クラウドソーシングで利用可能な日本語対話収集基盤

児玉 貴志 † Frederic Bergeron† 新 隼人 † 田中 リベカ † 坂田 亘 ‡† 黒橋 禎夫 †
† 京都大学 ‡ LINE 株式会社
{kodama,bergeron,atarashi,tanaka,kuro}@nlp.ist.i.kyoto-u.ac.jp
wataru.sakata@linecorp.com

1 はじめに

近年の対話システムの研究では，Twitter や Reddit 等の大規模対話データで事前学習を行い，それぞれのタスクに合わせた小規模対話データセットでファインチューニングを行うという手法が一般的になりつつある．この小規模対話データセットは，小規模と言えど，数万発話程度の規模が必要となるため，クラウドソーシングを利用して集められることが多い．

英語の対話研究では，Facebook が提供する ParlAI [1)] において，対話を行うワーカをリアルタイムでマッチングし，マッチングしたワーカ間の対話を収集・保存するフレームワーク（対話収集基盤）が公開されており，Amazon Mechanical Turk で対話収集のタスクを容易に実施できる．また，この対話収集基盤によって収集されたデータセットも数多く公開されており [1, 2, 3, 4]，対話研究の発展に大きく寄与している．しかし，Amazon Mechanical Turk の対応言語は英語のみであり，日本語の対話を収集するのは難しい．対話データは対話研究の礎である．対話データの分析を積み重ねることで新たな知見が得られるため，対話データ収集の重要度は高い．

そこで，我々はクラウドソーシングで日本語の対話収集を容易に行える対話収集基盤を構築し，公開した[2)]．本対話収集基盤は対話収集用サーバを用意するだけで使用することができる．指定した対話収集用 URL にワーカがアクセスするとペアのマッチングが行われ，ワーカがリアルタイムで対話するチャットルームの生成，およびそこでの対話の保存までを自動で行うことができる．また，ペアのマッチングおよびリアルタイム対話の実現という開発コストの高い対話収集システムの基盤部分を提供しており，少量の追加実装のみでタスク実施者が収集したい対話の形式に合わせてカスタマイズすることも可能である．この対話収集基盤を用いることで，タスク実施者は収集したい対話の形式や内容の検討のみに注力することができる．一方で，ワーカは URL をクリックするだけで対話収集に参加でき，対話のやりとりのためのアプリケーションやツールをインストールする必要がないため，ワーカ側の負担も少ないツールとなっている．また，特定のクラウドソーシングサービスに依存しておらず，汎用的に利用できる．

本稿では，公開した日本語対話収集基盤を利用した対話収集の流れを説明し，この対話収集基盤を実際に利用したクラウドソーシングでの対話収集事例を報告する．

2 関連研究

クラウドソーシングを用いた日本語の対話収集はこれまでにも取り組まれてきたが，「どのようにワーカ同士のマッチングを行うか」「どのように対話のやり取りを行う環境を提供するか」の 2 つが大きな課題とされてきた．

これらの課題に対して，2 者間の対話を，与えられた対話履歴に対する次の発話を作成するタスクに分解することで，複数の異なるワーカによって 1 つの対話データを作成する研究がある [5, 6]．不特定多数のワーカが並行して作業を進められるため所要時間の削減等のメリットがある一方で，あくまで疑似対話データであるため通常の対話とは違った性質を持っていると考えられる．

また，複数のアプリケーションやツールを組み合わせることで対話を収集するアプローチもある．塚原・内海 [7] はクラウドソーシングサイトで集めたワーカに Slack [3)] のチャンネルのアクセス用 URL をメールすることでワーカのマッチングを行い，対話を収集している．東中ら [8] は誰でも編集可能な Google Drive のスプレッドシートを用い，各ワーカに Skype の ID と対話可能な時間を記入してもらうことでマッチング及

1） https://github.com/facebookresearch/ParlAI
2） https://github.com/ku-nlp/ChatCollectionFramework
3） https://slack.com

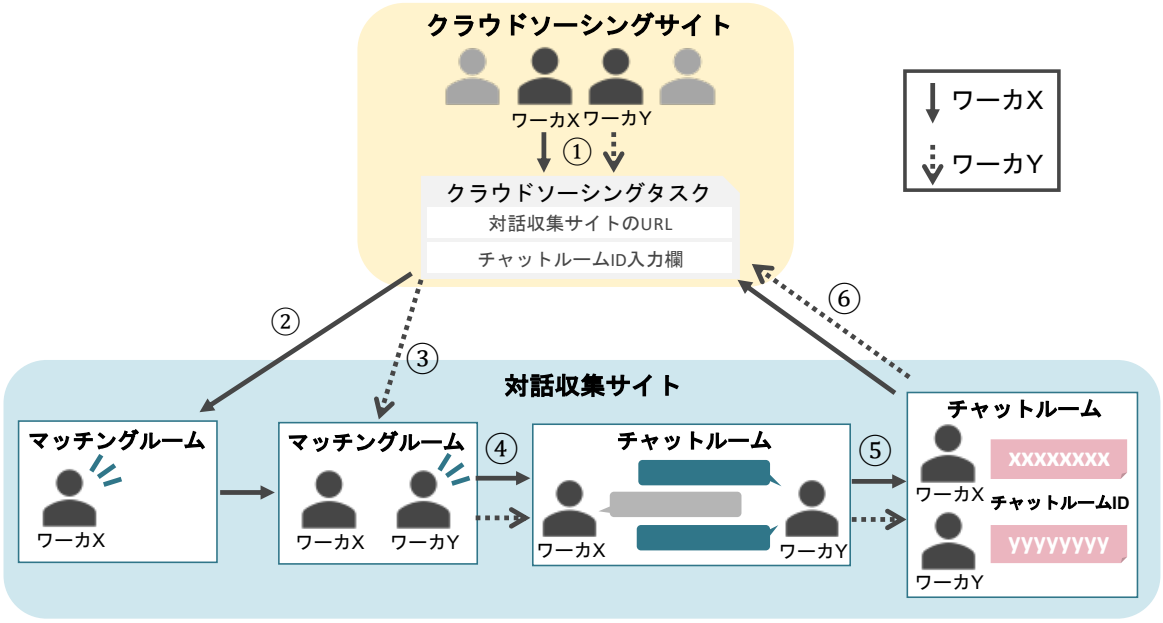

図 1　クラウドソーシングを利用した対話収集の流れ

び対話収集を実現している．ただ，こうした複数のツールを使う手法ではワーカ側にはアカウント作成やツールのインストールの手間が，タスク実施者側には複数のツールを管理する手間がかかり，負担が大きいと考えられる．

本稿で提案する対話収集基盤では，実際にワーカ 2 人がリアルタイムで対話をするため，疑似ではない通常の対話を収集できることに加え，ワーカのマッチング・対話・データ保存までを 1 つのツールで行えるため，ワーカとタスク実施者両者の負担が少ない．

3　対話収集の流れ

まず下準備として対話収集基盤の雛形から対話収集サイトを作成する．ワーカ同士が雑談をするという最も基本的な対話収集の設定の場合は設定ファイルを編集するだけで作成することができる．設定ファイルには対話データの保存場所，各ワーカの最低発話数，マッチングルームでの最大待機時間（後述）などの項目があり，容易に変更ができる．またチャットルームの HTML ファイルを変更すれば，タスク実施者が収集したい対話の形式に合わせてカスタマイズすることも可能である．

次にクラウドソーシングタスクを設計する[4]．クラウドソーシングタスクには作成した対話収集サイトの URL を掲載しておくのに加え，チャットルーム ID（詳細は後述）の入力欄を準備しておく．ワーカを集める目的でのみクラウドソーシングを利用するため，非常に簡潔なタスク設計となっており，一般的なクラウドソーシングサービスにも適用できる．

続いて実際の対話収集の説明に入る．図 1 にクラウドソーシングを利用した対話収集の流れを示す．対話収集のためのワーカはクラウドソーシングサイトを通じて集める（①）．タスクに参加するワーカ（ワーカ X，ワーカ Y）はタスクに掲載されている対話収集サイトの URL をクリックし，マッチングルームに入る．マッチングルームには最大で 1 人が待機している．まだ待機しているワーカがいなければ，ペアとなるワーカが来るまでマッチングルームで待機する（②）．対話収集サイト作成時に最大待機時間を設定することができる．最大待機時間を過ぎた場合は「現在，他のユーザーがいません．後でもう一度試して下さい．」というメッセージが表示され，マッチングルームから自動でワーカを退出させる．こうすることで不必要にワーカを長時間待機させない仕組みとなっている．我々の予備実験では最大待機時間を 120 秒に設定しているが，参加者が少ない時間帯などにタスクを実施する場合は調整が必要である．一方，マッチングルームに既にワーカが 1 人待機していた場合はマッチングルームに 2 人のワーカが揃い，マッチング成功となる（③）．

マッチングが成功したワーカペアは自動的にチャットルームに移され，そのチャットルーム内で対話を行う（④）．チャットルームのスクリーンショットを図 2 に示す．対話にはターン制を採用し，1 人のワーカが

4）本稿では Yahoo!クラウドソーシング (https://crowdsourcing.yahoo.co.jp/) を使用した

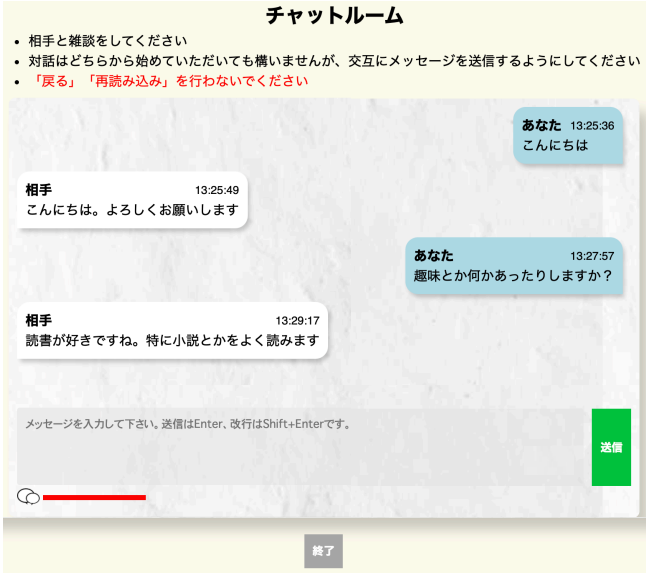

図 2　チャットルームのスクリーンショット

連続して複数回発話することはできないように設定している．対話の進捗は各ワーカの発話数で管理され，各ワーカは各々の発話数を発話入力欄下のプログレスバーで確認することができる．プログレスバーは，対話収集サイト作成時に設定した各ワーカの最低発話数に両者とも達していなければ赤色で表示される．対話の進捗とともに片方が達すれば黄色，両者が達すれば緑色へと次第に変化する．プログレスバーが緑色になる（＝ワーカの最低発話数に両方のワーカが達する）と対話を終了することができる．チャットルーム下部に用意された終了ボタンを押すと各ワーカに固有のチャットルーム ID が自動で発行される（⑤）．また，終了ボタンが押されたタイミングで，そのチャットルーム内で行われた対話が設定ファイルに記した保存場所に自動で保存される．

チャットルーム ID を確認したワーカはそれぞれクラウドソーシングサイトに戻って，発行されたチャットルーム ID を入力し，タスクを完了する（⑥）．このチャットルーム ID はワーカが対話を完了した証拠であるとともに，ワーカと対話の紐付けにも利用する．複数回対話収集に参加しているワーカの場合，そのワーカがこれまでにどの対話に参加したかが分かるため，話者 ID [9] のような役割も果たすことができる．

4　実際の収集例

本節では対話収集基盤を利用した実際の対話収集の例を 2 つ紹介する．

A: はじめまして

B: こんにちは。

A: こんにちは、今回はよろしくお願いします　さっそくですが、あなたのご趣味を教えてくださいますか？

B: 料理や音楽がしゅみです。

A: どちらも素敵なご趣味ですね　私はこういう時期ですので、最近は自炊が多くなりましたが料理はまだ苦手ですw　あなたは料理は何年くらいしてらっしゃるのですか？

B: もう２０年以上やってますねー

A: すごいですねー　失礼ですが、レパートリーはどのくらい・・？

B: う～ん色々作ってますが朝はパスタにしました、パスタ手間かからないので楽ですよ。

A: 私もパスタはときどき作って食べます　ソースはレトルトなんですけど・・w

B: パスタはいいですよ～いろんな種類買いすぎて今家に４０キロぐらいあります…

A: すごいですねw　パスタは保存も効くからいいですよね！　得意な料理のジャンルはございますか？

B: コロナで家にいるので保存食用に箱買してしまいました、得意は煮込み系でしょうか。

A: 煮込みは良いですね、ごはんにもお酒の肴にも　コロナのためにやはりご自宅での炊事は増えましたか？

B: 増えましたね～

A: そうですよね　私も仕方なく自炊するようになりましたし・・・w

B: そうですよね、慣れると簡単ですよ！ 学生さんですか？

A: いいえ、社会人ですw　今までほとんど自炊は母に任せてましたので・・・　それでは今回はありがとうございました

B: 実家いいですね～ありがとうございました！

図 3　趣味雑談対話の例．A が聞き手，B が話し手を指す．

4.1　趣味雑談対話

まず，趣味についての雑談対話を収集した例について紹介する．図 3 にその対話例を示す．各ワーカの最低発話数は 8 に設定した．この趣味雑談対話収集では概ね対話収集基盤をそのまま使用しているが，ワーカそれぞれに追加で役割を割り当てている．具体的には，片方のワーカに趣味の「話し手」，もう片方のワーカに趣味の「聞き手」の役割をランダムに割り当てて対話をしてもらった．図 3 の例では A が聞き手，B が話し手である．「聞き手」のワーカは自身の行動などの発話は行わず，相手のワーカの趣味を聞き出すようにインストラクションされている．この役割はマッチングルームに入ってきた順番を利用して割り当てており，対話収集基盤の雛形をベースに，役割やインストラクションの表示部分のみを追加で実装するだけで実

	対話	知識タイプ	知識
A:	どうぞよろしくお願いします	-	知識なし
B:	こちらこそ、よろしくお願いします	-	-
A:	今日の映画は戦国物語や武将が好きな方には気に入ってもらえると思います	レビュー	戦国物語や武将が好きな方には気に入ってもらえると思います
B:	割と好きな方なので、楽しみです。	-	-
A:	タイトルは「天と地と」と言いますがご覧になったことはありますか？	タイトル	天と地と
B:	聞いたことはありますが、見たことはないです。	-	-
A:	原作が何度か映像化されているみたいですが、今回のは1990年制作のものです	製作年度	1990年6月23日
B:	何度も映画化されたということは、創作意欲を掻き立てられる、おもしろい小説なんですね。	-	-

図 4　映画推薦対話の例

現することができる．このように，本稿で提案する対話収集基盤を用いることでワーカの役割が異なる形式の対話についても容易に収集することができる．この対話収集では，ワーカがタスクに 1 度しか参加できない設定で，50 対話を収集するのに平均約 4 時間程度かかった．

4.2　映画推薦対話

次に映画についての推薦対話 [10] の収集例を紹介する．この対話も趣味雑談対話と同様にワーカの役割が異なり，図 4 の例に示すように映画を勧める推薦者 A と映画を勧められる被推薦者 B に分かれて対話を行う．収集方法は Wizard of Wikipedia [2] と近い設定で，推薦者側は提示された外部知識を参照しながら対話を進めていく．この対話収集の際に使用したサイトも推薦者にのみ外部知識を提示する部分，及び推薦者が外部知識にチェックを入れてから発話を送信する機構のみを追加で実装した．

図 5 に映画推薦対話 200 対話を収集するのにかかった時間と各時点での累計の対話数の関係を示す．対話収集タスクは 2020 年 12 月 29 日の午前 11 時より開始し，同じワーカが 10 回までタスクに参加できる設定とした．また，各ワーカの最低発話数は 10 に設定した．図よりタスク開始後，約 3 時間で 100 対話，約 10 時間で 200 対話の収集が完了している．合計 20 発話以上の比較的長めの対話を非常に高速かつ大量に収集できていると言える．また，マッチングルームでの待機時間の間に相手とマッチングできた割合は 59%となっており，クラウドソーシングを利用したリアルタイムでのマッチングが実用レベルで成功していることが確認できた．

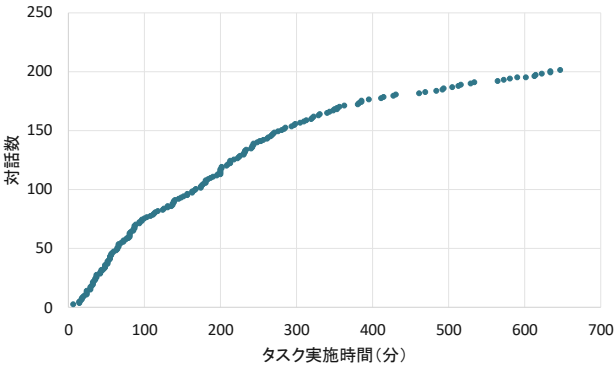

図 5　対話収集にかかった時間

5　おわりに

本稿では日本語の対話収集を容易に行える対話収集基盤を構築した．提案する対話収集基盤を用いることでワーカ間の対話収集をリアルタイムで行えるシステムを少ない負担で構築することができる．また，実際の収集例を通して，本対話収集基盤によって高速かつ大量に対話を収集できることを示し，タスク実施者の用途に合わせてカスタマイズできる拡張性の高さについても紹介した．この研究により対話データ収集の研究が増え，今後の日本語対話研究がさらなる発展を遂げることを期待している．

謝辞

この研究は国立情報学研究所 (NII) CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けて行った．

参考文献

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

[2] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[3] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019.

[4] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, 2020.

[5] Tetsuro Takahashi and Hikaru Yokono. Two persons dialogue corpus made by multiple crowd-workers. In *Proceedings of International Workshop on Spoken Dialogue Systems Technology*, 2017.

[6] 池田和史, 帆足啓一郎. 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. 人工知能学会全国大会論文集 第 32 回全国大会, 2018.

[7] 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, pp. 147–150, 2015.

[8] 東中竜一郎, 稲葉通将, 水上雅博. Python でつくる対話システム. オーム社, 2020.

[9] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, 2016.

[10] 児玉貴志, 田中リベカ, 黒橋禎夫. 外部知識に基づく発話生成に向けた日本語映画推薦対話データセットの構築. 言語処理学会第 27 回年次大会, 2021.