ゲーミフィケーションを利用した対話ログ収集における応答文の改善と対話ログの解析

叶内 晨*，尾形 朋哉，金子 正弘，河村 綾菜，
北川 善彬，黒田 紘司，齋藤 宏行，山本 豊，小町 守

首都大学東京

1 はじめに

近年，大規模なデータ収集が可能になると共に，データドリブンな対話システムの研究･開発が盛んに行われている．非タスク指向型対話システムにおいては，Twitter において Tweet と Reply の関係を大量に収集してきて対話コーパスとして利用する研究 [5, 8] や大規模な映画の字幕データを対話コーパスとして利用した研究 [1, 3] がある．しかし，英語では対話コーパスが充実しているものの，日本語における共通して利用可能な大規模な対話コーパスは存在していない．NTCIR Short Text Conversation 日本語タスクでは 100 万件の Tweet ID を公開しているものの，Twitter には誤字・脱字，文法誤りなどのノイズが数多く含まれる問題 [6, 10] や，単純に発話文と応答文のペアを収集した場合には，どの発話文に対してどの応答文が良かったのか，もしくは悪かったのかという教師データがなく，対話システムを定量的に評価するのが難しい問題がある．

一方，ゲーミフィケーションを利用することにより，対話ログを収集する研究がある [4, 7, 9]．叶内ら [9] は，自らチャットボットを作成してみたい人をターゲットに，容易にチャットボットを作成することのできるプラットフォームの構築を提案した．その際，ゲーミフィケーションを利用しユーザにチャットボットの育成してもらうことで，ログデータから対話破綻ラベル付きの対話ログを収集するシステムを提案した．しかし，生成した応答文候補の質の問題があった．

そこで本研究では，クラウドソーシングを利用することで，叶内ら [9] の対話ログ収集システムの応答文候補の質を改善し，評価する．さらに，実験協力者によって予備的に得られた雑談対話ログの解析を行うことで，各ユーザによる対話ログの揺れの問題について議論する．

*shin187nlp@gmail.com

2 ゲーミフィケーションを利用した対話ログ収集システムの概要

日本語における共通して利用可能な大規模な対話データは公開されていない．叶内ら [9] はゲーミフィケーションを利用することで，日本語における（1）大規模で（2）公開可能な（3）ラベル付きの雑談対話コーパスの開発を目指した．ゲーミフィケーションによってユーザにチャットボットの育成をしてもらう副産物として，雑談対話コーパスを作成するシステムを提案している．ゲームにおいて，各ユーザは与えられた発話文に対する応答文を選択していくことで，自分のチャットボットの応答の幅を増やしていく．スマートフォンでのプレイを想定し，タップだけで効率よくゲームが成り立つように，システム側であらかじめ応答文の選択肢を用意している．

図 1 に，実際にゲームにおいて応答文選択を行っている画面を示す．応答文を選択することで真ん中の画面から左の画面へ移行し，選択完了を押すことで育成データが追加される．適切な応答文が存在しない場合にはユーザが “応答文を自分で作成” を押すことで，自分で応答文を作成することができる．この操作を繰り返し育成データをユーザ毎に蓄積することで，ユーザ自身の学習データによるチャットボットが完成する．

しかし，叶内らのシステムは応答文の選択肢の質が低く，実用レベルに達していない問題があった．そこで本研究では，クラウドソーシングを利用することで応答文の質を改善し，評価する．その後，実験協力者により得られた対話ログの解析を行い，各ユーザによる対話ログの揺れの問題について議論する．

3 応答文の質の改善

叶内ら [9] の研究ではデータの公開を最終目標においているため，著作権に配慮し，応答文の生成を全て DoCoMo の雑談対話 API に頼っていた．表 1 に DoCoMo 雑談対話 API によって生成された複数の応答文を示す．例 1 と例 2 は良い応答文を含む例であり，発話文の特定

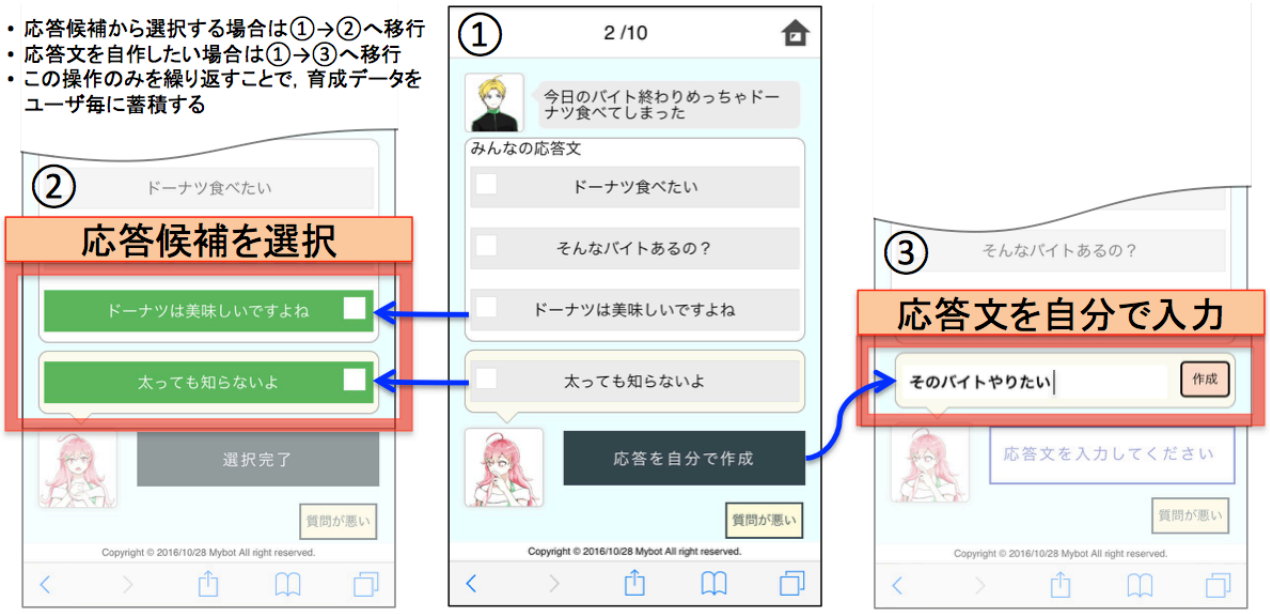

図 1: チャットボットの育成画面

表 1: DoCoMo 雑談対話 API と提案手法による応答文

番号	入力文	DoCoMo 雑談対話 API による応答文	提案手法により選択された応答文
例 1	**アーモンド**を定期的に摂取している	栄養価が高いだけでないですか 栄養価が低いのに高いですか 栄養価は高いですか 栄養価が高いだけでないのでしょうか？	アーモンドは大好物だよ 小さい頃食べすぎると鼻血出なかった？ それテレビでモデルが言ってた あとどれくらいでできるかな。
例 2	**アイス**を考えました	アイスはハーゲンダッツが好きですね なめらか系が好きですね greentea が好きですね アイスは好きですね	私もアイス食べたいよ こんな寒い季節にアイスはないよ 締めのアイスは美味しいよね 私も新しいアイスを見るとすぐ買っちゃう
例 3	**アイス**を買って冷凍専用に詰め込んだ	寒が過ぎますか ?? 寒が続くんですか？ 寒に耐えますか？ 寒が緩みます ??	締めのアイスは美味しいよね アイスはヤバいでしょう ‥ アイスの実はみかんがおいしいのにね アイスはそうだけどチーズは飲めない

のキーワードに対応する応答文を生成している．しかし，発話文に対してバラエティの富んだ複数種類の応答文は生成できておらず，全ての応答文が似通う傾向があった．例 3 は全ての応答文の意図がよく分からない例である．特定のキーワードを含む文に対しては，毎回同じように誤った応答文を生成する傾向が見られた．

本研究で改善を行うゲーム [9] は，発話文に対する応答文を選択して育成するゲームである．そのため，応答文候補が似通っていたり，全て明らかに間違えである場合にゲームの質が低いと感じる，というユーザ体験があった．しかし，DoCoMo 雑談対話 API の中身はブラックボックスであるため，改善することができない．

そこで本研究では，クラウドソーシングを利用することにより，ドメインを限定した発話文に対する応答文を一定量作成し，データベースに格納することでゲームに利用した．また，ユーザがゲーム内で過去に学習した結果を反映させるため，そのユーザのチャットボット育成のために蓄積したデータベースから応答文候補を選択するモデルを作成した．

3.1 クラウドソーシングによる応答文作成

本研究では，外部 API によらず応答文候補を一定数用意するために，発話文に対する応答文をクラウドソーシングにより作成した．発話文は Tweet を元にし，叶内ら [9] の前処理により生成した．クラウドソーシングはランサーズ [1] のタスク型のデータ入力に設定し，応答文をワーカーに入力してもらった．タスクを依頼する際には，ワーカーに以下の規則を提示した．

- SNS において，友達とチャットをしていることを想定して応答文を入力
- 8 文字以上 40 文字以下で入力
- 絵文字・顔文字の使用は禁止
- 標準語推奨で，敬語は禁止
- 一人称は「私」に統一
- 人名の使用は禁止（e.g. 太郎君）
- 一般的な人を指す表現は使用可（e.g. 友達，母）

キャラクター性を保つため，応答文の人称などはある

[1] http://www.lancers.jp

程度統一した．4,000 件の発話文に対してそれぞれ 3 人ずつに応答文を入力してもらうことで，合計 12,000 件の発話文と応答文のペアを作成した．なお，1 件あたり 3 円で作成を依頼し，38,880 円を要した．

クラウドソーシングを利用することにより，対話を行うドメイン毎に同様の作業が必要となるため，コストとのトレードオフの問題がある．しかし，本研究では，100 万件規模の全てのデータをクラウドソーシングによって作成した場合を高コストとした上で，ゲームを成立させるための最小限のコストは必要であると判断した．

3.2 データベースからの応答文選択

クラウドソーシングによって収集した発話文と応答文のペアをデータベースに格納することで，未知の発話文に対して，データベースからスコアリングし応答文を返した．スコアリングは，未知の発話文とデータベースの発話文において内容語のみの編集距離を計算し，最もスコアの小さい発話文から順に対応する応答文を出力した．なお，最小スコアが複数存在する場合は，さらに文字単位の編集距離を計算することで応答文を出力した．

表 1 にクラウドソーシングを利用した際の応答文候補を示す．クラウドソーシングにより 12,000 件の応答文を作成することで，発話文に対してバラエティに富んだ応答文を出力できている．

図 1 において，“みんなの応答文” とある 3 つの選択肢は，クラウドソーシングを利用して得た応答文である．4 つ目の応答文である「太っても知らないよ」は，そのユーザの過去の発話文と応答文の育成データを元に，スコアリングは同様な編集距離の計算によりユーザ自身のベストな応答文を出力している．これにより，自分の育成しているチャットボットが常にどのような応答をするのかを把握することができる．

3.3 応答文の質の評価

応答文の質がどの程度改善されたか，評価者による定量的な評価を行った．ある発話文に対して，先行研究と本研究による応答文をそれぞれ提示し，どちらのほうが適切に応答できているか選択してもらった．選択してもらう際，各応答手法は明記していない．選択肢として “A のほうが適切”，“B のほうが適切”，“どちらも同程度” の 3 つを用意した．評価者 2 名によって 50 件のデータを評価してもらった結果，提案手法のほうが良い結果となった（提案手法：先行研究：同程度 = 51 : 16 : 33）．評価者 2 名による *kappa* の一致率は $\kappa = 0.51$ であった．

4 収集したラベル付き対話ログの解析

本論文では 4 人の実験協力者に一定回数ゲームをプレイしてもらい，それによって得られたラベル付きの対話ログについて解析を行った．今回は実際のゲームプレイを想定したため，プレイする上で特定の指示は与えていない．図 1 におけるチャットボットの育成において，応答文を複数個選択可能である．これにより，発話文に対する応答文が絶対的に正しいかのラベル付きの対話ログを獲得した．実験協力者 4 人に，それぞれ発話文に対して応答文の選択を 2,000 件してもらうことで，合計 8,000 件の応答文を得た．聞き取り調査によると，プレイ時間は平均 2 時間であった [2]．

表 2: 実験協力者毎のゲームログ

実験協力者	適切な応答文（選択） [%]	不適切な応答文（非選択） [%]	質問が悪い [%]
A	46.4	46.7	6.9
B	37.1	53.2	9.7
C	29.4	68.7	2.0
D	25.6	73.6	0.8
平均	34.6	60.5	4.8

4.1 ユーザ毎の対話ログについて

本システムのゲーミフィケーションを利用することで自動で対話の成立・破綻ラベルを収集することが可能である．しかし，クラウドソーシングとは違い，ユーザはゲームにおいてアノテーションを全く意識しない．そのため，ゲームの設定次第で，獲得したいラベルと実際に収集されるログデータにはずれが生じると考えられる．

実験協力者毎の選択結果の比率を表 2 に示す．協力者 A の “適切な応答文” の比率は 46.4%であり，本ゲームでは毎回 4 件の応答文を提示しているため，毎回平均 2 件の応答文を選択している．一方，協力者 D の “適切な応答文” の比率は 25.6%であり，ユーザ毎に適切な応答文の選択比率の差が大きい．これにおいて，ユーザ毎の適切な応答文の許容範囲が違うという問題の他に，1 つ良い応答文があったら他の応答文の正しさが気にならなくなるというユーザ体験があった．解決策として，応答文選択の定義を見直すか，提示する応答文の候補数を減らすことで，ユーザ毎の選択数の比率が近づき，ラベルの質が向上すると考えられる．なお今回のアプリでは実装において，“自分のチャットボットに応答してほしい候補を選択する（複数選択可）” と定義した．

“質問が悪い” の欄を見ると，協力者 B は全体のうち 9.7%が発話文が不適切であると選択しているのに対して，協力者 D は 0.8%であり，その差は大きい．協力者からのフィードバックとして，ユーザ毎に発話文が悪いと思うか，もしくは，それがゲームの仕様だと考えて応答文が悪いと思うのかに違いがあることがわかった．例えば，意味不明の発話文に対して，単純に発話文が悪いとする以外に,「どういう意味？」などの応答文を作成する選択肢がある．

[2] ゲームを行う場所と時間帯は実験協力者の自由とした．

表 3: ユーザにより入力された応答文の例

応答文の例
美味しくできた?，ありがとう，それどんなもの?，お金持ちだね，お茶入れよう，辛いの苦手？，めで鯛

4.2 応答文の増加

発話文は無限に収集可能であるが，応答文はクラウドソーシングによるシードと，ユーザによる入力でしか増えない．今回，協力者 4 人によって入力された応答文の数は合計で 122 件であった．その例を表 3 に示す．ユーザ経験として，この量の応答文の入力であればゲームを進行する上で妨げにならないことがわかった．しかし対話コーパスを構築する上では，更なる応答文候補の作成が望まれる．今後の展望として，テンプレートの自動生成や，Sequence-to-Sequence などを利用した文生成により，応答文候補を自動で生成したい．

4.3 応答文における相づちの問題

本ゲームの最終的な目的は，他人のチャットボットよりも良い応答をする自らのチャットボットを作成することである．この目的を達成する上で，しばしば相づちが良い応答候補となる．少量の万能な相づちデータの入力によりゲームが成立していまう場合，ユーザにとって大量の育成データを作成するモチベーションは低下する．また表 3 に示すように，相づちでなくても，ユーザからの入力には万能な応答文が数多く見られた．そのため今後の展望として，ゲームにおいて相づちや万能な応答が使いにくくなる制約を設定する必要がある．

5 関連研究

日本語における，応答文の破綻をアノテーションした中規模な対話コーパスとして，Project Next NLP 対話タスクで収集された雑談対話コーパス [12] がある．雑談対話コーパスは人と対話システムにおける約 2 万ペアのコーパスで，対話破綻のアノテーションが付与されている．しかし対話生成において 2 万件は少なく，さらに大きい対話コーパスが必要である．

クラウドソーシングを利用することで，対話データを構築する研究がある [2, 7, 11]．Inaba ら [7] は，ある発話文に対してどの応答文が正しいかをクラウドソーシングを利用して選択する際に，選択結果からアノテータの対話力を診断するゲーム機能を導入することで，データの作成と品質管理を同時に行った．塚原ら [11] は人間同士の対話において，ワーカーに対話入力と同時にアノテーション付けと校正作業を同時に行う仕組みを提案した．しかし，大規模なデータを全てクラウドソーシングで作成するためには，金銭的なコストが必要となる．Bessho ら [2] はデータベースに適切な応答文が見つからない場合に，リアルタイムでクラウドソーシングを利用することで，ワーカーに応答文を作成させた．本研究においてもクラウドソーシングは利用するものの，ゲーミフィケーションによりコストを抑えつつ大規模なコーパスの構築を目指す．

6 おわりに

本論文では，叶内ら [9] によって実装された対話ログ収集システムの応答文の質を改善し，その後，実験協力者により得られた対話ログの解析を行った．応答文を改善するために，クラウドソーシングを利用することでドメインに対応した応答文を作成し，ゲームを成立させた．さらに，被験者による対話ログを解析を行うことで，ユーザ毎のデータの揺れと相づちの問題について議論した．今後の展望として，ゲーミフィケーションであることの利点を活かし，対話破綻ラベル以外の感情ラベルなどの自動獲得を行えるシステムを構築すると共に，ゲームを一般公開することで，大量の対話ログを収集したい．

参考文献

[1] Rafael E. Banchs. Movie-dic: A movie dialogue corpus for research and development. In *ACL*, pp. 203–207, 2012.

[2] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. Dialog system using real-time crowdsourcing and Twitter large-scale corpus. In *SIGDIAL*, pp. 227–231, 2012.

[3] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *ACL*, pp. 76–87, 2011.

[4] Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. Using role play for collecting question-answer pairs for dialogue agents. In *INTERSPEECH*, pp. 1097–1100, 2013.

[5] Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. Building a conversational model from two-tweets. In *ASRU*, pp. 330–335, 2011.

[6] Ryuichiro Higashinaka, Nozomi Kobayashi, Toru Hirano, Chiaki Miyazaki, Toyomi Meguro, Toshiro Makino, and Yoshihiro Matsuo. Syntactic filtering and content-based retrieval of Twitter sentences for the generation of system utterances in dialogue systems. In *Situated Dialog in Speech-Based Human-Computer Interaction*, pp. 15–26, 2016.

[7] Michimasa Inaba, Naoyuki Iwata, Fujio Toriumi, Takatsugu Hirayama, Yu Enokibori, Kenichi Takahashi, and Kenji Mase. Constructing a non-task-oriented dialogue agent using statistical response method and gamification. In *ICAART*, pp. 14–21, 2014.

[8] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of Twitter conversations. In *NAACL HLT*, pp. 172–180, 2010.

[9] 叶内晨, 小町守. ゲーミフィケーションを利用した効率的な対話ログ収集の試み. 信学技報, Vol.116, No.379, NLC2016-30, pp. 7–12, 2016.

[10] 稲葉通将, 神園彩香, 高橋健一. Twitter を用いた非タスク指向型対話システムのための発話候補文獲得. 人工知能学会論文誌, Vol. 29, No. 1, pp. 21–31, 2014.

[11] 塚原裕史, 内海慶. オープンプラットフォームとクラウドソーシングを活用した対話コーパス構築方法. 言語処理学会第 21 回年次大会発表論文集, pp. 147–150, 2015.

[12] 東中竜一郎, 船越孝太郎, 荒木雅弘, 塚原裕史, 小林優佳, 水上雅博. テキストチャットを用いた雑談対話コーパスの構築と対話破綻の分析. 自然言語処理, Vol. 23, No. 1, pp. 59–86, 2016.