A Coding Scheme for Conversational Style Classification in Online Microtasking

Sihang Qiu
Delft University of Technology
Delft, The Netherlands
s.qiu-1@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju-1@tudelft.nl

Alessandro Bozzon
Delft University of Technology
Delft, The Netherlands
a.bozzon@tudelft.nl

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.

Abstract

To improve satisfaction and engagement of workers in online crowdsourcing marketplaces, recent research in HCI has proposed the use of conversational interfaces to support the execution of a variety of tasks. Prior works in linguistics have shown that conversational styles have an important impact on human communication. Little is known about whether and how conversational styles of agents can be leveraged to improve outcomes in conversational microtasking. To this end, we propose a coding scheme to classify the conversational styles crowd workers and empirically observe how conversational styles of workers relate to work outcomes. Results show that crowd workers with a *highly involved* conversational style produced significantly higher quality results, exhibited a higher user engagement and perceived less cognitive task load in comparison to those in a *highly considerate* style. Our findings can have important implications on task design in microtask crowdsourcing.

Author Keywords

Conversational style; coding scheme; microtasks; crowdsourcing; performance; engagement

CCS Concepts

•Information systems $\rightarrow$ Chat; Crowdsourcing;
•Human-centered computing $\rightarrow$ Empirical studies in HCI;

Introduction

In crowdsourcing marketplaces, large batches of microtasks are concomitant with worker drop-outs during the course of task execution [3]. Several factors have been identified to effect worker engagement and cause task abandonment [8, 12, 13]. To tackle this problem, researchers have introduced conversational agents into the realm of crowdsourcing[1].

Previous works in the field of linguistics and psychology have shown the important role that *conversational styles* have on inter-human communication [6, 10, 11]. Being developed in the context of human conversations, the insights and conclusions of these works are not directly applicable to conversational microtasking, where workers typically aim to optimally allocate their effort rather as opposed to being immersed in conversations. To the best of our knowledge, current conversational agents have not exploited the conversation styles to improve the overall effectiveness of the crowdsourcing paradigm. Understanding the role of conversational styles in human computation can help us better adapt strategies to improve output quality and worker engagement, or better assist and guide workers in the training process. To this end, there is the need for novel methods for the classification of conversational styles in the context of microtask crowdsourcing.

Conversational Style Classification

Emulating particular conversational styles suitable to given contexts, or aligning the conversational style of an agent to the preferred style of workers, may help to improve worker engagement, satisfaction, and even output quality. To enable further research in this direction, we need a reliable method to estimate the conversational style of a worker. We introduce prior work on conversational styles, and present a novel coding scheme designed to label and classify conversation style of workers in conversational microtasking.

[1]We refer to crowd work executed through a conversational agent that can provide workers with a natural way to interact with a crowdsourcing system [1, 4, 5, 7, 9] as *conversational microtasking*.

High Involvement and High Considerateness

In this work we draw from Deborah Tannen's theory of conversational style [10, 11]. In Tannes's theory, conversational style can be classified broadly into two categories: *High Involvement* and *High Considerateness*. The High-Involvement style is described as follows: *"When in doubt, talk. Ask questions. Talk fast, loud, soon. Overlap. Show enthusiasm. Prefer personal topics, and so on"*. In contrast, she characterized the High-Considerateness style as follows: *"Allow longer pauses. Hesitate. Don't impose one's topics, ideas, personal information. Use moderate paralinguistic effects, and so on"*.

As per Tannen's theory, conversational styles emerge through the combined use of different linguistic devices. Tannen identifies nine *dimensions* of linguistic devices that are related to conversational styles: *Personal focus of topic*, *Paralinguistic features*, *Enthusiasm*, *Use of questions*, *Pacing*, *Use of repetition*, *Topic cohesion*, *Tolerance of silence*, and *Laughter* [11].

Coding Scheme of Conversational Style

While providing a conceptual framework for the definition and characterisation of conversational styles, Tannen's theory is not directly applicable to conversational microtasking. Tannen's work was developed (and tested) in the context of human conversations, which are typically long and articulated. In conversational microtasking, devices like "humor" and "the percentage of narrative turns" are clearly at odds with the need for workers to optimally allocate their effort. Moreover, Tannen's continua-based method for conversational style estimation does not have specific criteria to guide readers to distribute speakers on continua. For these reasons, a novel coding scheme for systematically classifying the conversational style is required, to enable the classification of coding styles, and guide the creation of ground truth data for conversation style estimation. This coding scheme builds upon a subset of the linguistic dimensions listed in the previous section. We exclude *Paralinguistic features*, *Use of repetition* and *Laughter*.

We include in the coding scheme *Tolerance of silence*, i.e. hesitation and silence occurring in conversations, but with some adaptation. In text-based chat, we measure tolerance of silence through editing actions (i.e., when users edit a message before it is sent). We calculate the percentage of deleted keys among all the keys pressed by the worker. The higher the percentage is, the more hesitation the worker has, implying longer silence during the conversation.

In our study, *Topic cohesion* refers to whether the answers that workers give to questions are topically coherent, and well linked. In some cases however, workers might directly ask questions to the conversational agent, referring to *4) Use of questions*, or express apologies to explain that they can not answer. Such questions or statements naturally deviate from the topic at hand. Therefore, we combine these two dimensions together as one factor in the coding scheme.

The resulting set of dimensions used to systematically analyze conversation styles are summarised in Table 1, and they include: *1) Personal focus of topic*, *2) Enthusiasm*, *3) Pacing*, *4) Tolerance of silence*, and *5) Topic cohesion & Use of questions*.

Each dimension is quantified using a binary score (either -1 or 1). A final $score$ is used to classify a conversation style as either *Involvement* or *Considerateness*. The score is calculated as a sum of scores corresponding to all the five dimensions. If final $score$ is greater than 0, the conversational style of a worker is classified as *Involvement*. If the final $score$ is less than 0, the conversational style of a worker is classified as *Considerateness*.

The coding scheme can be used for labeling ground truth data of the conversational style. To make the ground truth reliable, the coding process is done by multiple coders independently. First, the scores of all the dimensions are given by all the coders in the group independently. The cases having disagreement will be resolved through manual disambiguation. The reliability of the coding process is measured by Fleiss' Kappa [2].

Experiments

Conversational styles are independently labeled by multiple coders according to the coding scheme to understand how workers' conversational styles distribute among crowd workers. Afterward, we conduct a online crowdsourcing experiments with 180 unique workers and analyze the relationship between workers' conversational styles and their performance, engagement, and cognitive task load.

The coding process is conducted by three experienced experts, who have deeply studied the theory of conversational style and understood the concept of linguistic devices. The inter-rater reliability is measured by Fleiss' Kappa. Three coders are in complete agreement for 124 out of 180 crowd workers. The 56 cases having disagreement are disambiguated manually by coders. In total, 86 workers exhibited *Involved* style, while 94 workers showed *Considerate* style. Therefore the kappa κ value is 0.78.

We explored the behaviour of online workers with two conversational styles during conversational microtasking, and observed strong evidence that conversational style could

Table 1: Coding scheme for conversational style.

Dimension	Score	Criteria
1) Personal focus of topic	1	The worker prefers responding to the questions with personal opinions or personal anecdotes. For example, the worker uses first-person pronouns and phrases such as “I think”, “I like”, “my experience”.
	-1	The worker prefers responding to questions by using objective descriptions. For example, using impersonal phrases such as “it is”.
2) Enthusiasm	1	The worker demonstrates a willingness to converse with the conversational agent. For example, by responding positively to questions from the agent that would prolong the conversation.
	-1	The worker appears to be disinterested in the conversation with the agent. For example, by constantly seeking to end the conversation and responding with “no more”, “nothing else”, or similar phrases.
3) Pacing	1	Calculate the mean *pace* (typing rate) of all the workers. The score of the worker whose mean *pace* $\geq$ *median* is 1 (relatively faster pace).
	-1	Calculate the mean *pace* of all the workers. The score of the worker whose mean *pace* $<$ *median* is -1 (relatively slower pace).
4) Tolerance of silence	1	Calculate the mean *percentage of self-editing* (fractions of deleted characters among all the typed characters) of all the workers. The score of the worker whose mean *percentage of self-editing* $<$ *median* is 1.
	-1	Calculate the mean *percentage of self-editing* of all the workers. The score of the worker whose mean *percentage of self-editing* $\geq$ *median* is -1.
5) Topic cohesion & Use of questions	1	The worker prefers to express opinions directly linked to the topic or asks questions when in doubt.
	-1	The worker deviates from the topic without asking questions, but by responding respectfully to the conversational agent when in doubt.

bear relationship with quality of outcome for difficult tasks. We found that Involvement workers performed better in terms of quality-related outcomes in tasks with higher difficulty levels, with statistical significance. These results suggest that conversational style estimation could be a useful tool for output quality prediction. We found that workers with an Involvement style also tended to report significantly higher scores on UES-SF questionnaire. Analysis of cognitive task load revealed that workers of Involvement style perceived less task load with higher difficulty levels with statistical significance. Results found from the experiment imply the conversational style estimation can be used for worker performance prediction, to better enable adaptive crowdsourcing strategies.

Conclusions

In this work, we explored how the conversational style of a crowd worker could be reliably estimated during conversational microtasking, and the relationship between conversational styles and quality-related outcomes, worker engagement and cognitive task load. We found that workers' *Involvement* conversational style relates with higher output quality, higher user engagement and less perceived task load in tasks with higher difficulty.

REFERENCES

[1] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 33.

[2] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[3] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 321–329.

[4] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 295.

[5] Patrik Jonell, Mattias Bystedt, Fethiye Irmak Dogan, Per Fallgren, Jonas Ivarsson, Marketa Slukova, José Lopes Ulme Wennberg, Johan Boye, and Gabriel Skantze. 2018. Fantom: A Crowdsourced Social Chatbot using an Evolving Dialog Graph. *Proc. Alexa Prize* (2018).

[6] Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.

[7] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 151–162.

[8] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[9] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 243–251.

[10] Deborah Tannen. 1987. Conversational style. *Psycholinguistic models of production* (1987), 251–267.

[11] Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

[12] Ying Zhang, Xianghua Ding, and Ning Gu. 2018. Understanding Fatigue and its Impact in Crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 57–62.

[13] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today?: An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*. 373–382.