# 発話の交換を考慮した対話システムにおけるユーザ感情推定手法の検討

宮川由衣 [1]　加藤大貴 [2,1]　松平茅隼 [1]　平山高嗣 [3,1]　駒水孝裕 [1]　井手一郎 [1]
[1] 名古屋大学　[2] 名城大学　[3] 人間環境大学
{miyakaway, katoh, matsuhirac}@cs.is.i.nagoya-u.ac.jp
t-hirayama@uhe.ac.jp
taka-coma@acm.org
ide@i.nagoya-u.ac.jp

## 概要

非タスク指向対話システムは目標が明確でないため，システムの性能を評価することが難しい．システムの究極の目標はユーザの満足であるため，対話におけるユーザの感情遷移が重要な指針となる．そこで我々は，対話システムと人との対話の様子を収めたマルチモーダルコーパスを使用し，システム発話に対するユーザの感情推定モデルを構築している．本発表では，その際に 1 つの発話のみでなく，どれだけ過去の発話が重要であるか検討した結果を報告する．具体的には，心象アノテーションと発話の書き起こしを用いて BERT モデルを Fine-tuning した．実験の結果，システム発話とユーザ発話との対に加えて 1 つ過去のユーザ発話まで考慮することで，最高 0.89 の F1 スコアを達成した．

## 1　はじめに

自然言語処理や音声認識の発達に伴い，Amazon 社の Alexa [1] や Apple 社の Siri [2] などの音声対話システムが実用化されている．また，チャットボットやオープンドメイン対話システムなど、非タスク指向の対話システムの開発にも大きな関心が寄せられている．非タスク指向対話システムの品質向上はユーザの対話体験にとって重要であるが，タスクの達成のような目標が明確でないため，システムの性能評価が難しい．ユーザが対話に満足しているか，システムとストレスなく対話できているかは，対話システムの性能について考える際に重要な要素であることから，対話によるユーザの感情遷移を自動的に推定することができれば，重要な指針となると考えられる．

そこで本研究では，発話の書き起こしのみから，システム発話に対するユーザの感情を推定するモデルを提案する．一般に，対話においてはユーザの感情は 1 つの発話のみに明確に表れるとは限らず，それまでの対話の流れに影響される．そのため本研究では，過去の発話も考慮したモデルを構築する．そして，どれだけ過去の発話まで考慮した場合にユーザの感情を最も良く推定できるかを実験的に明らかにする．

以降，まず 2 節で関連研究を，3 節で使用するコーパスを各々紹介する．次に 4 節でコーパスへの感情ラベルの付与について，5 節で提案手法について各々述べる．最後に 6 節で実験について報告し，7 節で本論文をまとめ，今後の課題について述べる．

## 2　関連研究

Wei ら [3, 4] や Hirano ら [5] は，対話システムと人との対話の様子を収めたマルチモーダルコーパス [6, 7] を使用し，対話システムとの対話におけるユーザの内的状態を推定している．Wei ら [3] は，音響特徴，視覚特徴，言語特徴を RNN（Recurrent Neural Network），LSTM（Long Short-Term Memory），GRU（Gated Recurrent Unit）に入力し，システム対話におけるユーザの満足度を認識するマルチモーダルモデルを提案している．さらに，対話全体におけるユーザの感情と，1 対の受け答え（以下，交換）におけるユーザの感情の関係も分析している [4]．一方 Hirano ら [5] は，弱教師付き学習とマルチタスク学習を融合することで，心象アノテーションの信頼性の低さやデータ数の少なさを解決している．

しかし，これらの研究では，過去の発話を考慮することによるユーザの感情の推定性能への影響は検

討されていない．

## 3　対話コーパス

本研究では，先行研究 [3, 4, 5] で用いられている対話コーパス（Hazumi）[6, 7] に含まれる 3 つのデータセット Hazumi1712，Hazumi1902，Hazumi1911 を使用する．以下，これらについて紹介する．

### 3.1　対話の収録

対話システムを模擬して， Wizard-of-Oz 方式により別室から人が操作[1)] し，実験参加者が対話システムと雑談する様子を 1 名あたり約 15 分間収録している．各データセットは，ビデオカメラと Microsoft Kinect センサにより記録された参加者の行動や表情，さらにシステムと参加者の発話の書き起こしを収録している．

使用するデータセットのうち，Hazumi1712 では，事前にいくつかの話題に対する参加者の興味の有無を調査した上で，参加者の興味がある 3 話題と興味がない 3 話題をとりあげて対話が行なわれている．話題の例としてはスポーツ，ドラマ，芸能人，ゲーム，電車などがある．参加者は 20 代から 50 代の男女 29 名（男性 14 名，女性 15 名）である．

一方，Hazumi1902 と Hazumi1911 では，参加者が楽しんでいる時間が長くなるように話題を調整しながら対話が行なわれている．Hazumi1712 との違いは，参加者がある話題に興味を示さなかった場合に，対話システムが別の話題に変えている点である．参加者は 20 代から 70 代の男女 60 名（男性 25 名，女性 35 名）である．

### 3.2　心象アノテーション

これらのデータセットには，参加者とは別のアノテータ 5 名により心象アノテーションが付与されている．アノテーションの付与単位は発話における 1 交換（あるシステム発話開始時刻から次のシステム発話開始時刻まで）である．アノテータは対話の様子を記録した映像を始めから順番に視聴しながら，参加者のシステム発話に対する感情について，各交換に対して 7 段階で付与している．ここで，1 がネガティブ（楽しくない，話し続けたくない，不満，困惑など），7 がポジティブ（楽しい，話し続けたい，満足など）である．

1)　人間のオペレータが遠隔から，実験参加者の様子を見ながら専用のインタフェースを通じて，システムの応答を選択．

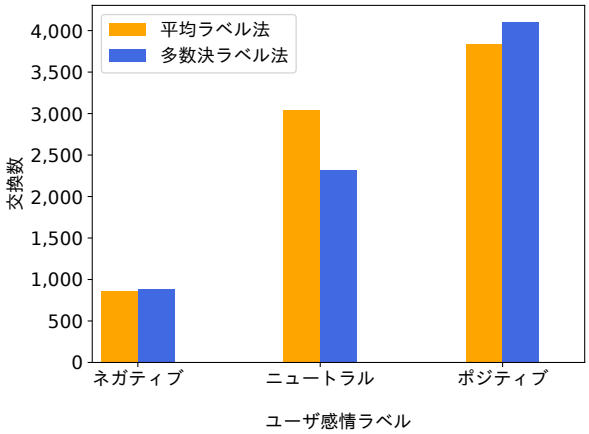


**図 1**　付与されたユーザ感情ラベルの分布

## 4　ユーザ感情ラベルの付与

本研究では，ネガティブ，ニュートラル，ポジティブの 3 クラス分類を行なう．そのために，アノテータ 5 名分の心象アノテーションを集約することで得られる 3 クラスのラベルをユーザ感情ラベルと定義する．各交換に対してユーザ感情ラベルを付与する際に，以下の 2 種類の方法を検討する．

**平均ラベル法**　アノテーションの平均をとる方法である．5 名分の心象アノテーションの平均をとり，先行研究 [5, 8] に倣い，3.5 未満はネガティブ，4.5 以上はポジティブ，その他はニュートラルに分類する．

**多数決ラベル法**　アノテーションの多数決をとる方法である．平均ラベル法でアノテーションがポジティブとネガティブに分かれた場合にニュートラルに分類されてしまう問題の解消を目的とする．5 名分の心象アノテーション各々について，3.5 未満はネガティブ，4.5 以上はポジティブ，その他はニュートラルに分類した後，3 つ以上一致したラベルに分類する．

7,743 交換に対して，平均ラベル法と多数決ラベル法各々によるユーザ感情ラベルの分布を図 1 に示す．なお，多数決ラベル法で 3 つ以上一致するラベルがなかった交換は削除したため，同法については合計 7,306 交換になった．

## 5　感情推定モデルの構築

感情推定モデルを構築する際に，学習データ数の少なさを考慮して，少数のデータでも高精度の分類が期待できる事前学習済み BERT（Bidirectional Encoder Representations from Transformers）[9] の Fine-

表 1　学習データの例

| 連結数 $N$ | ID | 発話 | 入力テキスト | 平均ラベル | 多数決ラベル |
|---|---|---|---|---|---|
| 2 | 1 | sys | コンサートとかには行きますか？ | ニュートラル | ポジティブ |
| | | user | [SEP] 行く時間がないので行けてないです, 行きたいのはありますが | | |
| 2 | 2 | sys | なまで演奏をきくと、迫力があって、感動するみたいですね！ | ポジティブ | ポジティブ |
| | | user | [SEP] そうですね, 全然違いますね, テレビとかとは違います | | |
| 4 | 1 | sys | コンサートとかには行きますか？ | ポジティブ | ポジティブ |
| | | user | [SEP] 行く時間がないので行けてないです, 行きたいのはありますが | | |
| | | sys | [SEP] なまで演奏をきくと、迫力があって、感動するみたいですね！ | | |
| | | user | [SEP] そうですね, 全然違いますね, テレビとかとは違います | | |
| 4 | 2 | sys | なまで演奏をきくと、迫力があって、感動するみたいですね！ | ポジティブ | ポジティブ |
| | | user | [SEP] そうですね, 全然違いますね, テレビとかとは違います | | |
| | | sys | [SEP] わたしも実際に、演奏をきいてみたいものです. | | |
| | | user | [SEP] ぜひ聴いてみてください | | |

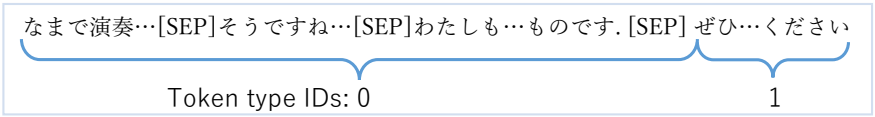


図 2　Token type IDs の使用例

tuning を行なう．

ここでは，あるシステム（ユーザ）発話開始時刻から次のユーザ（システム）発話開始時刻までを 1 発話と定義し，モデルへの入力単位は，連続する $N$ 発話を連結したものとする．$N = 1$ のとき入力単位は最後のユーザ発話のみとなり，過去の発話を考慮しない条件と等価である．システム発話とユーザ発話の間には文章の区切りを表す [SEP] トークンを挿入する．$N = 3$ 以上の場合には最後の交換に付与されたユーザ感情ラベルを使用する．表 1 は $N = 2$ 及び $N = 4$ における学習データの例である．ここで，3 列目の「sys」はシステム発話，「user」はユーザ発話を表す．

また，ユーザ感情ラベルには最後のユーザ発話の影響が大きいという仮説のもと，最後のユーザ発話を明確化することを目的とし，入力を Encode する際に Token type IDs を付与する．図 2 は表 1 における連結数 $N = 4$ の ID 2 に対する Token type IDs の例である．最後のユーザ発話に 1 を付与し，その他には 0 を付与する．

# 6　実験

4 節の方法で，対話の様子を収録した映像を視聴して付与された心象アノテーションに基づいて定義したユーザ感情ラベルについて，発話の書き起こしのみから推定するという問題設定で，分類実験を行なった．

## 6.1　実験条件

ユーザ感情ラベルを付与した書き起こしデータを 8 : 1 : 1 の比率に分け，それぞれ学習データ，検証データ，テストデータとした．この学習データと検証データを用いて，日本語の感情分析データによって事前学習された BERT モデル [10] を Fine-tuning した．

発話の連結数 $N$ について，$N = 1, 2, 3, 4, 6, 10$ の 6 通り測定して，結果を比較した．更に，直近のシステム発話 1 発話のみを入力とした場合の結果も比較した．

## 6.2　ベースラインモデル

ベースラインモデルとして，ML-Ask [11] を用いた．これは，感情表現辞典 [12] に基づく 2,100 の感情語によるパターンマッチングを行なうことで 10 種類の感情（怒，哀，怖，厭，昂，驚，恥，喜，好，安）を推定し，推定した感情に基づいて 3 クラス分類を行なうモデルである．具体的には，入力されたテキストはネガティブ（怒，哀，怖，厭），ニュートラル（昂，驚，恥），ポジティブ（喜，好，安）の感情語の個数に応じて 3 クラスのいずれかに分類される．

本実験では，感情語が存在しないために分類できなかったものは全てニュートラルに分類した．

連結数 $N = 1, 2$ のテストデータに対してユーザ感情ラベルを推定し，その分類性能を評価した．

**表 2**　3 クラスの平均 F1 スコア

| モデル | 連結数 $N$ | 平均ラベル | 多数決ラベル |
|---|---|---|---|
| ML-Ask | 1 (user) | 0.46 | 0.45 |
| （ベース） | 2 | 0.52 | 0.52 |
| BERT | 1 (sys) | 0.69 | 0.77 |
| （提案） | 1 (user) | 0.81 | 0.86 |
| | 2 | **0.83** | 0.87 |
| | 3 | **0.83** | **0.89** |
| | 4 | 0.82 | 0.88 |
| | 6 | 0.81 | 0.87 |
| | 10 | 0.81 | 0.84 |

### 6.3　実験結果

実験結果として，3 つの感情クラスの平均 F1 スコアを表 2 に示す．3 列目がユーザ感情ラベル付与に平均ラベル法を採用した場合の F1 スコア，4 列目が多数決ラベル法を採用した場合の F1 スコアである．また，「1 (sys)」はシステム発話のみを入力単位とした場合，「1 (user)」はユーザ発話のみを入力単位とした場合を表している．各方法で最高の F1 スコアを太字で示した．平均ラベル法では $N = 2$ と $N = 3$ の場合のスコア（0.83）が，多数決ラベル法では $N = 3$ の場合のスコア（0.89）が最高であった．一方，$N = 6$ 以上連結した場合や，ML-Ask モデルで分類した場合には性能が低かった．また，平均ラベル法と比較して多数決ラベル法の方が総じて良い性能を得られた．

### 6.4　考察

ML-Ask モデルの性能が低かったのは，テストデータ中に感情語が存在しない発話が多かったためと考えられる．1 交換（$N = 2$）において，感情語が存在しない交換は平均ラベル法では 774 交換中 513 交換，多数決ラベル法では 731 交換中 482 交換であった．辞書を使用する手法では，辞書に含まれない感情語を発している場合や明確にポジティブ／ネガティブな単語を発していない場合の感情推定に限界があるため，機械学習による手法より劣ると言える．

また，平均ラベル法と比べて多数決ラベル法の方が良い性能を得られた．これは，心象アノテーションを閾値処理して得た感情ラベルが 3 つ以上一致しないような，感情が明確でない交換が，多数決ラベルにおいて除外されたためであると考えられる．

また表 2 において，平均ラベル法と多数決ラベル法どちらの場合でも「1 (sys)」よりも「1 (user)」の方が性能が高くなった．このことから，ユーザ感情ラベルにはシステム発話と比較してユーザ発話の影響が大きいことが分かる．

**表 3**　4 発話における感情クラス別の F1 スコア

| 感情クラス | 平均ラベル | 多数決ラベル |
|---|---|---|
| ネガティブ | 0.46 | 0.57 |
| ニュートラル | 0.80 | 0.91 |
| ポジティブ | 0.88 | 0.90 |

表 3 に提案手法において $N = 4$ とした場合のクラス別の F1 スコアを示す．ポジティブやニュートラルと比較してネガティブの性能が低いことが分かる．この原因として，対話内で明確にネガティブな言葉（嫌い，つまらないなど）がほぼ発されないため，他クラスよりも推定が難しいことや，他クラスと比較してデータ数が少ないため，十分に学習できていないことなどが考えられる（図 1 参照）．ネガティブクラスに対する分類性能の向上は今後の課題である．

## 7　おわりに

本研究では，システム発話に対するユーザの感情を推定するモデルを提案した．

Hazumi データセットの心象アノテーションに基づいて，2 種類の方法で 3 クラスのユーザ感情ラベルを付与し，このラベルと発話の書き起こしを用いて BERT モデルを Fine-tuning した．入力単位とする発話の連結数を変えてモデルをそれぞれ学習し，感情クラス分類の性能を比較した．

実験の結果，平均ラベル法では入力単位を 2 発話及び 3 発話とした場合，多数決ラベル法では 3 発話とした場合に最高の F1 スコアが得られ，過去の発話を考慮することで分類性能が向上する可能性が示唆された．

分類性能向上のための今後の課題として，発話の連結方法の改善が挙げられる．例えば，Token type IDs の拡張による全ての発話の区切りの明確化や，最後の発話に対する重み付けが考えられる．

また，状況に応じた発話の連結数の切り替えも挙げられる．例えば，話題が切り替わった際には過去の発話を連結しない，発話が短い場合にのみ過去の発話も連結する，などが考えられる．

更に，今後は画像，音声，3 次元動作などのデータも加えてマルチモーダルな分析を行なっていきたい．

## 謝辞

本研究の一部は，JSPS 科研費 JP22H03612 の助成を受けたものである．

## 参考文献

[1] Amazon.com, Inc. Amazon Alexa, (2023-1-9 閲覧). https://www.amazon.co.jp/b?node=10406417051.

[2] Apple Inc. Siri, (2023-1-9 閲覧). https://www.apple.com/jp/siri/.

[3] Wenqing Wei, Sixia Li, Shogo Okada, and Kazunori Komatani. Multimodal user satisfaction recognition for non-task oriented dialogue systems. In Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 586–594, 2021.

[4] Wenqing Wei, Sixia Li, and Shogo Okada. Investigating the relationship between dialogue and exchange-level impression. In Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 359–367, 2022.

[5] Yuki Hirano, Shogo Okada, and Kazunori Komatani. Recognizing social signals with weakly supervised multitask learning for multimodal dialogue systems. In Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 141–149, 2021.

[6] 駒谷和範. マルチモーダル対話コーパスの設計と公開. 日本音響学会誌, Vol. 78, No. 5, pp. 265–270, 2022.

[7] 大阪大学産業科学研究所. マルチモーダル対話コーパス（Hazumi）, (2023-1-6 閲覧). https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/.

[8] Yuki Kubo, Ryo Yanagimoto, Hayato Futase, Mikio Nakano, Zhaojie Luo, and Kazunori Komatani. Team OS's system for dialogue robot competition 2022. Computing Research Repository arXiv Preprint, arXiv:2210.09928, 2022.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, 2019.

[10] daigo. bert-base-japanese-sentiment, (2023-1-6 閲覧). https://huggingface.co/daigo/bert-base-japanese-sentiment.

[11] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. A system for affect analysis of utterances in Japanese supported with Web mining. 知能と情報, Vol. 21, No. 2, pp. 30–49, 2009.

[12] 中村明. 感情表現辞典. 東京堂出版, 1993.