修士論文

マイクロブログからの対話コーパスの自動構築

関田　崇宏

主指導教員 白井　清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和元年三月

Abstract

In recent years, many studies of dialog systems that can chat with users are widely investigated. A dialog corpus, which is a collection of dialogs between humans, is necessary to develop such dialog systems. However, it is rather difficult to construct a large scale dialog corpus, since it requires much cost to record and transcribe conversation between human. On the other hand, several researchers attempted to automatically construct a dialog corpus by retrieving a large amount of sequence of tweets and replies, which is regarded as a pseudo dialog, from Twitter. However, such sequence of tweets and replies may not be a real dialog. One of the problems of the previous studies of automatic construction of a dialog corpus from Twitter was that they did not consider whether the retrieved pseudo dialogs were appropriate to be included in a dialog corpus. The goal of this thesis is to automatically construct a high quality and large scale dialog corpus by retrieving dialogs, i.e. sequence of tweets and replies, from a microblog (Twitter) and removing inappropriate ones from them.

The proposed method consists of three steps: collecting sequence of tweets and replies from Twitter as dialogs, removing inappropriate dialogs, and constructing a dialog corpus from the remain.

To collect dialogs from Twitter, first we search tweets by a keyword. If the tweet is a reply of another tweet, we retrieve both of them. We repeat this procedure unless the tweet is not a reply of another tweet. In this way, we collect sequence of tweets and replies as a dialog. Finally, we keep the dialog if its length (the number of the tweets) is greater than or equal to 3. The Twitter API is used to search and collect tweets. In order to collect natural dialogs, we use a list of 1,686 words whose word familiarity are high, such as "everyone", "reunion", and "lover", as search keywords.

In the next step, inappropriate dialogs are detected and removed from the collected dialogs. First, we analyze 100 dialogs and investigate what kinds of dialogs are inappropriate, how they are categorized, and how to detect them. As a result, we define four rules to remove inappropriate dialogs. The first one, $\mathbf{R_{short}}$, is a rule that removes dialogs containing a short tweet (utterance). Dialogs including an extremely short utterance are often not real dialogs. We remove dialogs if they contain a tweet that consists of only one Hiragana character (except for interjection), symbols such as punctuation, or emoji. The second one, $\mathbf{R_{line}}$, is a rule that removes dialogs including a tweet with multiple lines. Utterance including multiple lines does not usually appear in a real dialog, although it may appear in sequence of tweets and replies when users make their own stories on Twitter. We remove dialogs if they contain a tweet that fulfills the following conditions: (1) it includes multiple pairs of parentheses (lines are usually indicated by parentheses),

(2) the character length in parentheses is more than or equal to 6, (3) a word after the parentheses is not a case marker. The conditions (2) and (3) are set since parentheses are often used not to mark up a line but to emphasize a short noun. The third one, $\mathbf{R_{image}}$, is a rule that removes dialogs including images and URLs. If a tweet contains an image or URL of another web page, people cannot understand a dialog if they do not see and know the contents of the image or the linked web page. We distinguish dialogs where people can not understand them without image or URL and ones where people can understand even without image or URL. We aims at removing only the former. More specifically, we remove a dialog if it contains a tweet including an image or URL and there exists a demonstrative such as "this" or "that" around the image. The presence or absence of a demonstrative is checked to determine whether the tweet mentions the image. The fourth one, $\mathbf{R_{invite}}$, is a rule that removes dialogs if they start with a tweet that widely calls something to other Twitter users. A pseudo dialog is not a real dialog if it includes a call for many people such as Ogiri, which is a game where one user provides a question and other users reply funny answers. A list of users who run Ogiri is manually created in advance. A dialog is removed if the user of the tweet of the beginning of the dialog is included in the list.

Several experiments were conducted to evaluate our proposed method. Dialogs were collected from Twitter between June 8, 2019 and December 25, 2019. The number of collected dialogs was 92,207; the average length was 9.50. Thus we were able to collect a large number of relatively long dialogs. Next, we randomly selected 100 dialogs, and two subjects independently determined whether those dialogs were appropriate or not. The κ coefficient of two subjects was 0.60. In this way, two test datasets were prepared; each is 100 dialogs annotated with the judgment with one subject. Then, the performance of the detection of inappropriate dialog by the proposed method was measured on these two datasets. The precision was 0.75 and 0.75, the recall was 0.32 and 0.43, and the F-measure was 0.45 and 0.55. The major cause of low recall was that many inappropriate dialogs including images were failed to be detected.

Next, we evaluated the individual rules. Note that the rule $\mathbf{R_{invite}}$ was not evaluated because it used the manually created list of Ogiri users. Among dialogs that were judged as inappropriate by each rule, 50 dialogs were chosen as the test data. Two subjects independently judged whether they were inappropriate or not. The κ coefficient of the two subjects were 0.37 for $\mathbf{R_{short}}$, 0.47 for $\mathbf{R_{line}}$, and 0.77 for $\mathbf{R_{image}}$. The precision of $\mathbf{R_{short}}$, $\mathbf{R_{line}}$, and $\mathbf{R_{image}}$were 0.96/0.94, 0.74/0.76, and 0.78/0.78, respectively. It was found that the precision of detection of inappropriate dialogs of the proposed rules was relatively good.

In the future, it is necessary to refine the rules to detect inappropriate dialogs to

improve the precision and recall. It is also necessary to investigate another types of inappropriate pseudo dialogs and design methods to automatically remove them.

概要

近年、機械と人間が対話を行う対話システムに関する研究が盛んに行われている。対話システムの開発には、人間同士の対話を収録した対話コーパスが必要である。しかし、対話を録音したり書き起こしたりするのは多大なコストを要するため、大規模な対話コーパスを構築することは難しい。これに対し、Twitter からツイートとそれに対するリプライの組を疑似的な対話とみなし、これを大量に獲得することで対話コーパスを構築する試みが行われている。しかし、自動収集したツイートとリプライは対話として不自然なものも含まれるが、先行研究における対話コーパス構築では抽出した対話が適切であるかは考慮されていないという問題点がある。そこで、本研究ではマイクロブログ (Twitter) から疑似対話を抽出し、この中から対話として不適切なものを除去することで、良質かつ大規模な対話コーパスを構築することを目的とする。

提案手法は、Twitter からツイートとリプライの連鎖を対話として収集する、その中から不適切な対話を除去する、残された対話で対話コーパスを構築する、という 3 段階の手続きからなる。

対話の収集は、まずキーワードでツイートを検索し、それが別のツイートのリプライであるとき、元のツイートを辿っていき、長さ 3 以上の一連のツイートを対話として保存する。ツイートの検索や収集は Twitter API を用いる。自然な対話を収集するために、検索キーワードには「全員」「再会」「恋人」などのような単語親密度の高い 1,686 の単語を使用する。

次に、収集した対話の中から不適切な対話を除外する。まず、不適切な対話にはどのようなものがあるのかを調査するために、収集した対話 100 件を分析し、不適切な対話の分類や検出方法を検討した。その結果、不適切な対話を除去する 4 つのルールを考案した。一つ目は短いツイート (発話) を含む対話を除去するルール ($\mathbf{R_{short}}$) である。極端に短い発話を含む対話は対話として成立しない場合が多い。間投詞以外の一文字のひらがなのみ句読点などの記号のみ、絵文字のみのツイートが含まれている対話を除去する。二つ目は複数のセリフがあるツイートを含む対話を除去するルール ($\mathbf{R_{line}}$) である。1 つのツイートに複数のセリフが含まれている場合は、Twitter 上で物語を創作しているなど、対話として不適切なときが多い。括弧の組が 2 つ以上ある、括弧の中が 6 文字以上である、括弧の次の単語が助詞ではない、という 3 つの条件を満たすツイートを含む対話を除去する。三つ目は画像・URL を含む対話を除去するルール ($\mathbf{R_{image}}$) である。画像や他のウェブページの URL を含むツイートが対話の中に存在するとき、その画像やリンク先ウェブページの内容がわからなければ対話を理解できないことがある。画像を参照しないと内容を理解できない対話と、画像なしでも内容を理解できる対話を区別し、前者のみを除去する。具体的には、ツイートが画像や URL を含むこと、画像の周辺に「これ」「それ」などの指示語があること、などを除外の条件とする。指示詞の有無は、ツイートが画像に言及しているか否かを判断するためにチェックする。四つ目は不特定多数のユーザーへの呼びかけを含む対話を除去するルール

($\mathbf{R_{invite}}$) である。大喜利を発信するツイートとそれに対するリプライなど、対話の起点となるツイートが不特定多数への呼びかけであるときは対話として成立しない場合が多い。大喜利を運営しているユーザーのリストをあらかじめ人手で作成し、対話の起点となるツイートのユーザーがそのリストに含まれていれば、その対話を除去する。

提案手法の評価実験を行った。Twitter からの対話の収集は、2019 年 6 月 8 日から 2019 年 12 月 25 日にかけて実施した。収集した対話数は 92,207、平均対話長は 9.50 であり、比較的長い対話を大量に集めることができた。次に、ランダムに 100 件の対話を選択し、それらの対話が適切か不適切かを 2 名の作業者が独立に判定した。2 者の判定の κ 係数は 0.60 であった。これらを判定者毎に分けた 2 つの評価データとし、提案手法による不適切な対話検出を評価したところ、精度は 0.75 と 0.75、再現率は 0.32 と 0.43、F 値は 0.45 と 0.55 であった。再現率が低かった主な要因は画像を含む不適切な対話を検出できていなかったためであった。

次に、ルールを個別に評価する。ただし、$\mathbf{R_{invite}}$ は人手で作成した大喜利ユーザーのリストを用いているため、ここでは評価しない。各ルールによって不適切であると判定された対話 50 件を選択し、それらの対話が不適切であるかを 2 名の作業者が独立に判定した。2 者の判定の κ 係数は、$\mathbf{R_{short}}$ が 0.37、$\mathbf{R_{line}}$ が 0.47、$\mathbf{R_{image}}$ が 0.77 であった。不適切な対話検出の精度は、$\mathbf{R_{short}}$ は 0.96 と 0.94、$\mathbf{R_{line}}$ は 0.74 と 0.76、$\mathbf{R_{image}}$ は 0.78 と 0.78 となり、比較的良好な結果が得られた。

今後の課題として、個々のルールを改善して不適切な対話検出の精度、再現率を向上させることが挙げられる。また、今回想定した 4 つのタイプ以外にも不適切な対話があるかを調査し、これを自動的に検出する手法を検討する必要がある。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年、iPhone に搭載されている音声アシストシステムである Siri や、日本マイクロソフトが開発した対話ボットの一つである女子高生 AI りんなど、機械と人間が対話を行う対話システムが広く利用されるようになっている。対話システムは、大きく分けて以下の 2 つの種類がある。

一つ目は、あるタスクを達成することを目的に人と対話するタスク指向型対話システムである。具体例として、目的地までのルートや時間を案内するカーナビなどが挙げられる。タスク指向型対話システムでは、対話の内容がタスクによってある程度限定されており、システムによるユーザへの応答も定型的な文が多い。

二つ目は、普段日常で行われているような対話を行う非タスク指向型対話システム、いわゆる雑談システムである。具体例は、先ほど挙げた女子高生 AI りんなどが挙げられる。非タスク指向型対話システムでは、人間同士の日常対話のように、対話の内容は特に決まっているわけではなく、システムはユーザに対して様々な応答文を生成することが要求される。

対話システムは、タスク指向、非タスク指向ともに様々な場面で用いられている。近年では、雑談対話システムの研究が盛んに行われている。雑談対話システムは、例えば人間と対話を行う機会の少ない高齢者が雑談対話システムと対話を行うことで認知症を予防することや、雑談を行うだけで知識を獲得することができる知識伝達型対話システムなど、様々な応用例がある。

人間と対話を行う対話システムの開発には、人間同士の対話を収録した対話コーパスが必要である。しかし、対話を録音したり書き起こしたりするのは多大なコストを要するため、大規模な対話コーパスを構築することは難しい。これに対し、Twitter からツイートとそれに対するリプライの組を疑似的な対話とみなし、これを大量に獲得することで対話コーパスを構築する試みが行われている。しかし、自動収集したツイートとリプライは対話として不自然なものも含まれるが、先行研究における対話コーパス構築では抽出した対話が適切であるかは考慮されていないという問題点がある。

1.2 目的

本研究ではマイクロブログ (Twitter) から適切な疑似対話を抽出することで良質かつ大規模な対話コーパスを構築することを目的とする。Twitter からツイートとリプライの組を連結したものを対話として取得し、その中から対話として不適切なものを除外することで質の高い対話コーパスを自動構築する。また、対話を大量に除去して対話コーパスの規模が過度に小さくなるのは望ましくないため、適切な対話を出来るだけ多く残せるような手法を探求する。

1.3 本論文の構成

本論文の構成は以下の通りである。2 章では本研究の関連研究について述べ、これらの研究と本研究の違いについて論じる。3 章では Twitter から対話コーパスを構築する手法について述べる。特に、Twitter から収集された疑似対話の中から不適切な対話を除去するルールを詳述する。4 章では提案手法の評価実験について述べる。最後に 5 章では本論文のまとめおよび今後の課題を述べる。

第2章　関連研究

本章は、本研究の関連研究について述べる。これまで、自由対話システムの研究のためにTwitterから対話コーパスを構築する研究が行われている。2.1節では、Twitterもしくはチャットログから対話コーパスを構築することを目的とした研究を紹介する。2.2節では、対話コーパスの構築そのものが目的ではないが、Twitterから対話コーパスを構築し、これを対話システムの開発に利用した研究について述べる。2.3節では、対話システムの関連研究を紹介する。最後に、2.4節では、本研究と先行研究の違いを論じ、本研究の特色を明らかにする。

2.1　対話コーパスの構築に関する研究

Loweらは多様なトピックについて1対1のマルチターン会話が可能な対話エージェントを構築する問題について考察した[9]。彼らはUbuntuチャットログから抽出された約100万人の2人会話を収集し、Ubuntu dialogコーパスを構築した。会話はテキスト入力で行われ、会話の長さは平均8ターン、最低3ターンであった。チャットルームで行われた多人数会話から二項対話を抽出する方法を考案しデータセット(対話コーパス)を構築した。そのデータセットの価値を調査するために、RNNやLSTMを用いた対話システムを学習した。対話コーパスの対話から(応答前の対話、出力発話、出力発話が対話内の実際の発話であるか)といった3つ組を抽出し、それらを用いてRNNやLSTMを学習した。k個の最も可能性の高い回答を選択させ、そのk個の候補の中に真の回答が含まれているか(Recall@k)を評価指標とした。LSTMのRecall@5は92.6%であり、これが実験で得られた一番良い結果であった。また、応答選択アルゴリズムの評価のためのテストセットも作成した。

稲葉らは、非タスク指向型対話システムでの応答生成に用いることを前提に、Twitterから発話として使用可能な文を自動獲得する手法を提案した[8]。入力された「読書」「テニス」などの話題語でTwitterを検索し、話題語を含む、URLを含まない、単語数が6以上30未満である、ユーザー名を含まないという条件をすべて満たす文を対話システムの応答に使う文として抽出した。さらに、抽出した文に対し、以下の7つのルールを用いてフィルタリングを行った。一つ目は話題語と名詞が連続している文を除外するルールである。話題語が「アメリカ」である場合における「アメリカザリガニ」など、話題語ではない複合名詞を含む文が

抽出されるの妨げるために設けられた。二つ目は人名、代名詞が含まれている文を除外するルールである。人名、代名詞が含まれている場合はその発話だけで意味が理解できないことが多いため設けられた。三つ目は先頭の単語の品詞が助詞、助動詞、接続詞の文を除外するルールである。四つ目は末尾の単語の品詞が助詞-格助詞、助詞-係助詞、助詞-接続助詞、助詞-並列助詞、名詞 (名詞-形容動詞語幹は除く) の文を除外するルールである。この研究では、記号をセパレータとしてツイートを文に分割しているが、三つ目と四つ目のルールは、文分割が不適切であるときに生じる不自然な文を削除する働きをする。五つ目は文末以外に助詞-終助詞が含まれている文を除外するルールである。このような文には句読点の打ち間違えや誤字が多く見られたため、これを除外するルールが設けられた。六つ目は時間を特定する語や数値が含まれている文を除外するルールである。「今日」や「明日」などの時間を特定する語が含まれている文は、限定された時間でしか使用できない文であることが多く、数値も日付などを指定する際に多く使われていたため、このルールを設けた。七つ目は不十分な比較が含まれている文を除外するルールである。「〜より〜のほうが」という表現は比較の際に用いられるが、「より」と「ほうが」の一方しか含まれない文は不完全な文の可能性が高いため、これを除外する。これらのフィルタリングにより対話として適切な発話を獲得した。

次に、得られた発話に対して、それが対話での応答文としてどれだけ適切かを評価するスコアを算出し、これが低い発話を除外する。ここで、応答文として適切な発話を「正解発話」、不適切な発話を「不正解発話」と定義する。正解発話と不正解発話では出現する単語が異なると仮定し、正解発話中に出現しやすい単語には高い点数を、不正解発話に出現しやすい単語には低い点数を付与した。例えば、「降る」は天候を述べる際に使用されることが多いため、この単語を含んだ発話は使用できる時間・空間が限定された発話となる。ゆえに、「降る」は不正解発話に出現しやすいと考えられる。単語の点数は、その単語が正解発話の総単語数に占める割合と不正解発話の総単語数に占める割合の比を用いた。そして文中に出現する単語の点数を掛け合わせることで文の点数を求め、このスコアが低い発話を除外した。最後に、発話文として使用可能な形にするため、文の語尾を整形した。

適切な発話の選別について、上記の提案手法と SVM を用いた発話の点数付けによる手法を比較した。SVM を用いた手法では、発話が正解発話である事後確率を推定し、これをスコアとした。様々な話題語に関する合計 500 発話を対象とした評価では、提案手法の正解率が SVM による手法を 10%以上上回った。

2.2 Twitter を利用した対話システム関する研究

Ritter らは、発話の対話行為を推定する 3 種類の教師なし機械学習手法を提案した [11]。対話行為推定モデルの学習データとして、Twitter から 3 回以上連鎖するツイートとリプライを対話として収集し、対話コーパスを作成した。Twitter 上では様々なトピックについて対話が行われており、対話全体を「食べ物」「音楽」

などのトピックに分割すると、話題が遷移していくことが予想される。対話行為の推定結果を視覚化し解釈するために、対話行為遷移図等を検討した。対話のみの視覚化モデルよりも対話とトピックの両方を考慮した視覚化モデルの方が解釈しやすいことがわかった。

Ritter らは、発話に自動的に応答するタスクに対し、統計的機械翻訳に基づくいくつかのデータ駆動型のアプローチを示した [12]。130 万の対話データから、対話の最初の 2 発話を抽出し、最初の発話を次の発話に翻訳するモデルを学習することで、発話に対して応答文を生成するシステムを構築した。

東中らは、Twitter から自動収集した対話コーパスから、入力された発話に対して応答を返す対話モデルを構築する手法を提案した [5]。彼らは対話コーパスとしてツイートとリプライの 2 つの発話の組を収集し、次に内容が似ている発話の組を連結することで、ターン数が 2 回を超える長い対話をモデル化した。ターン数が 2 回の対話データとそれを連結して作成したターン数が 3 回以上の対話データから学習したモデルを用いて、別途収集したターン数 3 回以上の対話データをいかに説明できるかを調査することで、提案手法の対話モデルを評価した。評価尺度として対数尤度とケンドールの τ を用い、ターン数が 2 回の対話のみから構築した対話モデルの有効性を示した。

別所らは、Twitter 大規模コーパスとリアルタイムクラウドソーシングの枠組みを利用した対話システムを構築した [3]。Twitter から 120 万の発話対を対話コーパスとして収集し、ユーザの発話が入力されたとき、その発話と似ているツイートを検索し、それに対するリプライを応答として生成した。また、適切な応答文を Twitter コーパスから見つけることができなかったときは、応答文の作成をリアルタイムでクラウドソーシングに依頼するシステムを構築した。100 万を超える発話に対して平均 2.54 秒の応答時間で応答を生成し、またクラウドソーシングの枠組みの統合がシステム応答の自然さよりシステム全体の面白さの向上に寄与することを示した。

福田らは、バックチャネル (相槌) と呼ばれる、「うん」や「はい」などの聞き手側が行う会話中の短いリアクションを適切に生成する手法を提案した [7]。Twitter から他のツイートに対するリプライを収集し、ハッシュタグや URL、「おはよう」などの定型的な挨拶を含むものを除いて、約 1500 万件のリプライからなる実験データを作成した。人手によってリプライ中のバックチャネルのタイミングが「適切」「どちらともいえない」「不適切」であるかを判定し、LSTM を用いた自動判定結果と比較したところ、LSTM によるモデルでは不適切なバックチャネルが生成されにくいことを示した。

2.3 対話システムに関する研究

Banchs と Li は、ベクトル空間モデルフレームワークに基づいたチャット指向の対話システムである IRIS(Informal Response Interactive System) を提案した [2]。

IRIS は対話の大規模なデータベースを持ち、これを用いて特定のユーザー入力に対する応答を生成する。現在のユーザー入力とデータベース内の全ての既存の発話と、現在の対話履歴のベクトル表現とデータベース内の対話のベクトル表現の 2 種類をそれぞれ比較し、ユーザの入力に対する応答をデータベースから取得する二重検索戦略を用いた。対話の最初にユーザーの名前を聞き出し、その名前を保存する。ユーザの名前が既に保存されていた場合は保存された対話履歴が読み込まれ、そうでない場合は対話データベースから 1 つの対話ベクトルをランダムに選択する。その後、IRIS はユーザーに何をしたいのかを尋ねる。そして、対話データベースや用語を保存する Vocabulary Learning リポジトリに保存されていない用語 (Out of Vocabulary: OOV) が対話に現れた場合は、その意味をユーザーまたは外部の情報源から収集し、意味を理解したことをユーザーに伝える。収集されなかった場合は二重検索構造を用いてベクトルの類似性を計算し、対話データベースから類似性の高い発話の中の 1 つをランダムに選択し、ユーザーに返す。IRIS では、Movie-DiC と呼ばれるインターネットムービースクリプトデータコレクションから抽出された対話コーパスが用いられており、実装にはコメディ、アクション、家族のジャンルに属する 153 本の映画の台本から構成される対話集を使用した。新しい用語を学習することで、それらを既に保持している知識に意味的に関連付け、ユーザーへの応答発話の質を向上させた。

福田らは、対話履歴要約が対話システムの応答生成に有効であると仮定し、過去に話した内容と関連のある発言を行った際に、対話履歴の発話の要約を用いて応答するシステムを提案した [6]。対話データとして、人間同士の 1 対 1 による 30 分のチャット対話を合計 1025 発話収集した。被験者実験では、主要な話題となった単語に着目して応答を生成することで、被験者の話を聞いているかという項目において、ベースシステムと比較して倍以上の高い評価が得られた。

大竹らは、ウェブから収集した情報と複数の言語資源の知識を用いた高齢者向け対話システムを提案した [10]。入力された発話に対して前処理を行い、入力文を解析した後、応答文を生成する。応答文は「関連語を用いた発話」「5W1H 法」「意見感想の促し」「ポジティブな評価」「共感」「相槌」の 6 種類であり、それぞれを生成するモジュールを実装した。被験者実験では、「話を聞いてもらえたと感じたか」という項目に関して、その他の項目と比較して平均評価値がやや高いことが分かった。

江頭らは、通常の発話生成や質問応答などの異なる機能を個別の応答生成モジュールとして扱い、発話に対してそれらを適切に選択することで雑談対話を実現する手法を提案した [4]。発話を生成するための主な情報源としてウェブ上の最新のニュースを用いた。強化学習によって、対話を重ねることによってユーザの満足のいく対話を行うことができる対話戦略を学習することができた。

山口らは、対話に応じて適切な相槌をうつために、相槌の形態と発話の統語構造や韻律的特徴の関係について分析し、これらの発話の特徴を手がかりに相槌を予測したり生成したりする手法を提案した [13]。対話コーパスとして、日常の悩み

や困りごとに関する対話を収録したコーパスを用いた。機械学習によって発話の特徴から相槌を適切に予測するモデルを学習できることが示された。

2.4 本研究の特色

Twitter から取得したツイートとリプライの組の中には対話として不自然なものも含まれるが、先行研究では Twitter から取得した対話が適切であるかどうかはほとんど考慮されていなかった。これに対し本研究では、抽出したツイートの組が対話として適切であるかを判定し、不適切なものを除去することで対話コーパスの品質を向上させることを狙う。高品質かつ大規模な対話コーパスは、それを元に構築した対話システムの品質向上に繋がるため、実用上の意義も大きい。

第3章　提案手法

本章では、Twitter から対話コーパスを構築する提案手法の詳細を述べる。3.1 節では、不適切な対話にはどのようなものがあるかを調査した予備調査について述べる。3.2 節では、提案手法の概要について述べる。3.3 節では、Twitter から対話を収集する方法を述べる。3.4 節では、3.1 節の予備調査の結果を踏まえて設計した不適切な対話を除去するルールについて述べる。3.5 節では、3.4 節のルールによって不適切な対話を除去して構築した対話コーパスについて述べる。

3.1　予備調査

まず、Twitter から自動収集された対話のうち、対話として不適切なものにはどのようなものがあるのかを調査した。3.3 節に後述する手法で Twitter から疑似対話を収集し、その中からランダムに 100 件の疑似対話を選んだ。これらに対し、その疑似対話が対話として適切かどうか、不適切な場合はその要因を考察した。予備調査の結果を表 3.1 に示す。この表は、適切ならびに不適切と判定した対話数を示し、また不適切と判定した対話についてはその要因毎に対話数を示している。

表 3.1: 予備調査の結果

不適切さの要因	対話数
(1)　画像や URL を含む対話	38
(2)　セリフを含む対話	2
(3)　不特定多数への呼びかけを含む対話	3
(4)　短い発話を含む対話	1
(5)　絵文字を含む対話	2
適切な対話	54

適切な対話の数が 5 割程度しかないことから、Twitter から収集した対話の中には不適切なものが多いことが分かった。本研究は Twitter 上の疑似対話の中から不適切なものを除去することを目的としているが、これが質の高いコーパスを構築するために必要な処理であることが確認された。不適切な対話の中では、画像や URL を含む対話が一番多かった。表 3.1 に示した対話の不適切さの要因の詳細は 3.4 節で述べる。

以上の予備調査の結果を踏まえ、Twitter から収集した対話から不適切なものを除外する手法を考案した。具体的には、不適切さの要因毎に、不適切な対話を除去するルールを設計した。以降の節では提案手法の詳細について説明する。

3.2 提案手法の概要

提案手法の概要を図 3.1 に示す。まず始めに Twitter から対話を収集する。次に、不適切な対話を 4 つのルールを用いて除去する。そして、除去されなかった対話から対話コーパスを構築する。以降の節では、それぞれのステップの詳細を順に説明する。

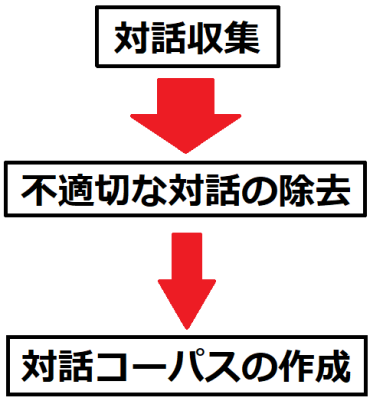

図 3.1: 提案手法の概要

3.3 対話収集

Twitter では、あるユーザが投稿したテキスト (ツイート) に対して返答することができる。このようなユーザの返答をリプライと呼ぶ。ツイートとそれに対するリプライは、ユーザー間の疑似的な対話とみなすことができる。また、二者もしくはそれ以上のユーザが、互いのツイートに対してリプライを繰り返すとき、これらの一連のツイートは長い疑似的な対話とみなすことができる。本研究では、このようなリプライの連鎖を対話として収集する。

Twitter から対話を収集する手続きを以下に示す。

1. キーワードでツイートを検索する。
2. 得られたツイート t_1 が別のツイート t_2 のリプライであり、かつ t_1 のユーザ ID と t_2 のユーザ ID が異なるとき、t_2 を取得する。

3. 2. の操作をツイート t_i が別のツイートのリプライでなくなるまで繰り返す。

4. 得られたツイートを逆順に ($t_n t_{n-1} \cdots t_1$ の順に) 並べ、対話として保存する。n が 2 のとき、対話は一つのツイートとそれに対するリプライのみから構成され、対話としては短い。本研究では、なるべく長い対話を獲得するため、長さ 3 以上 ($n \geq 3$) の対話のみを保存する。

5. 1.〜4. の操作を繰り返す。

Twitter では、他者ではなく自分のツイートに対してもリプライすることができる。このとき、ツイートとリプライの組はともに同一人物による投稿なので、対話とはみなせない。上記の手順 2. において、t_1 と t_2 のユーザ ID が同じでないことを確認しているのは、同一ユーザによるリプライを誤って対話として取得しないためである。

ツイートの検索や収集は Twitter API を用いる。Twitter からキーワード「全員」を用いて対話を収集する例を図 3.2 に示す。Twitter API を用いて「全員」を含むツイートを検索し、③のツイートを得る。これは、②のツイートに対するリプライであるため、②のツイートも取得する。同様に①のツイートも取得する。①のツイートは他のツイートに対するリプライではないため、収集した①-②-③のツイートを 1 つの対話として保存する。

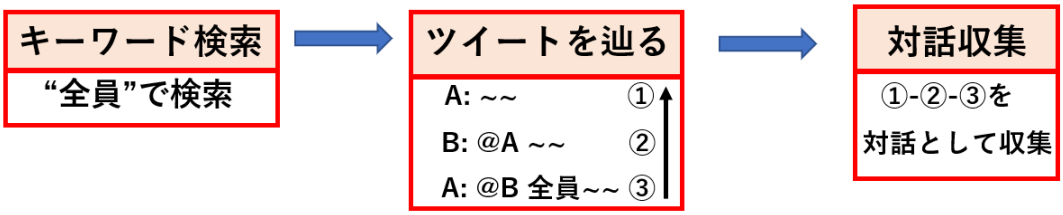

図 3.2: 対話の収集例

上記の手順 1. で用いる検索キーワードとして、単語親密度 [1] の高い単語を使用する。単語親密度とは、単語を「知っている」「書く」「読む」「話す」「聞く」の 5 つの観点について、その傾向の強さを 5 段階で評価したものである。「知っている」という観点については多くの人がその単語を知っている度合を、「書く」「読む」「話す」「聞く」についてはそれぞれの行為における単語の使用頻度を評価している。本研究では文献 [1] で構築したデータベースに登録されている約 10 万の単語の中から「知っている」についての単語親密度が 2.28 以上である名詞を使用する。単語親密度の高い単語は、多くの人にとってなじみのある単語であるために使用頻度が高く、多くの対話を収集するのに適しており、また親密度が高い単語

を含む対話は自然な対話になりやすいと考えたためである。本研究では、「全員」「再会」「恋人」などの 1,686 個のキーワードを使用する。その一部を表 3.2 に示す。

表 3.2: 対話の収集の際に使用したキーワード (抜粋)

全員、再会、恋人、入社、ストロー、トンネル、常備薬、三人がかり、授業、リップクリーム、混ぜ御飯、ネクタイ、原因、今月、ピーマン、シェア、餓死、発砲、投球、こけし、水道、フラワー、絶対音感、韓国、自宅、決め手、二十世紀、灯油、頂上、起床、飲み物

Twitter API の仕様では、15 分間で 180 ツイート以上の検索ができない。そのため、間隔を空けてツイートを検索し、対話を継続的に収集するクローラーを実装した。また、同じ検索キーワードを用いても、時間がたてば検索されるツイートの集合は異なる。実装したクローラーでは、1,686 個のキーワードを巡回させ、同じキーワードを時間をおいて何回も検索することで、多くの対話を収集する。この際、既に対話として収録したものと同じツイートが検索されることがある。そのため、既に対話として取得したツイートの ID を保存し、これと同じ ID のツイートは取得しないようにすることで、同じ対話を重複して抽出しないようにした。

3.4 不適切な対話の除去

収集した対話に対し、それが不適切な対話であるかを判定する。適切であると判定された対話のみ対話コーパスに収録する。3.1 節で述べた予備調査の結果を踏まえ、不自然な対話を除去する 4 つのルールを設ける。以下、それぞれのルールの詳細を説明する。

3.4.1 複数のセリフがある発話を含む対話の除去

1 つのツイートの中に他者のセリフが含まれるとき、対話として成立しないことが多い。3.1 節の予備調査では、表 3.1 に示した不適切さの要因 (2) に相当し、100 件のうち 2 件の対話が該当した。例えば、以下の対話 D1 では、1 つの発話に複数人のセリフがあり、対話としては不自然である。

D1

u_1: 客「何？ここ禁煙なの？」ワシ「禁煙にさせて頂いてますすみません～」客「チッ」

セリフは括弧で囲まれて表記されることが多いため、一つのツイートの中に括弧の組が複数存在するかをチェックすることで、複数のセリフがあるために不適切な対話を除去する。しかし、括弧はセリフを用いる際に使われるだけでなく、単語を強調する際にも用いられる。以下の対話 D2 と対話 D3 はその例である。

D2
u_1: 座右の銘は「猪突猛進」「切磋琢磨」です。
u_2: へぇぇ〜 !!　いいね !!

D3
u_1: 好きな言葉は「明日は明日の風が吹く」と「失敗は成功のもと」かな。
u_2: わかる〜

これらの対話例に示すように、「猪突猛進」「切磋琢磨」のような名詞を強調したり、「明日は明日の風が吹く」「失敗は成功のもと」のような格言を強調するときにも括弧が使われる。このような場合は、括弧が複数使われていても対話として不自然ではないので、除去するべきではない。

上記を踏まえ、複数のセリフがある発話を含む対話を除去するルールを図 3.3 に示す。これ以降、このルールを $\mathbf{R_{line}}$ と記す。条件 1. は複数の括弧の有無をチェックする。条件 1. のみでは、D2 のような括弧内がセリフではなく短い単語であり、括弧が強調に用いられている対話も除去してしまうため、条件 2. を加える。また、対話例 D3 のように、括弧内が長い文であってもセリフではない場合があるが、このようなときは括弧の次に助詞が続くことが多い。そのため、括弧の次の単語が助詞ではないという条件 3. を設定し、D3 のような対話を除去しないようにする。

以下の条件を全て満たす発話 (ツイート) があるとき、その対話を除外する。

1. 1 つの発話中に複数の括弧 (“「” と “」”) を含む。
2. 括弧の中の文字が 6 文字以上である。
3. 括弧の次の単語が助詞以外である。

図 3.3: 複数のセリフがある発話を含む対話を除去するルール $\mathbf{R_{line}}$

3.4.2　短い発話を含む対話の除去

極端に短いツイートを含む対話は対話として成立しない場合が多い。3.1 節の予備調査では、表 3.1 に示した不適切さの要因 (4)(5) に相当し、100 件のうち 3 件が該当した。以下の対話 D4 はその例である。発話 u_2 は「を」というひらがなだけであり、意味を理解できないため、対話として不適切である。

D4
u_1: 少し寝ようと思って寝たらこの時間つんだ
u_2: を

予備調査やそれ以降に実施した調査の結果、ひらがな以外にも、全角スペース、句点、読点のみの発話が含まれる対話が確認された。これらの対話も意味を理解することができないため、対話として不適切と言える。ただし、1 文字程度の極端に短い発話が含まれていても、不自然な対話にならないときもある。例を D5 に挙げる。

D5
u_1: 寝て起きたら朝だった
u_2: 笑

上の対話例における「笑」のような単語や、疑問符や感嘆符、間投詞に関しては、相手の発話に対して反応を示していると考えられる。対話としても不自然ではないので、除去するべきではない。

一方、絵文字のみのツイートも注意を要する。絵文字は何らかの意味を表しているので、絵文字のみの発話があっても、対話全体の内容を理解できることも多い。しかし、本研究で実装した対話収集システムでは、文字コードを UTF-8 に変換してから保存する仕様になっているため、絵文字を正常に表示することができない。したがって、絵文字のみのツイートは文字化けされた状態で表示されるため、意味を理解できない。一方、対話コーパスとして利用する場面を考えると、絵文字はテキストではないため、前処理の段階で削除されることも多い。絵文字のみのツイートに対してこのような前処理を行うと、ツイートが空文字列になってしまうという問題もある。したがって、本研究では、絵文字のみからなるツイートを削除する。例を D6 に挙げる。「??」は文字化けしている絵文字であり、実際にはスマイルマークである。

D6
u_1: 無事大学に合格しました！
u_2: ??

上記を踏まえ、短い発話を含む対話を除去するルールを図 3.4 に示す。これ以降、このルールを $\mathbf{R_{short}}$ と記す。条件 1. ではひらがな 1 文字のみのツイートのうち、間投詞を除いている。これは、間投詞はひらがな 1 文字であっても意味を理解できることが多いためである。

以下の条件のいずれかを満たす発話 (ツイート) があるとき、その対話を除外する。

1. 1 文字のひらがなのみのツイート。ただし、間投詞となる「あ」「え」「お」を除く。
2. 全角スペース、句点、読点のみのツイート。
3. 絵文字のみのツイート。

図 3.4: 短い発話を含む対話を除去するルール $\mathbf{R_{short}}$

3.4.3 画像・URL を含む対話を除去するルール

画像を含むツイートが対話の中に存在するとき、対話として不自然な場合が多い。3.1 節の予備調査では、表 3.1 に示した不適切さの要因 (1) に相当し、100 件のうち 38 件の対話が該当した。画像を参照しなければ対話の内容を理解することができない対話の例を D7 に示す。

D7

u_1: この目を光らせるとバイザーが立たず、バイザー立てると
　　他が真っ暗 … だれか一緒に光り物撮影しよ！ IMAGE
u_2: かっこいいやつだ !!!

IMAGE は画像を表す。D7 は、ウェブブラウザ上では画像を含めて図 3.5 のように表示される。

図 3.5: 画像を含むツイートの例

D7 の u_2 の発話「かっこいいやつだ!!!」は、表示されている画像の内容を知らなければ理解することができない。本研究では、テキストのみを保存し対話コーパスを構築することを仮定としているため、画像を見ないと理解できない対話は除去する。一方、画像を含むツイートがあっても、その画像なしでも対話の内容を理解できる場合がある。例を D8 に挙げる。

D8	u_1: お祭り楽しい IMAGE u_2: いいなー

この例では IMAGE にはお祭りの画像が表示されているが、「お祭り楽しい」「いいなー」というやりとりは、画像がなくとも理解できるし、自然な対話である。画像を参照する対話を除去するのは簡単である。Twitter API で保存したツイートでは、画像はその参照先の URL が保存されているので、ツイート内に URL(”http” で始まる文字列) があるかをチェックすればよい。しかし、Twitter では画像を含むツイートが多く、画像を含む対話をすべて除去してしまうと、最終的に構築される対話コーパスの量が小さくなりすぎる可能性がある。できれば、対話例 D7 のように画像を参照しないと理解できない対話のみを除去し、対話例 D8 のような画像を参照しなくても理解できる対話は残したい。

なお、Twitter 上では、URL は画像ではなく他のウェブページのアドレスを表すこともある。このとき、画像と同様に、参照先のウェブページの内容を知らなければ対話を理解することができない場合があり、不適切な対話として除去するべきである。本研究では、不適切な対話かどうかを判定するためには、URL の参照先が画像であっても他のウェブページであっても、参照先の画像もしくはウェ

ブページの内容が対話の理解に必要かどうかを判断する際、これらを同様に扱う。すなわち、URL が画像を参照するか他のウェブページを参照するかを区別せず、URL を含むツイートがあったとき、それを含む対話が不適切であるかを同じルールで判定する。

上記を踏まえ、画像・URL を含む対話を除去するルールを図 3.6 に示す。

以下の条件のいずれかを満たす発話 (ツイート) があるとき、その対話を除外する。

1. URL と指示語 (「これ」など) を含む。
2. URL を含むツイートに対するリプライに「それ」などの指示語を含む（「その通り」「そのうち」を除く）。
3. URL のみ、又は URL とハッシュタグのみのツイート。

図 3.6: 画像・URL を含む対話を除去するルール $\mathbf{R_{image}}$

条件 1. は、ツイート内に URL に加えて指示語があるときに、それを含む対話を不適切と判定する。不適切な対話の例 D7 では指示語「この」が含まれているが、適切な対話の例 D8 では含まれていない。D7 のように「これ」「この」などの指示語が含まれていると画像の内容を参照する可能性が高いため、URL と指示語の両方を含むときに対話を除去するように条件を設定した。条件 1. で使用した指示語の一覧を表 3.3 に示す。

表 3.3: 条件 1. で用いた指示語

これ、ここ、この、こう、こちら、こっち、こんな

次に条件 2. について説明する。条件 2 に該当する対話の例を D9 に示す。

D9

u_1: 面白かった IMAGE
u_2: その本いいよね

D9 では、画像を含むツイート u_1 は指示語を含まないが、その次のツイート u_2 には「その」という指示語が含まれている。この場合も条件 1. と同様に、画像を含むツイートに対するリプライに「それ」「その」などの指示語が含まれていると、その画像の内容を参照している可能性が高い。ただし、指示語が含まれている場合であっても、画像の内容を参照する可能性が低い場合がある。例を D10 と D11 に挙げる。

D10
u_1: お金は大事 IMAGE u_2: その通り

D11
u_1: また行きたいなー IMAGE u_2: そのうちね

対話例 D10 や D11 では、画像を含むツイートに対するリプライが指示語である「その」を含んでいても画像の内容に言及しているわけではない。つまり画像の内容を知らなくても対話の内容を理解できる。上記を踏まえ、条件 2. では、リプライに指示語を含み、かつ「その通り」「そのうち」という単語を含まないときに対話を不適切と判定している。条件 2. で使用した指示語の一覧を表 3.4 に示す。

表 3.4: 条件 2. で用いた指示語

それ、そこ、そちら、そっち、その

条件 3. では、まず URL のみのツイートを含む対話を除外する。URL は対話中の発話としては完全に不適切である。また、URL とハッシュタグのみのツイートを含む対話も除外する。ハッシュタグは対話中の発話としては不自然であり、またハッシュタグは画像やリンク先のウェブページの内容と関連している可能性があり、これらを参照しないと対話の内容を理解できない可能性が高いためである。

3.4.4 不特定多数のユーザへの呼びかけを含む対話の除去

不特定多数への呼びかけから始まる対話は、対話として成立しないことが多い。3.1 節の予備調査では、表 3.1 に示した不適切さの要因 (3) に相当し、100 件のうち 3 件の対話が該当した。以下の対話 D12 はその例である。

D12
u_1: 名前誰か僕につけてくれない？ u_2: フクロウ

u_1 は不特定多数へ呼びかけている発話である。この他に、大喜利のお題を発信するツイートや告知をしているツイートなどが不特定多数への呼びかけの例として挙げられる。通常、対話とは二者もしくは少人数による発話のやりとりを指す。したがって、多くのユーザを対象とした呼びかけは対話として適切ではない。

不特定多数への呼びかけを含む対話を検出する方法を Twitter 上で投稿された実際のツイートを参考にして検討したところ、対話の起点となるツイートへのリプライ数が多いとき、そのツイートが不特定多数への呼びかけである場合が多いことがわかった。このことから、対話の起点となるツイートに対し、それへのリプライ数が閾値以上のときに不適切な対話として除去する方法が考えられる。しかし、Twitter API ではリプライ数を取得することができない。そこで、あるツイートに対するリプライ数を取得するための手法をいくつか検討した。

手法 1. スクレイピングを利用する手法

Twitter に投稿されたツイートをウェブブラウザ上で表示すると、そのツイートに対するリプライ数が表示される。そのため、ツイートを表示した HTML ファイルを取得し、これを解析して、リプライ数を得ることができる。このように HTML ファイルを解析して情報を抽出する技術はスクレイピングと呼ばれている。しかしながら、Twitter 社は自社のサーバに対するスクレイピングを禁止している。このため、この手法を本研究に適用することはできない。

手法 2. 疑似的にリプライ数を取得する手法

Twitter では、あるユーザ A のツイートにリプライしたときには、その先頭に「@ A」という形で元のツイートのユーザ名が表示される。したがって、「@ A」という文字列を含むツイートを検索すれば、ユーザ A のツイートのリプライ数を見積もることができる。この手法の詳細な手続きを図 3.7 に示す。ツイート①をリプライ数を取得する対象のツイートとする。このツイートを投稿したユーザは A である。まず、「@ A」という文字列を含むツイートを検索する。Twitter API のキーワード検索では、約 1 週間以上前のツイートを検索することができない。そこで、1 週間以上前のツイートを検索することができる Getoldtweets というライブラリを用いた。ただし、ツイート① 以前に投稿されたツイートは明らかにリプライではないため、ツイート①のタイムスタンプをチェックし、それ以降に投稿されたツイートのみを取得する。図ではユーザ B、C、D のツイートが取得される。次に、Twitter API を用いて、取得されたツイートのリプライ元をチェックし、それがツイート① であることを確認する。なぜなら、「@ A」を含むツイートは、ユーザ A の別のツイートへのリプライの可能性があるためである。図の例では、ユーザ B と C のツイートのリプライ元はツイート①であるが、ユーザ D のツイートは異なる。最終的に、ツイート①のリプライ数は 2 となる。ただし、上記の手法で取得できるのはリプライ数の見積もりであり、正確なリプライ数ではない。また、上記の手法を実装したところ、ツイートの検索やリプライ元のチェックに時間を要し、ツイートのリプライ数を効率よく推測することはできなかった。一方、本研究では取得した大量の対話に対して、その最初の発話が不特定多数への呼びかけであるかをチェックすることが求められるため、リプライ数を高速に推測することも求められる。このため、手法 2 を実際に適用することは難しいと判断した。

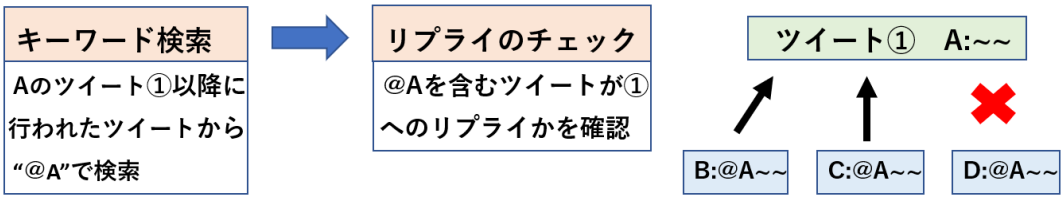

図 3.7: リプライ数を取得する手順の例

検討した手法 1. も手法 2. も実現が難しかったため、本研究では、除去の対象とする不特定多数への呼びかけを含む対話を大喜利のみとする。大喜利は特定の Twitter ユーザがアカウントを開設して運営していることが多く、大喜利のお題のツイートは比較的容易に検出できる。まず、ユーザ名やプロフィールに「大喜利」というキーワードを含んだユーザの ID を取得し、大喜利アカウントであるかを人手で判定し、大喜利アカウントのリストを作成する。結果として、61 個の大喜利ユーザからなるアカウントリストを得た。そのリストを用いて、リストに含まれているユーザ ID と収集した対話の最初のツイートのユーザ ID を比較する。ユーザ ID が一致している場合は大喜利ツイートとみなして除去する。

上記を踏まえ、不特定多数のユーザへの呼びかけを含む対話を除去するルールを図 3.8 に示す。これ以降、このルールを $\mathbf{R}_{\mathbf{invite}}$ と記す。

以下の条件を満たす発話 (ツイート) があるとき、その対話を除外する。

1. 対話の起点となるツイートのユーザが大喜利アカウントのとき、その対話を除去する。

図 3.8: 不特定多数のユーザへの呼びかけを含む対話を除去するルール $\mathbf{R}_{\mathbf{short}}$

3.5　対話コーパスの整備

Twitter から取得した対話に対し、3.4 節で説明した 4 つのルールを用いて不適切な対話を除去した後、残りの対話を収録した対話コーパスを構築する。対話をコーパスに収録する際には、対話にメタデータを付与する。ここではその構想を示す。

収集した対話に対して以下のメタデータを付与する。

対話のメタデータ

1. 対話 ID
   対話を識別するためのユニークな番号である。

2. 収集に用いたキーワード
   対話を収集する際に用いたツイートの検索キーワード。3.3 節で述べた単語親密度の高い 1,686 個の単語のうちの 1 つである。

3. 対話長
   収録した対話の長さ (対話を構成するツイートの数) である。

4. 対話の開始が特定のユーザに向けた発話であるか
   Twitter では、あるユーザ (@abc とする) のツイートに対してリプライすると、「@ abc ツイート本文」のように、冒頭に元のツイートのユーザ ID が表示される。一方、Twitter ユーザは特定のユーザ (@def とする) に対するメッセージを新規に送信する際に、メッセージの冒頭にユーザ ID を付与し、「@def ツイート本文」のような形式でツイートを投稿することがある。このようなツイートは別のツイートに対するリプライではないが、リプライと区別することが難しい。稀ではあるが、対話の起点となるツイートが特定のユーザに向けたメッセージであり、その冒頭にユーザ ID が表示されていることがあるが、リプライから開始される不自然な対話であるように見える。そのため、対話の起点となるツイートの冒頭がユーザ ID であるかを区別するフラグをメタデータとして付与する。

一方、対話中の個々の発話データに対しても、メタデータを付与する、その詳細を以下に示す。

発話のメタデータ

1. ツイート ID
   Twitter API による取得されるツイートの識別番号である。

2. ユーザー ID
   ツイートを投稿したユーザの識別番号である。これも Twitter API によって取得される。対話コーパスでは、ユーザ ID は話者の ID として利用できる。

3. タイムスタンプ
   ツイートが投稿された時刻の情報である。

4. 画像へのリンクである URL を含むか
   本研究で提案するルール $\mathbf{R_{image}}$ は画像を含む不適切な対話を除去するが、不適切な対話であっても適切であると誤って判定することもある。対話コーパスの利用者の利便性を考慮し、判定が誤っている可能性があることを明示するために画像の有無をメタデータとして付与する。

5. ウェブページへのリンクである URL を含むか
   画像を含む対話同様、ウェブページへのリンクである URL を含む対話も誤って適切と判定され、対話コーパスに収録される場合がある。その可能性があることを明示するためにこのメタデータを付与する。

画像へのリンクである URL とウェブページへのリンクである URL に対して別々のメタデータを付与するのは、両者で処理を分けたいユーザの利便性を考量したためである。なお、Twitter では、画像の URL のドメインは常に pbs.twing.com であるため、URL のリンク先が画像であるかウェブページであるかは容易に識別できる。

第4章　評価実験

本章では、提案手法の評価実験について述べる。まず、4.1 節では、Twitter から対話を収集し、大規模な対話コーパスを構築する実験について報告する。4.2 節では、構築した対話コーパスから、提案手法によって不適切な対話を除去した結果を報告する。4.3 節では、提案手法によってどれだけ正確に不適切な対話を除去できたかを評価する。ここでは、提案する 4 つのルール全てを使ったときの結果を評価する。4.4 節では、不適切な対話を除去するルールを個別に評価する。

4.1　Twitter からの対話の収集

3.3 節で述べた手続きに従い、Twitter から対話を収集した。収集は、2019 年 6 月 8 日から 2019 年 12 月 25 日にかけて実施した。収録した対話の概要を表 4.1 に示す。

表 4.1: Twitter からの対話収集

対話数	平均対話長	平均発話長
92,207	9.50	67.73

合計対話数は 92,207 であり、多くの対話を収集することができた。また、平均対話長は 9.50 であり、比較的長い対話を獲得できたことがわかった。実際に収集した対話の例を D13 に示す。

D13
u_1: チロル並みに血色ある唇してる
u_2: えっそれはやばい
u_3: 初めてこんなにガサガサしてるし赤くムラになって…痛いん…
u_4: リップクリームいっぱい塗って〜〜 !!!!

4.2　不適切な対話の除去

4.1 節で収集した対話に対し、3.4 節で提案した 4 つのルールを用いて不適切な対話を検出した。結果を表 4.2 に示す。

表 4.2: 各ルールによって除去された対話数

ルール	除去された対話数
R_{short}	297
R_{line}	669
R_{image}	16289
R_{invite}	0

一番多く不適切な対話を検出したルールは $\mathbf{R_{image}}$ であった。Twitter から収集した対話には画像を含んだツイートが含まれることが多いことや、$\mathbf{R_{image}}$ によって適切な対話も誤って除去してしまった可能性があることから、不適切な対話の検出数はその他のルールと比較してかなり多い。一方、$\mathbf{R_{short}}$ と $\mathbf{R_{line}}$ によって除去された対話の数は 0.3%もしくは 0.7%程度と少ないことが分かった。これは、真に不適切な対話の数が少ないためか、もしくは、本研究では不適切な対話を除去するだけではなく適切な対話をできるだけ除去しないことも考慮してルールを設計したため、真に不適切な対話であっても適切な対話と判定されて検出できなかったため、という 2 つの要因が考えられる。$\mathbf{R_{invite}}$ で除去された対話は 1 つもなかった。しかしながら、除去されなかった対話の中には 3.4.4 項で挙げた対話例 D12 のような不特定多数へ呼びかけている発話も少なからず含まれていたと思われる。そのため、不特定多数への呼びかけから始まる不適切な対話を除去する別の方法を考える必要がある。例えば、3.4.4 項で検討したようなリプライ数に基づいて不特定多数への呼びかけを検出する手法が考えられる。このとき、ツイートに対するリプライ数を効率的に求める技術が必要である。また、他者へ呼びかけている際によく用いられている単語、句、表現を調査し、これらを検出の手掛かりとするアプローチも考えられる。

4.3 不適切な対話の除去の評価

本節では、本論文で提案した 4 つのルールによって、不適切な対話をどれだけ正確に除去できたかを評価する。

4.3.1 実験の手順

まず、評価用データを作成する。収集した対話のうち、ランダムに 100 件の対話を選択し、テストデータとする。次に、テストデータのそれぞれの対話に対し、それが対話として適切か不適切か、言い換えれば対話コーパスに収録するのに適した対話かどうかを人手で判定する。判定作業は 2 名の作業者が独立に行った。2 名による判定の一致率と κ 係数を表 4.3 に示す。

表 4.3: テストデータに対する 2 者の判定の一致率と κ 係数

一致率	κ 係数
0.82	0.60

判定が一致しなかった原因を分析したところ、主に画像を含む対話に対して判定が割れていた。1 名の作業者は URL 先の画像を見て対話として成立しているかを判定していたが、もう 1 名の作業者は画像を閲覧することなく対話文のみで判定したため、対話の内容が画像とどれだけ関係しているかの判断が分かれたことが要因として考えられる。また、以下の対話 D14 のようなチケットや物品の売り買いを求める発話が起点となる対話について、適切かどうかの判定が分かれていることが多かった。

D14
u_1: うらたぬきさんワンマン
【譲】期限内のお支払い【求】たぬワン大阪指定席 2 連
こちら切実に求めております。なるべく住所変更できる方がいいです。
u_2: 初めまして。こんばんは。検索より失礼致します。検討違いかと思いますが大阪 2 連のスタンディングを所持しております。なるべく良番を回せるようにさせて頂きますのでご検討いただけますと幸いです。

テストデータの対話に対し、提案した 4 つのルールを全て適用し、個々の対話が適切かどうかを自動的に判定する。この結果を人手による判定結果と比較することで、提案手法の性能を評価する。提案手法は取得した対話から不適切なものを検出するが、本研究の目的は適切な対話を残して対話コーパスを構築することであるため、適切な対話と不適切な対話のそれぞれの検出の精度、再現率、F 値を評価指標とする。6 つの評価指標のそれぞれの定義を式 (4.1), (4.2), (4.3), (4.4), (4.5), (4.6) に示す。OK は適切な対話、NG は不適切な対話を検出するタスクの精度、再現率、F 値であることを表す。

$$\text{精度 (NG)} = \frac{\text{不適切と正しく判定された対話数}}{\text{提案手法によって不適切と判定された対話数}} \tag{4.1}$$

$$\text{精度 (OK)} = \frac{\text{適切と正しく判定された対話数}}{\text{提案手法によって適切と判定された対話数}} \tag{4.2}$$

$$\text{再現率 (NG)} = \frac{\text{不適切と正しく判定された対話数}}{\text{テストデータにおける不適切な対話数}} \tag{4.3}$$

$$\text{再現率 (OK)} = \frac{\text{適切と正しく判定された対話数}}{\text{テストデータにおける適切な対話数}} \tag{4.4}$$

$$\text{F 値 (NG)} = \frac{2 \cdot \text{精度 (NG)} \cdot \text{再現率 (NG)}}{\text{精度 (NG)} + \text{再現率 (NG)}} \tag{4.5}$$

$$\text{F 値 (OK)} = \frac{2 \cdot \text{精度 (OK)} \cdot \text{再現率 (OK)}}{\text{精度 (OK)} + \text{再現率 (OK)}} \tag{4.6}$$

4.3.2 実験結果と考察

実験結果を表 4.4 に示す。「作業者 1」「作業者 2」の列は、それぞれの作業者による判定を正解としたときの精度、再現率、F 値を示している。

表 4.4: 不適切な対話の検出手法の評価

	作業者 1	作業者 2
精度 (NG)	0.75	0.75
再現率 (NG)	0.32	0.43
F 値 (NG)	0.45	0.55
精度 (OK)	0.70	0.81
再現率 (OK)	0.94	0.94
F 値 (OK)	0.80	0.87

不適切な対話の再現率は 0.32 もしくは 0.43 と低い。反対に、適切な対話を検出するタスクについては、最低でも作業者 1 の精度が 0.70 と全体的に数値が高い。これは、適切な対話を適切であると判定できたが、不適切な対話を不適切であると判定できなかった場合が多いことを表す。

次に、不適切な対話を検出するタスクにおいて、提案手法による判定と正解判定の対応をまとめた対応表を表 4.5 に示す。「作業者 1」と「作業者 2」はそれぞれの作業者の判定を正解としたときの対応表を表す。

表 4.5: 不適切な対話の検出の対応表

作業者 1

		(正解) NG	(正解) OK
(判定)	NG	12	4
	OK	25	59

作業者 2

		(正解) NG	(正解) OK
(判定)	NG	12	4
	OK	26	68

False Negative(不適切な対話を検出できなかった誤り) が多い。これにより、表 4.4 に示すように、不適切な対話検出の再現率 (再現率 (NG)) が低くなっている。その要因の多くは画像を含む不適切な対話を検出できていなかったためである。対話 100 件の中で、画像を含んだ対話は 45 件であり、作業者 1 による判定を正解としたときは、その中の False Negative の数は 13 であった。これは全体の False Negative の数（25）の約半数である。画像を含む対話の除去ルール $\mathbf{R_{image}}$ を改善することが必要であると考えられる。画像を含む対話のうち、実際には不適切ではあるが検出されなかった対話について考察する。

以下の対話例 D15 は、顔が写っている画像を紹介する発話とそれに対する反応を含む対話である。

D15
u_1: 死んだ様に寝てる　IMAGE
u_2: 一瞬ひっ！てなるけど安心してぐっすり眠っている顔じゃあ

発話 u_1 では「顔」という名詞は含まれていないが、発話 u_2 は顔という単語を使って画像の内容に言及している。このような対話は、画像を含んだ発話に続く発話に「顔」や「山」などの画像に関連しうる名詞が含まれているかどうかによって判定するという方法や、画像を含んだ発話とそれに続く発話で用いられている名詞を比較し、前者の発話には含まれていなかった名詞が後者の発話に含まれているかどうかによって判定するという方法が考えられる。具体的な判定方法の例を以下に挙げる。

1. 特定の名詞（「顔」「山」「景色」などの画像に写されやすい名詞）が画像を含んだ発話の次の発話に含まれている場合は除去する。

2. 画像を含んだ発話の次の発話に、対話の中で初めて用いられた名詞が含まれる場合は除去する。

1. については、画像に写されやすく、画像を含んだ発話の次の発話に現れやすい名詞を調査するなどの準備が必要である。このような名詞をどの程度用意するかによって除去する対話の量をある程度コントロールできる。キーワードとする名詞を多く用意すれば、それだけ不適切と判定される対話の量が多くなる。

2. については、対話で用いられた名詞を記録しておき、これらと現在の発話（画像を含む発話の次の発話）内に含まれる単語と比較することで判定できる。1. と

比較すると多くの対話がこの条件を満たすことが予想され、適切な対話を過度に不適切と判定してしまう危険性がある。

1. と 2. のどちらの手法がより適切であるかを検討し、このどちらかのルールを $\mathbf{R_{image}}$ に組み込むことによって、不適切な対話検出の再現率を向上させることができると考えられる。

また、名詞だけでなく動詞についても、画像への参照を暗に示唆する単語が存在する。以下の対話例 D16 には、「描いた」という動詞を含む発話がある。

D16	u_1: フォロワーさんから案を頂いて描いた　IMAGE u_2: 夏！って感じで涼しげでいいね〜

この対話では、描いたものが画像として表示されており、その発話の返答も画像の内容について触れている。このような対話を除去するための方法として、以下が挙げられる。

- 画像を含んだ発話もしくはその次の発話に、画像に関連しうる動詞（「描く」「撮る」など）を含む場合は除去する。

「描く」や「撮る」という動詞が含まれていてかつ画像がある場合、その画像の内容は描いたものや撮ったものである場合が多いと考えられる。また、その発話に対する返答の中に「描く」や「撮る」といった動詞が含まれるときも画像の内容に触れている場合が多いと考えられる。このように、画像への参照を暗に示唆する動詞のリストをあらかじめ用意し、それを含む対話を除去することで、不適切な対話の検出の再現率を向上させることができると考えられる。

4.4　個々のルールの評価

本節では、不適切な対話を除去するルールを個別に評価する。4.3 節の実験では、ランダムに選択した 100 個の対話のみを評価の対象としていた。しかし、ルールによっては、100 個の対話の中から 1 個も不適切な対話を検出していないものもあり、個々のルールの性能を評価するには不十分である。ここでは、各ルールについて、そのルールによって不適切な対話と判定された対話の中から、ランダムに 50 件を選択し、それらの対話が不適切であるかを人手で判定した。人手による正解判定は 2 名の作業者が行った。ルールの評価基準は精度とする。精度の定義を式 (4.7) に示す。

$$\text{精度} = \frac{\text{ルールによって検出されかつ不適切である対話数}}{\text{ルールによって検出された全対話 (50)}} \tag{4.7}$$

$\mathbf{R_{invite}}$ については、人手で作成した大喜利ユーザーのリストを用いており、常に正しく判定できるとみなせる。また、表 4.2 に示したように、$\mathbf{R_{invite}}$ によって実際に検出された対話がないため、ここでは評価はしない。結果を表 4.6 に示す.

表 4.6: 各ルールによる不適切な対話の検出精度

作業者 1

ルール	精度
$\mathbf{R_{short}}$	0.96 (48/50)
$\mathbf{R_{line}}$	0.74 (37/50)
$\mathbf{R_{image}}$	0.78 (39/50)

作業者 2

ルール	精度
$\mathbf{R_{short}}$	0.94 (47/50)
$\mathbf{R_{line}}$	0.76 (38/50)
$\mathbf{R_{image}}$	0.78 (39/50)

3 つのルールとも精度が 7 割を超えており、良好な結果であったと言える。精度が一番高かったのは $\mathbf{R_{short}}$ であったが、このルールで除去された対話の大多数は絵文字を含む対話であった。また、$\mathbf{R_{line}}$ や $\mathbf{R_{image}}$ の正解率も高いが、修正するべき問題がいくつか見つかった。各ルールによる判定の誤りの要因の分析は後で報告する。

それぞれのルールの評価データに対する判定の作業者 2 名の一致率と κ 係数を表 4.7 に示す。

表 4.7: 個々のルールの評価における 2 者の判定の一致率と κ 係数

	一致率	κ 係数
$\mathbf{R_{short}}$	0.94	0.37
$\mathbf{R_{line}}$	0.80	0.47
$\mathbf{R_{image}}$	0.92	0.77

一致率が一番高いのは $\mathbf{R_{short}}$ であったが、κ 係数は一番低かった。$\mathbf{R_{short}}$ は、精度が高いことからも分かるように、検出された 50 個の対話の大多数が不適切な対話であったため、少数の判定の不一致によって κ 係数が大きく下がったためである。例えば、「ん」のみの発話を含む対話の判定が分かれていた。1 文字で意味を成すかの見解が個人で分かれるようなひらがなをどのように扱うべきか考える必要がある。$\mathbf{R_{line}}$ では、一つの発話に複数のセリフと発話者自身のコメントが含まれている対話の判定が分かれていた。その後の対話の流れなども含め、それが対話として成立しているかの判定に相違があった。$\mathbf{R_{image}}$ では、4.3.2 項で述べた判定の一致率と κ 係数に関する考察と同様に、発話が画像の内容には触れているがその画像がなくても対話として成立していると判定した場合などで判定が揺れることがあった。

次に、各ルールについて、その精度を向上させるために修正すべき点を考察する。

4.4.1 短い発話を含む対話を除去するルール ($\mathbf{R_{short}}$) の考察

このルールで除去された対話の多くが、絵文字だけからなる発話を含む対話であった。その割合は検出された対話の 9 割ほどであった。既に述べたように、本研究で保存された対話コーパスでは絵文字は正常に表示されないため、絵文字だけの発話は意味を理解することができない。したがって、これらについては、不適切な対話を正確に除去できたと言える。

一方、適切な対話を誤って除去した例もあった。例えば、以下の対話のように、単語を省略し返答を一文字で表しているような場合については、対話としては成立しているが $\mathbf{R_{short}}$ によって誤って除去されていた。

D17
u_1: もう 2 時とかま？
u_2: ま

D18
u_1: 海外に行ってみたい
u_2: ね

対話例 D17 における「ま」は「まじ」が、対話例 D18 における「ね」は「だね」を省略したものと考えられるため、対話としては成立している。このような誤検出を防止するために、1 文字でも発話として意味を成すひらがなとそうでないひらがなを分け、後者のひらがな 1 文字だけの発話が含まれるときのみ対話を除去する方法が考えられる。1 文字だけでも発話として成立するひらがなのリストを得るためには、以下のようなアプローチが考えられる。

1. 一文字のひらがなの発話を含む対話を収集する。
2. 収集した対話から適切なものを人手で抽出し、その中の 1 文字発話で使われている文字を収集する。
3. 収集された頻度の高い文字を 1 文字だけでも発話が成立するひらがなとする。

以上を踏まえ、$\mathbf{R_{short}}$ を以下のように再定義する。

$\mathbf{R_{short}}$: 以下のいずれかの条件を満たすツイートを含む対話を削除する。

1. 1 文字のひらがなのみのツイート。ただし、1 文字でも発話として成立する文字が使われている場合を除く。
2. 全角スペース、句点、読点のみのツイート。
3. 絵文字のみのツイート。

修正されているのは条件 1 である。 この修正により、ルールの精度がどのように改善されるかを検証するのは今後の課題である。

4.4.2 複数のセリフがある対話を除去するルール ($\mathbf{R_{line}}$) の考察

このルールで誤検出された対話の中には、以下のような対話例 D19 のように、括弧内が長い文ではあるがセリフではない場合があった。このような場合は、括弧が複数回使われていても、対話中の発話として自然である。

D19 | u_1:「夢中になれることにこだわりがある」のは職人です。大抵他の業種は「マニュアル通り」なんで。

誤検出の割合を減少させるには、括弧で囲まれた文がセリフかそうでないかを識別しなくてはならない。しかし、括弧内の文がセリフか否かを判断するのは一般には難しい。そこで、括弧内のテキストではなく、括弧の後に続く単語を手掛かりに括弧内がセリフかどうかを判定することが考えられる。対話例 D19 の u_1 の括弧の後に用いられている単語は名詞の「の」である。現在の除去ルールでは、括弧の後に続く単語が助詞以外であれば除去するため、D19 の対話も除去されていた。しかし、エラー分析では、括弧の後に来る単語が「の」である場合は、括弧内がセリフではない場合が多かった。

そこで、不適切な対話の誤検出を減らすため、3.4 節に示した $\mathbf{R_{line}}$ を以下のように修正することが考えられる。

3. (修正前) 括弧の次の単語が助詞以外である。

3. (修正後) 括弧の次の単語が助詞または「の」以外である。

しかし、以下の対話例 D20 のように括弧の次の単語が「の」の場合でも、括弧の中がセリフを表すことがある。したがって、括弧の後が「の」である発話を抽出し、実際に除去してよい対話の割合が多いかどうかを検討した上で、ルールを修正する必要がある。

D20 | u_1: 私「朝から何も食べてないから何か食べたいな」の「グミならあるよ」

4.4.3 画像・URL を含む対話を除去するルール ($\mathbf{R_{image}}$) の考察

このルールによって適切な対話であるが誤って除去されてしまった対話例 D21 を以下に挙げる。

D21
u_1: この度「くまカフェ」の福岡での開催が決定しました！ IMAGE
u_2: 行きます！

この対話は指示語「この」と画像を含んでいたため除去されていたが、画像の内容が分からなくても (画像がなくても) 対話としては成立している。この対話で用いられている指示語「この」が画像以外のことを指しているのが誤検出の要因だと考えられる。また、本研究では URL を含むツイートに対するリプライが指示語を含む場合には不適切な対話として検出するが、例外的に「その通り」と「そのうち」を含んでいるときは検出していない。しかし、この 2 つ以外にも画像や URL を参照しない指示語もあると考えられる。このような例外的な指示語を網羅的に列挙し、ルールを精緻化する必要がある。

また、以下の対話例 D22 のように、指示語が画像を指している場合であっても対話としては成立している場合があった。

D22
u_1: 兄貴に誕プレを買ってもらい、これで仕事頑張れよって言われて 少し感動しちゃいました！ IMAGE
u_2: いいお兄さんやん！

この対話のように、画像を含んだ発話に対する返答が画像について触れていない場合があり、対話としては適切である。しかし、このような対話は不適切な対話と区別することが難しい。画像や URL を含む発話に対するリプライの発話を大量に収集し、それを分析することで、画像や URL を参照している文の言語的特徴、参照していない文の言語的特徴を明らかにした上で、その特徴を考慮したルールの設計が必要である。

第5章　おわりに

5.1　まとめ

本研究では、Twitter からツイートとリプライの連鎖を対話とみなして収集し、その中から不適切な対話を除去するためのルールを 4 種類 ($\mathbf{R_{line}}$、$\mathbf{R_{short}}$、$\mathbf{R_{image}}$、$\mathbf{R_{invite}}$) 考案した。$\mathbf{R_{line}}$ は、複数のセリフがある対話を不適切な対話とみなし、括弧が複数ある、括弧の中の文字が 6 文字以上である、括弧の次の単語が助詞以外である、という条件をすべて満たすツイートを検出することでそれらを除去するルールである。$\mathbf{R_{short}}$ は短い発話を含む対話を不適切な対話とみなし、間投詞以外の一文字のひらがなのみ、句読点などの記号のみ、絵文字のみのツイートが含まれている対話を除去するルールである。$\mathbf{R_{image}}$ は画像・URL を含む対話のうち、画像を参照しないと内容を理解できない対話を不適切な対話とみなし、ツイートが画像や URL を含むこと、画像の周辺に「これ」「それ」などの指示語があることなどを検出の条件とし、それらを除去するルールである。$\mathbf{R_{invite}}$ は対話の起点となるツイートが不特定多数への呼びかけである対話を不適切な対話とみなし、これを削除するルールである。当初は、対話の起点となるツイートに対するリプライ数が多いときに不特定多数への呼びかけと判定する手法を検討したが、技術的な問題によって実現が難しいことがわかった。本研究では、大喜利を運営しているユーザーのリストをあらかじめ人手で作成し、対話の起点となるツイートのユーザーがそのリストに含まれている場合に、その対話を除去するという簡単な手法を採用した。

提案手法の評価実験を行った。実際に Twitter から対話を収集し、提案する 4 つのルールで不適切な対話を除去し、その精度、再現率、F 値を測った。適切な対話を除去しないという観点では、$\mathbf{R_{invite}}$ 以外のルールについては良好な結果が得られた。しかし不適切な対話を除去するという観点では、全体的に再現率が低く、特に $\mathbf{R_{image}}$ の改良が必要であることがわかった。とはいえ、不適切な対話を除外しつつ、平均対話長が 10 程度の比較的長い対話をおよそ 76,000 程度含む大規模な対話コーパスを構築することができた。

5.2 今後の課題

実験の結果、複数のセリフを含む対話を除去するルールや、画像・URL を含む対話を除去するルールについては、十分な精度が得られなかった。今後の課題として、これらのルールを改善することが挙げられる。特に、画像・URL を含む対話を除去するルールについては、画像を参照しないと対話全体を理解できるかできないかを判定しなければならないが、これは画像とテキストの関連性を測るという難しい問題を解決する必要がある。今後、画像や URL を含む多くの対話を分析し、ルールを精緻化したい。また、本研究では除去すべき対話を調査する予備実験に用いた対話が 100 対話のみであったため、不適切な対話のタイプを網羅的に集めることができなかった可能性がある。より多くの対話を調査することで、不適切な対話を包括的に分析する必要がある。このために、対話データの考察と合わせて、対話の収集を今後も続けることで対話コーパスを増強し、様々な対話を分析することで、より正確に不適切な対話のみを除去できるようなルールを考案していきたい。

謝辞

本研究を進めるにあたり、研究の方針を始めとする熱心な御指導を頂いた白井清昭准教授に心より御礼申し上げます。また、本論文を審査してくださった飯田弘之教授、岡田将吾准教授、池田心准教授にこの場を借りて御礼申し上げます。

参考文献

[1] 浅原正幸. クラウドソーシングによる単語親密度の推定. 言語処理学会第 25 回年次大会, pp. 45–48, 2019.

[2] Rafael E. Banchs, Haizhou Li. IRIS: a Chat-oriented Dialogue System based on the Vector Space Model, Proceedings of the ACL 2012 System Demonstrations, pp. 37-42, 2012.

[3] 別所史浩, 原田達也, 國吉康夫. リアルタイムクラウドソーシングと Twitter 大規模コーパスを利用した対話システム. 情報処理学会研究報告 自然言語処理（NL）206 巻 13 号, pp. 1-8, 2012.

[4] 江頭 勇祐, 柴田 知秀, 黒橋 禎夫. 雑談対話システムにおける強化学習を用いた応答生成モジュールの選択, 言語処理学会第 17 回年次大会発表論文集, pp. 654–657, 2012.

[5] 東中竜一郎, 川前徳章, 貞光九月, 南泰浩, 目黒豊美, 堂坂 浩二, 稲垣博人. 2 ツイートを用いた対話モデルの構築. 言語処理学会第 18 回年次大会, pp. 567–570, 2012.

[6] 福田彩子, 荒木健治, ジェプカラファウ. 対話システムにおける対話履歴要約の有効性について, 情報処理学会研究報告 自然言語処理（NL）195 巻 11 号, pp. K1-K6, 2010.

[7] 福田 拓也, 若林 啓. 雑談システムにおける Twitter データからの統計的バックチャネル応答抽出手法, 人工知能学会論文誌 33 巻 1 号, pp. DSH-H_1-10, 2018.

[8] 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,29 巻 1 号, pp. 21-31, 2014

[9] Ryan Lowe, Nissan Pow, Iulian Serban, Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems,SIGDIAL 2015, pp. 285-294, 2015.

[10] 大竹 裕也, 萩原 将文. 高齢者のための発話意図を考慮した対話システム, 日本感性工学会論文誌,11 巻 2 号, pp. 207-214, 2012.

[11] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations，In HLTNAACL, pp. 172–180, 2010.

[12] Alan Ritter, Colin Cherry, William B. Dolan, Data-Driven Response Generation in Social Media, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 583-593, 2011.

[13] 山口 貴史, 井上 昂治, 吉野 幸一郎, 高梨 克也, Ward, N. G., 河原 達也. 傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成, 人工知能学会論文誌, Vol. 31, No. 4, pp. C-G31_1-10, 2016.