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

クラウドソーシングで構築した疑似対話データの分析

Analysis of Pseudo Dialogue Data Constructed using Crowdsourcing

横野 光 *1　高橋 哲朗 *1
Hikaru Yokono　Tetsuro Takahashi

*1株式会社 富士通研究所
Fujitsu Laboratories Ltd.

In this paper, we discuss aspects of "pseudo" dialog data, which is created by multiple crowd worker assigned to one of two participants in a dialogue. Comparison with the dialogue created in a common way shows that the pseudo dialogue has difference in terms of dialogue behavior.

1. はじめに

対話データは対話システムの構築において，対象としている対話にどのような現象が起こりうるかなどの分析や実際のシステム開発に利用される必要不可欠なものである．対話システムが想定する対話は想定される状況によって様々である．例えば，観光案内におけるガイドと観光客の対話とカスタマーサポートにおけるオペレータとユーザの対話では目的が異なり，それによって行われる対話も異なる．従って，対話システム構築に必要なデータはシステムの状況に応じて用意する必要がある．

対話データの構築は，実際に想定している状況においてなされる対話を収集する以外にも，実験的な状況を設定し，それに基づいて作業者にロールプレイ的に対話を行ってもらい収集するという方法がある．対話データ構築とその他の言語データ構築における相違点として，構築に複数の作業者 (発話者) が必要となり，さらに作業者たちは同じ作業空間に存在しなければいけない，ということがある．この制約のために一般的なアノテーション作業と比べて，作業者のスケジュールを調整するなどといった作業者の管理のためのコストが必要となる．

近年，大量のデータ構築の方法の一つとしてクラウドソーシングが注目されている [馬場 17]．クラウドソーシングでは不特定多数の作業者に対してタスクを割り当てることができる．作業者の管理のためのコストをあまりかけずに安価に大量のデータを構築することができるが，作業者の統制を取ることが困難であり，データの品質の保証や複雑なデータ構築が困難という問題もある．クラウドソーシングに適している作業は作業者が単独で行えるものであるため，対話データ構築のように複数人の作業者が必要となるタスクは通常とは異なる作業が必要となる．

クラウドソーシングを用いた対話データ作成としては，作業者を集めるという部分でクラウドソーシングサービスを利用するというものがある [塚原 15]．これに対して我々はクラウドソーシングに適したタスク形式に対話データ構築を分解するという手法を提案した [Takahashi 17]．これによってクラウドソーシングを用いて安価に対話データを構築することが可能となるが，個々の作業者による発話生成というタスクで対話を構築しているため，通常の手段で構築される対話データと異なる特徴を持つと考えられる．本論文は，我々の手法で構築した対話データを通常の手法で構築した対話データと比較することでその特徴を分析する．

連絡先: 横野光, 株式会社富士通研究所，神奈川県川崎市中原区上小田中 4-1-1, yokono.hikaru@jp.fujitsu.com

2. クラウドソーシングを用いた対話データ構築

対話データの構築では対話の参加者となる作業者を同時刻に作業空間 (e.g. 実際の部屋やチャットスペースなど) に配置する必要があるため，独立したタスクを不特定多数の作業者に割り当てられるクラウドソーシングの利点を生かすことが困難である．この問題に対して，我々は対話データ構築を，これまでの対話の履歴とその対話の参加者に関する情報が与えられた上で，次の発話を作成するというタスクに分解し，クラウドソーシングに適用可能な形式とする手法を提案した [Takahashi 17](図 1)．

この方法では対話の参加者を複数人の作業者で担当して発話を作成する．作業者にはこの対話がどのような状況で行われるものかについての情報と，対話の参加者に関する背景知識などが与えられる．作業者が発話を作成する際には，これまでになされた対話の履歴が表示され，作業者はそれに続く発話を作成する．作業者にはこの形式のタスクが与えられるため，個人の作業においては他の作業者を考慮する必要がないという点で，クラウドソーシングに適した形になっている．

通常の対話データ構築と異なる点は主に以下の 2 点である．

- 対話の参加者を複数人の作業者が担当する
- 実際にリアルタイムで対話を行っていない

これらの要素が生成される対話データにどう影響するのかを把握する必要がある．

2.1 不動産対話コーパス

不動産屋に物件を探しに来た客と，その客に応対するスタッフという設定の対話データを構築した．この対話では，客の目的は自身が持っている背景を満たす物件を見つけたい，というものであり，スタッフの目的はその客の要望にあった物件を見つけるための情報を引き出す，というものである．この対話では，実際の物件検索は行わず，検索に必要な条件を対話によって抽出することがゴールとなる．作業はテキストチャットシステム上で行った．

作業内容は以下の通りである．

- 作業者は客かスタッフのいずれかの立場を指定され，その立場で発話を作成する

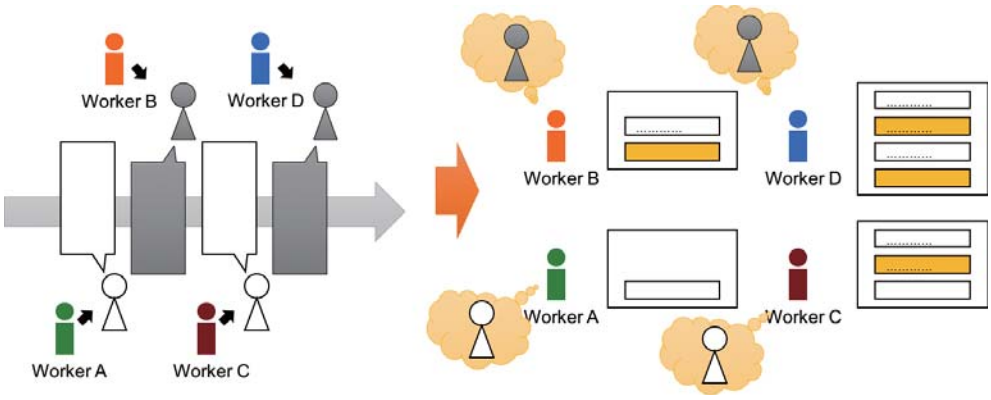

図 1: 個々のタスクに分割した対話データ作成

- 客の役割が与えられた作業者には図 2 のような背景が提示される
- スタッフの役割の作業者には，客から検索条件を引き出すという目的のみが提示される

- 対話はスタッフと客が交互に行う
- スタッフ側の作業者が検索に必要な条件を収集できたと判断したらそこで対話を終了する
- 与えられた対話の履歴が不自然で対話の継続が困難な場合，“不適切” とマークし，その理由を記述して対話を愁傷する
- 対話はスタッフの “どのような物件をお探しですか?” という発話を受けた客側の発話から開始する

2 年ほど付き合った彼氏と同棲することになり、これまでのワンルームから引っ越すことになった女性。これを機に料理に力を入れたいため、コンロが多く使いやすいキッチンがある物件を希望している。

図 2: 背景の例

この設定において，我々の手法と作業者を同時刻に配置し実際に対話を行うという一般的な手法の 2 通りでデータを構築した．どちらの場合も作業者は物件を探した経験があることを条件としており，実際に業務に携わった経験の有無は問うていない．以降，我々の手法で構築した対話データを**疑似対話データ**，一般的な手法で構築した対話データを**リアルタイム対話データ**と呼ぶ．疑似対話に関しては 1 対話 20 発話まで，リアルタイム対話に関しては 1 対話 15 分までで打ち切りとした．

疑似対話データ，リアルタイム対話データの例をそれぞれ図 3，4 に示す．

3. 疑似対話データの分析

構築したデータの対話数を表 1 に示す．疑似対話データに関しては，作業者が与えられた対話の履歴から次の発話を作成することが困難と判断した場合，“不適切” と記入してそこで対話を終了している．

スタッフ：どのような物件をお探しですか？
客：これから子供が出来ることも考え広めの家を探しています。
スタッフ：７畳ぐらいだと子供部屋としては狭いですか？
客：逆に大きすぎると思います。
スタッフ：リビングの間仕切りを変えて広さを変えれますが、いかがでしょうか？
客：でしたら、それでも構いません。
スタッフ：リビングはどのくらいの広さを希望されますか？
客：１２畳が希望です。
スタッフ：場所はどの辺りをご希望ですか？
客：えっと、ごめんなさい。それはリビングの場所のことですか？それとも物件の場所のことですか？
スタッフ：ああ、すみませんでした。物件がある場所の事です。
客：港区周辺を希望しています。
スタッフ：周辺ということですが、電車やバスでどのくらいまでの距離だったら大丈夫ですか。
(以下省略)

図 3: 疑似対話の例

スタッフ：どのような物件をお探しですか？
客：駐車場と駐輪場がある物件を探しています。
スタッフ：一軒家をお探しですか？
客：軒家でもマンションでも構いません。
スタッフ：駐輪場についてはこだわりはございますか？
客：広めだと嬉しいです。
スタッフ：物件の間取りについては何かご希望はございますか？
客：広さにはこだわりません。
スタッフ：物件の間取りについては何かご希望はございますか？
客：広さにはこだわりません。
スタッフ：立地についてご希望はございますか？
客：特にありません。
スタッフ：お家賃についてご希望はございますか？
客：　特に決めてません。
(以下省略)

図 4: リアルタイム対話の例

3.1 “不適切” と判断されたデータの分析

疑似対話データは複数人の作業者で対話を行っていくため，これまでの対話中の発話がどのような意図で行われたのかなどが共有されず，対話を続けることが困難となる状況が起こりうる．不動産対話コーパスの構築ではそのような状況になった時に作業者は “不適切” とマークしその理由を記述している．今回のデータ構築ではどのような時に “不適切” とマークするかについての詳細な指示は行っておらず，直感的に不自然と感じたらマークするという指示に留めている．

実際に不適切と判断され途中で終了した対話は 113 件存在し，それらを人手で分類した．

表 1: 構築した対話データのサイズ

疑似対話データ	765
正常終了した対話	652
“不適切” で終了した対話	113
リアルタイム対話データ	979

前述の通り，業務に従事した経験は不問としたため，不適切と判断した理由の多くは “いきなり予算のことは聞かないため”，“こういう順番で聞いてほしい” といったような，不動産屋ではあまり行われないような対話であるというものであった．また，具体的な判断基準を提示していなかったために，“家族持ちということは、年齢がすごく上に見られたと思い不快に感じる “判断した者の個人的な嗜好に基づいていると思われる理由も存在した．

それ以外の理由として顕著だったものは，“前と同じことを尋ねている” といった発話の繰り返しと “唐突に話題が転換している”，“客の提示した条件に対して関係ない質問をしている” といった不整合な発言というものであり，それぞれ 14 件と 29 件存在した．発話の繰り返しに関しては，通常の対話ではあまり起こりえない現象である．我々の手法ではこれまでの対話の履歴は他の作業者によって作られたものとなるため，実際にどのようなやり取りが行われたかは履歴を読まないと理解できない．この状況は作業者が対話を最初からではなく，作成する発話の直近のみしか読んでいなかったために起こってしまったと考えられる．

不整合に関しては，発話の解釈誤りなどで一般的な対話でも起こりうる．しかし，実際の対話で起こった時はそれを修復するための対話が行われるが，今回の作業では “不適切” を選んで終了，という選択肢を用意したこともあり，修復を試みる前に対話が終了されたと考えられる．

3.2 疑似対話データとリアルタイム対話データの比較

疑似対話データとリアルタイム対話データの統計を表 2 に，1 対話あたりの品詞の平均含有率を表 3 に示す．形態素解析には IPA 辞書を用いた MeCab*1 を利用している．表 2 中の内容語は名詞，動詞，形容詞，副詞を指している．

1 対話あたりの平均発話数は疑似対話データの方がリアルタイム対話データより少なく，一方で，1 発話あたりの文字数は若干ではあるが疑似対話データの方が多くなっている．これは疑似対話データ構築が対話の相手が発話者の発話を待っているという状況ではなく，発話者は自身の発話の生成に時間的な制約を受けていないからだと考えられる．一方で，リアルタイム対話データでは，相手が待っているという状況から時間をかけて自身の発話を考えることが困難となるため，1 発話あたりの文字数が少なく，その代わりに対話が長くなっている．

1 対話あたりの平均文字数，平均単語数はリアルタイム対話データの方が多くなっているが，疑似対話では発話数で対話長を制限したのに対して，リアルタイム対話では時間で制限しており，同じ条件ではないため単純に比較することはできない．しかし，疑似対話の 20 発話という制限に対して，実際の平均発話数は 12.414 であり限界まで到達していない対話が多いことから，疑似対話では長い対話を作成するためにはそのための指示を用意する必要があると考えられる．

発話あたりの文字数，単語数について，疑似対話では客側の発話の方が多く，リアルタイム対話ではスタッフ側が多くなっている．今回の設定のような店のスタッフと客の対話では，主にスタッフが主導を取って客から情報を引き出していくというやり取りが多いと考えられる．そのような場合，スタッフの発話に対して客は返答をするというやり取りが表れ，その結果としてスタッフの方が発話量が多くなると考えられる．疑似対話では，リアルタイムの対話ではなく，また，対話の全てを一人の作業者が担当するわけではないため，内容をどういう順番で提示するかといったような発話の戦略が立てられずに，自身の情報をまずは提示するということが行われたために客側の発話量が多くなったと考えられる．

どのような役割の発話が行われたかについて，それぞれの発話に対してその発話意図を推定する代わりに名詞を含む発話がどの程度存在するかを求めた *2．これは，名詞が含まれている発話は何かしらの情報伝達の役割を持っており，逆に名詞を含んでいない発話は相づちや単なる返事などの役割を持っていることが多いと考えたからである．

わずかではあるが，リアルタイム対話の方が名詞を含む発話の割合が小さく，また，スタッフ側に比べて客側の発話の方

表 2: 対話データの比較

	疑似対話データ	リアルタイム対話データ
1 対話あたりの平均発話数	12.414	28.574
1 対話あたりの内容語の異なり数	51.130	84.002
1 対話あたりの平均文字数	278.169	567.689
1 対話あたりの平均単語数	170.917	344.002
1 発話あたりの平均文字数	22.407	19.867
スタッフ側の発話	20.222	21.725
客側の発話	24.267	18.120
1 発話あたりの平均単語数	13.768	12.039
スタッフ側の発話	12.556	13.269
客側の発話	14.799	10.882
名詞を含む発話の割合	0.984	0.953
スタッフ側の発話	1.000	0.993
客側の発話	0.971	0.916
代名詞を含む発話の割合	0.101	0.113
スタッフ側の発話	0.169	0.179
客側の発話	0.043	0.050
直前の発話と重複する内容語を持つ発話の割合	0.415	0.374
スタッフ側の発話	0.241	0.234
客側の発話	0.588	0.515

表 3: 1 対話あたりの品詞の平均含有率

	疑似対話データ	リアルタイム対話データ
名詞	0.307	0.234
動詞	0.098	0.105
形容詞	0.038	0.030
副詞	0.009	0.008
連体詞	0.008	0.005
助詞	0.294	0.276
助動詞	0.116	0.129
接頭詞	0.023	0.024
接続詞	0.002	0.002
感動詞	0.003	0.008
記号	0.102	0.113

*1 http://taku910.github.io/mecab/

*2 具体的には IPA 辞書の品詞体系で非自立ではない名詞のみを対象としている．

が割合が小さいことから，リアルタイム対話ではスタッフ側の提案に対して，客側が応答している，といったようなやり取りが含まれていると考えられる．この点において，疑似対話データでは実際の対話におけるやり取りの表れ方に違いが出てきている．

直前の発話と重複する内容語を持つ発話の割合は疑似対話データの方が大きい．これも発話にかける時間の制約の有無が影響していると考えられる．つまり，疑似対話では発話にかける時間を考慮する必要がないため，多少冗長な表現になっていると考えられる．客側の発話に重複が多いのは，スタッフからの質問を受けてそれに答えているためと考えられる (e.g. スタッフ: 「家賃はどうですか?」，客: 「家賃は安い方が良いです」)．また，疑似対話では対話の参加者を複数人で担当しているため，対話の参加者の背景は共有されているが，対話が行われている時の対話の相手と共有される知識が作業者間で共有されることがないというのも影響していると考えられる．このことは照応や省略という現象として発話文に表れると考えられるが，表 2 に示すように代名詞の出現の傾向からはわずかにリアルタイム対話の方が割合が大きいものの，それほど顕著な差は表れなかった．

表 3 に示す対話中の品詞の分布に関しては，名詞の割合とそれに伴って助詞の割合が疑似対話データで多くなっているが，それ以外の品詞に関しては大きな差は見られなかった．

内容のやり取りという点では疑似対話データとリアルタイム対話データでは大きな差は見られないが，実際の対話で発生するやり取りに関してはリアルタイム対話データと違いがあると考えられる．

4. おわりに

本稿では，クラウドソーシングに適用するために対話データの構築を対話履歴に対する発話文の作成というタスクに分割するという手法で構築した対話データの特徴について，不動産屋でのスタッフと客との対話データを基に分析を行った．

我々の手法で構築した疑似対話データはリアルタイムデータと比べてそれほど大きな違いは見られなかったが，我々の手法は作業者に時間的な制約を与えていなかったために，発話長などのように時間に関わる要素で差が見られた．また，作業者間では対話の参加者間で構築される共有知識が共有されないため，前の話題に戻るといったようなことが起こりにくい．このことから，我々の手法による疑似対話データは対話のやり取りを捉えるためには作業者への指示などにそれらを意識して用いさせるといったこと行う必要があるが，伝達される内容に関しては一般的な対話データとして用いることが可能であると考えられる．

参考文献

[Takahashi 17] Takahashi, T. and Yokono, H.: Two persons dialogue corpus made by multiple crowd-workers, in *Proceedings of International Workshop on Spoken Dialogue Systems Technology (IWSDS2017)* (2017)

[塚原 15] 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)

[馬場 17] 馬場雪乃：ヒューマンコンピュテーションとクラウドソーシング, 言語処理学会第 23 回年次大会チュートリアル (2017)