The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

クラウドソーシングを利用した 非同期チャットによる対話シナリオ収集方式の提案

Asynchronous Chat System for Dialogue Collection on Crowdsourcing

池田 和史[*1]
Kazushi Ikeda

帆足 啓一郎[*1]
Keiichiro Hoashi

[*1] KDDI 総合研究所
KDDI Research, Inc.

In this paper, we design a crowd-powered system to efficiently collect data for training dialogue systems. Conventional systems assign dialogue roles to a pair of crowd workers, and record their interaction on an online chat. In this framework, the pair is required to work simultaneously, and one worker must wait for the other when he/she is writing a message, which decreases work efficiency. Our proposed system allows multiple workers to create dialogues in an asynchronous manner, which relieves workers from time restrictions. We have conducted an experiment using our system on a crowdsourcing platform to evaluate the efficiency and the quality of dialogue collection. Results show that our system can reduce the necessary time to input a message by 68% while maintaining quality.

1. はじめに

Siri や Google Assistant などの対話エージェントは，高度化するシステムの入力インターフェースとして重要性を高めている[McTear 2002]. 一部の人手によるエージェントシステム[Lasecki 2013a, Huang 2016]を除いて，多くの対話エージェントシステムは，事前に作成したルールに基づいて応答したり [Weizenbaum 1966, Bennacef 1996] ，機械学習によって応答を決定する[Raymond 2007, Henderson 2013]. そのため，網羅的かつ正確な対話を実現させるためには，想定される対話のやり取り(シナリオ)を記述した大規模な文章データ(コーパス)が必要となる.

質問応答システムのように既存のデータをコーパスとして利用できる場合もあるが[Kiyota 2002]，雑談のような一般的な対話を行うには網羅性の観点で課題がある. 適切な応答が既存のコーパス中に見つからない場合，クラウドソーシングなどを利用して不足するシナリオを補う処理が必要となる[Bessho 2012].

クラウドソーシングで募集した 2 名のワーカペアにチャットを行わせることで，コーパスを構築する手法[Lasecki 2013b, Tsukahara 2015]が提案されている. 用途に応じた高品質なシナリオを構築可能なメリットがある一方，2 名のワーカが同時刻に作業をする必要があるという時間的な拘束や，相手の入力中はワーカが待ち状態になるという効率面での課題があった. 所要時間の削減はコストの削減に重要であり[Krishna 2016]，時間単価が上がれば多くのワーカを獲得できる [Mason 2009].

我々はクラウドソーシングに適した対話コーパスの構築システムを設計した. 提案システムでは，従来と同様にチャット形式でシナリオを入力するが，同一のシナリオを複数人で非同期に作成することが可能なため，ワーカを拘束する必要がなくなり，作業効率を向上させることが可能となる.

提案システムの有効性を確認するため，シナリオ作成の所要時間やコストといった効率，および作成されたシナリオにおける話題の多様性や会話のしやすさといった品質について評価した. 提案システムは，既存の 2 名のペアにチャットを行わせる方式と比較して，同程度の品質を保ちつつ，作成に要する時間を68%削減することが確認された.

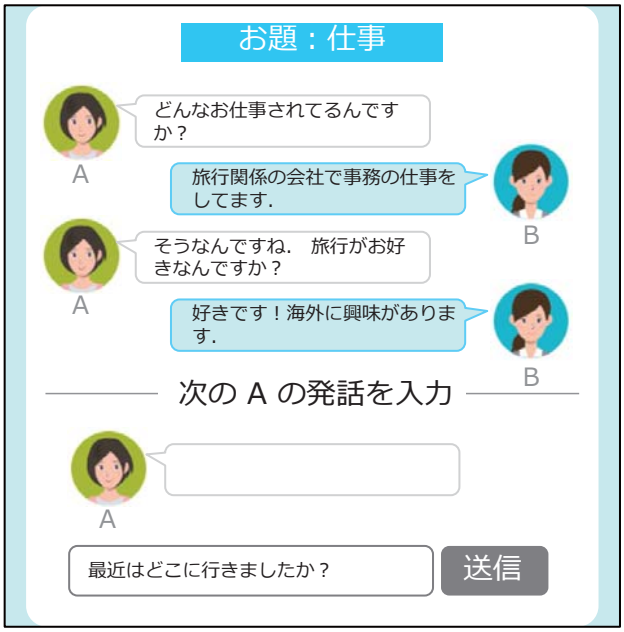

図 1 提案システムのインターフェース

2. 提案手法

本稿で提案するシナリオ作成システムの画面を図 1 に示す. チャット形式のインターフェースに 1 つのシナリオが表示される. 従来手法では 2 名のワーカがペアとなって会話を行うことでシナリオを構築する. 提案手法では，クラウド上の複数のワーカを 2 名の発話者(図 1 の A または B)のいずれかに割り当てる. ワーカがタスクを実施するためにシステムにアクセスすると，途中まで入力された会話が提示される. 図 1 では，役割 A のワーカがシステムにアクセスしている.「仕事」というテーマで A と B がそれぞれ 2 発話を入力しているが，この 4 発話はそれぞれ別の

ワーカが入力したものである. ワーカはこれまでの発話の流れを読んだ上で, 次の発話を入力し, 送信を押す. 入力が完了すると, ワーカには次の別のシナリオが提示される. これを繰り返すことで, ワーカは相手の発話を待つことなく, 次々とシナリオを入力することが可能となる.

シナリオにおける発話が長くなるにつれて, 過去の会話を確認するために要する時間が増加することが想定される. 会話の流れを把握するために必ずしも全ての発話を提示する必要はないため, 直近 N 件の発話(例えば図 1 の場合は 4 件)を提示するものとし, 最適な N の値については実験により評価する.

3. 実験設計

提案手法の有効性を評価するため, クラウドソーシング上で対話シナリオを構築する実験を行う. 以下ではシナリオ作成タスク, 比較手法, 評価指標, ワーカについて説明する.

3.1 シナリオ作成タスク

本稿では, ユーザとの雑談が可能な対話システム用のシナリオ作成を目的とする. 雑談を行う上では, 話題や対話システムの属性(見た目が人間風かロボット風か, など), ユーザとの関係性などを具体化する必要がある. これらは対話システムのコンセプトや利用シーン, 想定するユーザに応じて決定される. 一例として本実験では, 女性のユーザを想定し, ユーザと親しく会話が可能な, 女性の見た目を持つ対話システムをデザインする.

このような対話システムに適したシナリオを収集するため, 知人程度の関係の女性同士で, お互いに良好な関係を構築したい, という設定で 2 名の女性の役割をワーカに割り当てる. 対話システム役は, 年齢, 性別, 家族構成, 職業など 75 項目からなる詳細なペルソナを設定し, ワーカはその役割になり切って会話を行う. ユーザ役は, 同世代の女性という設定のみとし, 会話する上で属性情報が必要となる場合は, ワーカの属性を基準とするよう指示する. これにより, 属性の異なるユーザを想定した対話シナリオを収集できる. また, このような関係においてよく話される話題を 4 名の女性の被験者から募集し, 代表的な 50 のテーマ(休日の過ごし方, 好きなスポーツ, 料理など)について, 30 シナリオずつ, 計 1500 シナリオを構築する.

雑談では, 完了状態を明確に定義することは困難なため, 作成するシナリオの長さは固定長で 16 発話とする. 各発話の入力文字数は 10 文字以上とし, 既存の Web 文章のコピーを行わないこと, などその他の詳細な条件をワーカに指示する.

3.2 比較手法

提案手法と従来手法の比較評価を行う. 加えて, ワーカが一人で二人分の発話を入力する手法も考えられる. これを参考手法と呼び, 提案手法, 従来手法と参考手法の 3 条件で評価を行う. 各手法の特徴を表 1 に示す.

従来手法は, 二人のワーカがペアとなり, 同期して作業を実施する. 各ワーカは役割 A または役割 B のいずれかが割り当てられる. 報酬は待ち時間に対する固定報酬と, 1 発話を入力するごとに出来高報酬が提供される.

提案手法では各ワーカは非同期的にシナリオを入力する. 各ワーカは役割 A または役割 B のいずれかが割り当てられる. 報酬は出来高報酬のみを提供する. 過去 N 発話を表示するが, N=2,6,16 の 3 条件について評価を行う.

参考手法では各ワーカは非同期的にシナリオを入力する. 各ワーカは役割 A と役割 B の両方を実施し, 1 つのシナリオに対して連続して 16 発話分の入力を行う. 報酬は出来高の報酬を提供する.

表 1 比較手法の特徴

手法	同期/非同期	役割	報酬
従来	同期	A または B のいずれか	固定報酬 4 円/分 出来高報酬 4 円/発話
提案	非同期	A または B のいずれか	出来高報酬 4 円/発話
参考	非同期	A と B の両方	出来高報酬 4 円/発話

3.3 評価指標

シナリオ作成の効率とシナリオの品質について評価を行う. 効率の指標として, 1 発話当たりの入力に要する時間を評価する. 従来手法では, 相手の待ち時間を含める. 加えて, 待ち時間に対する報酬を含めた 1 発話当たりのコストを評価する.

品質の指標として, 発話長とシナリオの情報量, 一貫性, 回答しやすさを評価する. 発話長は最低 10 文字以上と指定しているが, ワーカが作業を短時間で終わらせるために, 短い文章を入力する可能性がある. 情報量については, 相槌ばかりの文章や同一の発話を繰り返すよりも, より多様な話題を取り入れた会話が望ましい. シナリオの形態素解析を行い, 一定量の形態素中に含まれる形態素の種類数に基づいて情報量を評価する. 一貫性はシナリオ内の前後の発話に矛盾がないか, 担当する役割のペルソナと比べて矛盾がないか, を 4 名のアノテータに評価させる. 回答しやすさは, システムとの対話を活性化させる上で重要な指標である. シナリオの一部をアノテータに提示し, 次の発話の入力しやすさを 5 段階で評価させる.

3.4 ワーカ

従来手法, 提案手法(N=2,6,16 の 3 条件), 参考手法の計 5 条件について, それぞれ 60 名の参加者を国内の大手クラウドソーシングサイトで募集した. 設定した役割とワーカの属性を合わせるため, 20 代～40 代の女性を募集した. 従来手法と提案手法では, 役割 A に 30 名, 役割 B に 30 名をそれぞれ割り当てた. ここで, 従来手法はワーカがペアを構築する必要があり, 24 時間作業可能な状態では, ワーカのシステムへのアクセスが分散し, 待ち時間が増加することが懸念される. このため, 作業時間を 12 時～15 時に限定した. 実験期間は 2 週間とした.

4. 実験結果

4.1 効率

ワーカ当たりの 1 発話の平均入力時間を表 2 に示す. 従来手法における平均入力時間は 133 秒であった. 提案手法では, 平均入力時間は N の値によって異なり, N=16 の時は 52 秒, N=2 および N=6 の時は, 42 秒であった. 参考手法の平均入力時間は 29 秒であった. 従来手法では, 2 名のワーカがペアで作業をするため, 1 名のワーカがシナリオを入力する時間を t 秒とすると, もう 1 名の待ち時間 t 秒が加算され, 計 2t 秒を要する. この点を考慮しても, 従来手法は提案手法の N=2,6 や参考手法と比較して, 3 倍以上の時間を要している.

Kruskal-Wallis test を用いた有意検定において, 比較した 5 条件間で平均入力時間に有意差が確認された($H=7407$, $p<.01$). ${}_5C_2$ 通りの条件間の組み合わせに対し, Bonferroni 法を用いた名義的有意水準($\alpha'=.005$)において, Wilcoxon rank sum test による有意検定を実施したところ, 入力数 n が十分に大きいため, 全ての条件間で有意差が確認された. 以降, 表中で有意差を示している項目については同様の方法で有意検定を実施した.

表 2 ワーカ当たりの平均入力時間

	平均 (秒)	標準偏差	件数 *n*	p 値
従来	133	124	16,704	
提案(N=16)	52	51	22,500	
提案(N=6)	42	43	22,500	$2.2e^{-16}$ **
提案(N=2)	42	47	22,500	
参考	29	38	21,447	

表 3 作成された発話数, シナリオ数と費用(円)

	発話数	シナリオ数	総費用	総費用/発話数
従来	16,704	1,058	175,148	10.5
提案(N=16)	22,500	1,500	90,000	4
提案(N=6)	22,500	1,500	90,000	4
提案(N=2)	22,500	1,500	90,000	4
参考	21,447	1,500	85,788	4

表 4 平均発話長

	平均 (文字)	標準偏差	件数 *n*	p 値
従来	28.3	12.9	16,704	
提案(N=16)	24.2	10.2	22,500	
提案(N=6)	20.7	7.7	22,500	$2.2e^{-16}$ **
提案(N=2)	22.3	8.9	22,500	
参考	19.2	10.5	21,447	

表 3 は作成された発話数, シナリオ数, 総費用, 発話当たりの費用を表す. 提案手法と参考手法では, 2 週間の実験期間内に 50 トピックについて各 30 シナリオを構築することができた. 従来手法では, 作業時間を限定したために, 一部のシナリオは期間内に完成しなかった. 従来手法における 1 シナリオ当たりの待ち時間の平均は 93 秒であった. これに加えて, 相手の入力を待つ時間に対しても 1 分当たり 4 円の報酬が提供されるため, 発話当たりの報酬は平均で 10.5 円となった. 出来高報酬が 4 円であることと比較すると, 高コストといえる.

4.2 品質

表 4 は平均発話長を表す. 従来手法が最も長く 28.3 文字, 参考手法が最も短く 19.2 文字であった. 参考手法では, N の値による顕著な傾向は見られず, 従来手法と参考手法の中間であった.

表 5 は各手法で作成されたシナリオの情報量と一貫性の欠如率を表す. 情報量はシナリオ中の 10 万形態素あたりに出現する形態素の種類数を算出しており, 参考手法が少ない値となった. 参考手法では, 相槌のみの発話や指定されたトピックから逸脱して, 他のトピックと同じ内容を使い回すワーカも見られた.

一貫性の欠如については, 提案手法の N=2 や参考手法でやや増加する傾向が見られた. 前後の文脈と矛盾する場合と, 設定したペルソナと矛盾する場合に分けてアノテーションを実施しており, 提案手法(N=2)では前者が, 参考手法では後者が多く見られた. 提案手法(N=2)では発話の文脈が十分に提示されないために矛盾が生じたと考えられる. 参考手法では, ワーカが一人二役を担うため, 混乱した可能性がある.

表 6 はシナリオの回答しやすさを表す. 各手法で作成したシナリオからランダムに 100 件を選択し, 各シナリオの最初の発話から途中の発話までを 4 名のアノテータに提示し, 次の発話の入力の容易さを 5 段階で評価させた. 参考手法において, 平均スコアの低下が確認された. 参考手法のシナリオを確認したところ, 作成したワーカ以外には継続が困難なシナリオが見られた.

表 5 形態素の種類数(情報量)と一貫性欠如率

	形態素の種類数	一貫性欠如率
従来	1,752	0.07
提案(N=16)	1,759	0.09
提案(N=6)	1,708	0.03
提案(N=2)	1,731	0.14
参考	1,583	0.12

表6 回答のしやすさ

	平均	標準偏差	件数 *n*	p 値
従来	3.13	1.28	400	
提案(N=16)	3.18	1.17	400	
提案(N=6)	3.10	1.20	400	$2.2e^{-16}$ **
提案(N=2)	2.93	1.06	400	
参考	2.79	1.31	400	

5. 考察

提案手法における N の最適値については, N=2, 6 と比べて N=16 の入力時間が増加した. これは過去の発話を読むために時間を要したためと考えられる. 効率面では N=2, 6 が有望であるが, 品質面では N=2 において, 一貫性が低下する課題が見られた. 総合的に判断すると, N=6 が最適と考えられるため, 以降は提案手法(N=6)について考察を行う.

効率面では, 提案手法は従来手法と比較し, 所要時間を 68%, コストを 62%削減可能であることが確認された. これは, 相手の入力を待つ時間を削減したことによる効果が大きい. 実験後に各手法を担当したワーカに(1)報酬の満足度, (2)作業を楽しいと感じたか, (3)もう一度同じ作業を受託したいかをアンケートにより 7 段階で回答させた. 表 7 に回答をまとめる. 提案手法および参考手法における報酬の満足度および再受託を希望する度合いは従来手法よりも高いことが確認された. 本結果は, 提案手法で作業を実施したワーカは十分な報酬を得たと感じていることを示しており, 報酬単価をさらに下げることや従来手法よりも多くのワーカを参加させることも可能と考えられる.

品質面では, 参考手法は情報量や回答しやすさといった対話シナリオに重要な指標において, 従来手法よりも品質の劣化が確認された. 提案手法では, 発話長を除いて従来手法と同等の品質が得られていることが確認された. 表 8 に各手法で作成されたシナリオの例を示す. 従来手法では, 役割 B の 3 発話目において, 直前の発話をさらに掘り下げる発話が見られた. 従来手法では, 同じワーカペアが会話を行うため, 前の発話の情報を補足するような発話が見られた. また, リアルタイムに会話を行うため, 相槌などが多くなる傾向も見られた. 提案手法では, 過去の発話を踏まえた発話が行われるが, 背景となる情報を他のワーカと共有できていないため, 過去の発話を掘り下げるような発話はあまり見られなかった. これらの点が発話長に影響したと考えられる.

一方, 参考手法では, 10 文字以上の制約を満たすために不要な相槌を入力したり, 指定されたトピックを逸脱して同じシナリオを繰り返して入力するワーカが複数見られた. 従来手法や提案手法のように他者との協調作業ではないため, 他者が回答しやすい内容や自然な会話を行うといった意識が低下したものと思われる. ワーカが作業時間を短縮するために入力パターンを定型化した結果, 情報量の欠如や他者にとって回答しにくいシナリオが作成されるといった致命的な問題を発生させている.

表 7　ワーカへのアンケート結果

	従来	提案	参考	p 値
報酬の満足度	**4.63**	**4.99**	**5.45**	**0.047***
作業の楽しさ	5.88	5.92	6.11	0.892
再度受託したい	**6.22**	**6.54**	**6.70**	**0.042***
件数 n	32	39	44	

表 8 トピック「好きな街」で作成されたシナリオの例

従来手法で作成されたシナリオ

A: いつもどのあたりで買い物しますか？
B: だいたいは，勤務先の品川か地元横浜ですねぇ
A: 好きなお店があるのですか？？
B: 服でいえばアースミュージックエコロジーが好きですねっ
A: 私はファッションのことに疎いのですが，どんな系統のファッションですか？
B: ん～～なんていうんだろう‥どんな場所でも対応できるような清楚というかそういった服装です 笑
A: 清楚女子憧れますね～！！何処でも対応できるのは便利ですしね！私はいつもネットで服を買っているので‥
B: ネットでどういった服を買ってらっしゃるんですか？

提案手法で作成されたシナリオ

A: いつもどのあたりで買い物しますか？
B: 勤務先の品川あたりが多いです．
A: そうなんだ．品川辺りでオススメのショップは？
B: 好みがありますからね．何処のお店も良いですよ！
A: そうなんですね．品川はオフィス街のような認識でした．
B: 私も最近は仕事終わりにブラブラしたりしています．
A: 仕事帰りにブラブラできていいですね．私は最近忙しくって．
B: 駅ビルが充実してるので，フラっと行けるのが気楽でいいんですよね．

参考手法で作成されたシナリオ

A: いつもどのあたりで買い物しますか？
B: その時によりますねー
A: へぇ．そうなんですね?
B: A さんはどうですか？
A: 私も似たような感じです
B: なるほど．そうなんですね
A: そうなんですよね?.
B: 私たち，気があいますね

6. まとめと今後の課題

本稿では，クラウドソーシングを利用して効率的に対話シナリオを作成する方式を提案した．従来のリアルタイムチャットによるシナリオ作成手法では，ワーカを時間的に拘束することが課題であった．これに対し，提案方式では同一のシナリオを複数人で非同期に作成可能とすることで，作業効率を向上した．実験の結果，提案手法は方式と同程度の品質を維持しつつ，所要時間を 68%，コストを 62%削減可能なことを確認した．

今後の課題として，他のワーカの作業状況を引き継ぐ仕組みの導入を検討している．これにより，より踏み込んだ内容の雑談や，タスク型の対話シナリオの作成など，多様なシナリオの作成に本システムの適用が可能になると考えられる．

参考文献

[Bennacef 1996] Bennacef, S., Devillers, L., Rosset, S., and Lamel, L.: Dialog in the RAILTEL Telephone-Based System. In Proceedings of ICSLP, pp. 550-553, 1996.

[Bessho 2012] Bessho, F., Harada, T., and Kuniyoshi, Y.: Dialog System Using Real-Time Crowdsourcing and Twitter Large-Scale Corpus. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 227-231, 2012.

[Henderson 2013] Henderson, M., Thomson, B., and Young, S. J.: Deep Neural Network Approach for the Dialog State Tracking Challenge. In Proceedings of SIGDIAL Conference, 467-471, 2013.

[Huang 2016] Huang, T. H. K., Lasecki, W. S., Azaria, A., and Bigham, J. P.: "Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, 2016.

[Kiyota 2002] Kiyota, Y., Kurohashi, S., and Kido, F.: Dialog Navigator: A Question Answering System based on Large Text Knowledge Base. In Proceedings of the 19th International Conference on Computational Linguistics, 1: 1-7, 2002.

[Krishna 2016] Krishna, R. A., Hata, K., Chen, S., Kravitz, J., Shamma, D. A., Fei-Fei, L., and Bernstein, M. S.: Embracing Error to Enable Rapid Crowdsourcing. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 3167-3179, 2016.

[Lasecki 2013a] Lasecki, W. S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J. F., and Bigham, J. P.: Chorus: A Crowd-Powered Conversational Assistant. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, 151-162, 2013.

[Lasecki 2013b] Lasecki, W. S., Kamar, E., and Bohus, D.: Conversations in the Crowd: Collecting Data for Task-Oriented Dialog Learning. In Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, 2013.

[Mason 2009] Mason, W., and Watts, D. J.: Financial Incentives and the "Performance of Crowds". In Proceedings of HCOMP, 77-85, 2009.

[McTear 2002] McTear, M. F.: Spoken Dialogue Technology: Enabling the Conversational User Interface. ACM Computing Surveys (CSUR), 34(1): 90-169, 2002.

[Raymond 2007] Raymond, C., and Riccardi, G.: Generative and Discriminative Algorithms for Spoken Language Understanding. In Proceedings of Interspeech, 1605-1608, 2007.

[Tsukahara 2015] Tsukahara, H., and Uchiumi, K.: System Utterance Generation by Label Propagation over Association Graph of Words and Utterance Patterns for Open-Domain Dialogue Systems. In Proceedings of 29th Pacific Asia Conference on Language, Information and Computation, 323-331, 2015.

[Weizenbaum 1966] Weizenbaum, J.: ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. Communications of the ACM, 9(1): 36-45, 1966.