglow of sweetness."*
23:00	U_2	(Change the tagline edit field F) 香りの余韻。時間の余韻。甘みの余韻。	*Afterglow of aroma. Afterglow of time. Afterglow of sweetness*

Table 6: Example of extracted utterances and tagline on text field F. U_1, U_2 respectively represent the two users. Shaded rows indicate the edits made to the tagline. The utterances were originally in Japanese and have been translated into English by the authors.

necessity for grounding utterances in taglines.

6. Summary and Future Work

In this study, we designed a tagline co-writing task and collected dialogue data for analysis to lay the groundwork for developing a dialogue system that can work with users in creative collaboration. Specifically, we created an interface for the task and collected 782 dialogues from 105 participants. We performed utterance clustering and mined frequently used expressions in highly rated dialogues. In addition, we analyzed the workflow and also the interplay between the utterances and taglines. We found that our new dataset has a sufficient level of complexity to pose the challenges required for future dialogue systems to engage in collaborative dialogue.

In future work, we plan to develop a system using LLMs (e.g., GPT-4) that can generate utterances or tagline edits when given the context, including the dialogue history, created taglines, and product descriptions. We will also further analyze the behavior of users in our data to gain more insight into human creative collaboration. We are also interested in the effect of cultural and linguistic background (including biases) on the interactions, and we plan to investigate how specific types of utterances or dialogue patterns contribute to taglines, although annotations (e.g., linking utterances with taglines) will be necessary for this investigation. Finally, after further analyses and preparations, we plan to release the data and the code.

7. Acknowledgments

This work was supported by JSPS KAKENHI Grant number 19H05692 and JST Moonshot R&D Grant number JPMJMS2011. The computation was carried out using the Flow supercomputer at the Information Technology Center, Nagoya University.

8. Ethical Considerations

All evaluations were approved by the research ethics committee of our institution. We employed workers using a crowdsourcing service and made sure they were paid above the minimum wage.

9. Bibliographical References

Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. 2019. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of 16th Conference of the Europian Chapter of the Association for Computational Linguistics*, pages 515–531.

Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. 2020. Sequential attention GAN for interactive image editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4383–4391.

Ethan A. Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Trenton Chang, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, Yutong He, Hari Sowrirajan, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, Jillian Tang, Avanika Narayan, Giovanni Campagna, and Christopher Manning. 2022. Neural generation meets real people: Building a social, informative open-domain dialogue agent. In

Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 376–395.

Christopher Clark, Jordi Salvador, Dustin Schwenk, Derrick Bonafilia, Mark Yatskar, Eric Kolve, Alvaro Herrasti, Jonghyun Choi, Sachin Mehta, Sam Skjonsberg, Carissa Schoenick, Aaron Sarnat, Hannaneh Hajishirzi, Aniruddha Kembhavi, Oren Etzioni, and Ali Farhadi. 2021. Iconary: A pictionary-based game for testing multimodal communication with drawings and text. In *Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1864–1886.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th Congerence on Neural Information Processing Systems*, volume 34, pages 8780–8794.

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

Xiaoxuan Fang, Davy Tsz Kit Ng, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2023. A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies*, 28:3953–3975.

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

Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Zhexue Huang et al. 1997. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining*, pages 21–34.

Takuma Ichikawa and Ryuichiro Higashinaka. 2022. Analysis of dialogue in human-human collaboration in Minecraft. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 4051–4059.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a minecraft dialogue. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2589–2602.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435.

Dmitry Kobak and Philipp Berens. 2019. The art of using t-SNE for single-cell transcriptomics. *Nature communications*, 10(1):5416.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2023. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6850–6861.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw me a flower: Processing and grounding abstraction in natural language. *Transactions of the Association for Computational Linguistics*, 10:1341–1356.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. 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*, pages 355:1–:34.

Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga, and Sen Yoshida. 2022. 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*, pages 5749–5758.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the Association for Computational Linguistics*, pages 5405–5415.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings

using siamese BERT-networks. *CoRR*, abs/1908.10084.

Charles Rich, Candace L. Sidner, and Neal Lesh. 2001. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22(4):15.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Allen Roush, Sanjay Basu, Akshay Moorthy, and Dmitry Dubovoy. 2022. Most language models can be poets too: An AI writing assistant and constrained text generation studio. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 9–15.

Takuma Udagawa and Akiko Aizawa. 2021. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432.

Xianchao Wu. 2022. Creative painting with latent diffusion models. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 59–80.

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